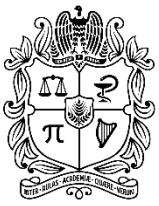


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Relación entre los eventos públicos y los accidentes de tránsito a partir del análisis de datos en Bogotá

Yefersson Adrian Rojas Real

Universidad Nacional de Colombia
Departamento de ingeniería de sistemas e industrial
Bogotá, Colombia
2020

Relación entre los eventos públicos y los accidentes de tránsito a partir del análisis de datos en Bogotá

Yefersson Adrian Rojas Real

Trabajo de investigación presentado como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:

PhD. Cesar Augusto Pedraza

Codirector:

MsC. Camilo Albeiro Gutiérrez

Línea de Investigación:

Computación aplicada

Grupo de Investigación:

PLaS

Universidad Nacional de Colombia

Departamento de ingeniería de sistemas e industrial

Bogotá, Colombia

2020

(Dedicatoria)

A Dios, a mis padres, a mis hermanos por su ayuda incondicional y apoyo, a mi novia por su paciencia y amor, para ellos este trabajo.

Yefersson Adrian Rojas Real

Resumen

El estudio de los factores involucrados en los accidentes de tránsito y la movilidad es un tema de gran interés. Dado que afectan la vida cotidiana de los habitantes de cada gran ciudad. Los eventos públicos, como conciertos y eventos deportivos, pueden tener un gran impacto, debido a la existencia de un mayor riesgo de accidente en las aglomeraciones públicas. Este estudio busca identificar analíticamente la relación entre eventos públicos y accidentes de tráfico, mediante datos extraídos de páginas web sobre eventos públicos celebrados en la ciudad de Bogotá entre septiembre 2018 y julio 2019. Con respecto al caso de accidentes de tránsito, se obtuvo información de la aplicación Waze.

Para comprender la relación entre eventos públicos y accidentes, se propuso un análisis de espacio-tiempo para eventos públicos y accidentes de tránsito utilizando una variedad de técnicas geoestadística, un sistema de información geográfica (SIG) y una base de datos espacial. Entre los resultados obtenidos se tiene la variedad de la ocurrencia de los accidentes en cuanto a la presencia o ausencia de los eventos públicos. En los diferentes análisis se tiene en cuenta la franja horaria de la ocurrencia de los eventos, al igual la distancia entre el accidente y el lugar de ocurrencia del evento Público. En la obtención de resultados se utilizaron algunas técnicas de agrupación espacial como el índice de Moran, índice de vecino más cercano para el análisis de identificaciones de patrones de concentración o dispersión y métodos estadísticos que caractericen el comportamiento de los accidentes.

Palabras clave: Análisis espacial, Sistemas de información geográfica, geoestadística, accidentes de tránsito, eventos públicos.

Abstract

The study of the factors involved in traffic accidents and mobility is a topic of great interest. Since they affect the daily lives of the inhabitants of each large city. Public events, such as concerts and sporting events, can have a major impact, due to the existence of an increased risk of accidents in public agglomerations. This study seeks to analytically identify the relationship between public events and traffic accidents, using data extracted from web pages about public events held in the city of Bogota between September 2018 and July 2019. With regard to the case of traffic accidents, information was obtained from the Waze application.

To understand the relationship between public events and accidents, a space-time analysis for public events and traffic accidents using a variety of geostatistical techniques, a geographic information system (GIS) and a spatial database was proposed. Among the results obtained is the variety of the occurrence of accidents in terms of the presence or absence of public events. The different analyses take into account the time frame of the occurrence of the events, as well as the distance between the accident and the place of occurrence of the Public event. Some spatial clustering techniques such as the Moran index were used to obtain results, nearest neighbour index for the analysis of identification of concentration or dispersion patterns and statistical methods characterising accident behaviour.

Keywords: Spatial analysis, Geographic information systems, Geostatistics, Traffic accidents, Public events

Contenido

Lista de figuras

Pág.

Lista de tablas

Introducción

Una serie de componentes y eventos influyen en las incidencias y accidentes de tránsito. Por ejemplo, el estado de las vías, condiciones del automóvil, estado del conductor, factores climáticos y demanda presencial de personas en ciertas áreas públicas causan congestión y accidentes de tránsito [1], [2].

La distribución espacial de los accidentes en Bogotá, D.C. (Colombia) va relacionada con la caracterización de la población en cuanto a la generación de concentraciones o dispersión de personas y la presencia de eventos públicos. Las concentraciones y aglomeraciones de personas generan congestión vehicular y conllevan a un mayor riesgo de accidentes de tránsito. También cabe añadir que las zonas urbanas poseen características diferentes, referente a lo comercial, residencial y demás puntos recreativos enmarcados en un nivel económico, en lo que implica que los factores que influyen en los accidentes de tránsito no son los mismos, según se mostró en los diferentes trabajos y enfoques de investigaciones [1], [3]–[5].

Abordando el tema de los eventos públicos estos se comprenden como aglomeraciones de público o eventos masivos, el cual hace referencia a la concentración de un gran número de personas productos de una convocatoria individual o colectiva, de forma general y abierta [9]. En este contexto se enmarcan eventos públicos como eventos deportivos, festivales, conciertos entre otros. Se presentan una serie de problemáticas en cuanto a los eventos públicos y los accidentes de tránsito. Además de lo mencionado y a la revisión de literatura, se tiene que los estudios que abordan la identificación de los eventos públicos como posible causa de los accidentes de tránsito son pocos a comparación a los dedicados a la caracterización e identificación de la gravedad de las lesiones de los accidentes de tránsito, como se muestra en la en la tabla 1. Algunos estudios también presentan baja precisión en modelos y congruencia en los resultados con

las diferentes técnicas de procesamiento, [28]. Al igual la calidad de datos que se han utilizado para resolver problemas similares procedentes a la investigación de los eventos y los accidentes de tránsito, a la falta de fuentes de información de eventos públicos, en la cual se tenga datos de geolocalización ubicación y fecha de los eventos [8]. A partir de las problemáticas mencionadas anteriormente se plantea la siguiente pregunta de investigación: ¿Cómo determinar la relación de los eventos públicos en los accidentes de tránsito a modo de factor influyente en la ciudad de Bogotá?

Para solucionar la pregunta formulada se propuso como objetivo general:

Identificar la relación entre los eventos públicos y los accidentes de tránsito por medio de la extracción de información de páginas web con técnicas de análisis de datos.
Como objetivos específicos.

1. Diseñar un método de raspado web para la extracción de datos de páginas web sobre eventos públicos en la ciudad de Bogotá.
2. Determinar el grado de asociación entre las ocurrencias de los eventos públicos y accidentes de tránsito en la ciudad de Bogotá
3. Identificar patrones de agrupación o dispersión de los eventos accidentales y públicos en las diferentes zonas y puntos críticos de la ciudad de Bogotá.

Para llevar a cabo los objetivos se utilizó en esta investigación un enfoque cuantitativo para el análisis de la relación de los eventos públicos con los accidentes de tránsito, que permita establecer la relación entre estas dos variables, el cual hace uso de la geoestadística y técnicas de análisis de datos cuantitativos [38]. Con un diseño de investigación no experimental de tipo explicativo, puesto que no se va a inferir en la modificación de variables o elementos, sino se basa en la obtención datos geoespaciales de cierto tiempo que sustente la relación entre los eventos públicos y accidentes vehiculares [12].

En el proceso de análisis de datos implica el uso de técnicas de "*Scraping Web*" herramientas de raspado de información web para la obtención de los datos [39]. La información de los eventos públicos, se compone de listas diarias de acontecimientos de la ciudad de Bogotá entre el mes de septiembre 2018 y julio de 2019. Entre la información obtenida, tanto de los eventos públicos como de los accidentes de tránsito, se tiene en

cuenta datos como ubicación, fecha y hora de ocurrencia, al que se hace frente a las problemáticas de información incompleta. Con el
propósito de extraer elementos que permitan el desarrollo de la distribución espacial de los accidentes de tránsito y los eventos públicos [8]. En desarrollo de la metodología se contempla tres fases principales: extracción de datos, preprocesamiento de datos y análisis exploratorio de datos. La extracción de datos infiere en el raspado de información de páginas web para luego en el procesamiento realizar la estructuración, filtrado y el almacenamiento de datos, en la tercera fase se tiene la aplicación de métodos geoestadísticos, análisis y evaluación de resultados.

Finalmente, este documento se estructura con cinco capítulos. El primer capítulo 1 se presentan los antecedentes del tema que consiste en los estudios que se han realizado, los enfoques que estos presentan y las técnicas de minería de datos utilizadas; el segundo capítulo toca las dimensiones teóricas las cuales se emplean durante la investigación; el tercer capítulo se presenta el desarrollo del primer objetivo en lo que infiera en el método de extracción de datos sobre los eventos públicos; en cuarto capítulo se analiza en detalle los métodos y procesos en el análisis espacial; el quinto capítulo incorpora el análisis estadístico espacio-temporal entre los eventos públicos y los accidentes de tránsito. Por último, la conclusión donde se realiza la discusión de los resultados obtenidos.

1. Antecedentes

La distribución espacial de los accidentes de tránsito y los eventos públicos, permite identificar patrones, concentración o dispersión de los eventos, así como, conocer la frecuencia y ocurrencias de estos eventos en los diferentes corredores viales y áreas de una ciudad [7]. Para comprender la relación entre los eventos públicos y los accidentes, Cerquera [8] propone el análisis espacial para estos dos eventos utilizando una serie de técnicas de estadísticas y métodos matemáticos para analizar geográficamente las relaciones espaciales de dichos eventos. Por su parte, Bailey y Gattrel [9] definen en términos amplios el análisis espacial como el estudio cuantitativo de fenómenos que están localizados en el espacio. En el mismo sentido, Legendre et al. [10] indica que el análisis espacial comprende estudiar cuantitativamente datos espaciales o datos que contengan información de localización. Por lo tanto, los accidentes de tránsito y los eventos públicos al tener una ocurrencia en espacio y en tiempo, se ubican geográficamente y son susceptibles de ser analizados espacialmente. En este capítulo se presentan los antecedentes de los estudios que analizan accidentes de tránsito y las técnicas computacionales empleadas para dicho análisis, con el propósito de tener como base el enfoque y los resultados que se han obtenido, para impartir en las problemáticas que se han presentado y presentar alternativas en el desarrollo de la investigación.

1.1.1 Estudios realizados en los accidentes de tránsito

Los accidentes de tránsito son un fenómeno que implican la interacción de varios componentes, algunos como el conductor, vehículo y la carretera que se complementa de otros elementos esenciales, causantes de un déficit en la seguridad vial. Algunos

investigadores como Martin, Baena, Garach, López y Juan de Oña [3] se centralizan en la identificación de los puntos peligrosos y elementos de la red vial como parte causal de los accidentes. Por su parte, Xi un, Zhonghao Zhao, Li Wei y Wang Quan [5] establecieron un análisis de los elementos causales con mayor influencia en los accidentes de tránsito, en las que identifican la probabilidad del accidente, a partir del estado operativo en que se encuentran estos elementos. Estos y otros estudios solo se han enfocado en cuantificar el impacto que tienen algunos factores directos a los accidentes de tránsito como causa principal, como fallos mecánicos o del conductor y no se han tenido en cuenta factores externos, como señales de tránsito, estado de las vías o eventos que intervienen en la movilidad, entre otros. Por ejemplo, Investigadores de la Universidad de Novi Sad [1] en Serbia han asociado el clima como un factor de gran causalidad de los accidentes, en donde las condiciones climáticas aumentan la probabilidad de accidentes con gravedad de lesiones altas. Sin embargo, estos estudios no son aplicables a todas las ciudades, dado que las características del estado climático son diferentes.

Algunas investigaciones en donde se tiene como objetivo la caracterización y la predicción de los accidentes, identifican posibles patrones y tendencias causales de estos [2], [11] Los diferentes enfoques en tema de accidentalidad y los objetivos propuestos en cada uno de los estudios extraídos de la literatura se presentan en la **Tabla 1-1**.

Tabla 1-1. Estudios tratados en accidentes de tránsito en los últimos 10 años

Enfoque	Objetivo	Autores
Caracterización	Describir las variables que intervienen en los accidentes de tránsito	Zhang, He, Gao & Ni (2018) [12].
		He , Liu, Zhou & Zhong (2017) [2].
		Tiwari (2017) [13].
		Serna, Gerrikagoitia & Ruiz (2017) [14].
		Akomolafe, & Olutayo (2013) [4].
	Identificar puntos críticos y elementos susceptibles de mejora de las vías	Martín, Baena & de Oña (2014)[3].
		Akomolafe & Olutayo (2013).
Gravedad de las lesiones	Identificar el grado de severidad de las lesiones causadas por la accidentalidad vehicular	Sameen & Pradhan (2017). [15]
		Olutayo & Eludire (2014) [16].
		S.Shanthi & Geetha Ramani (2012) [17].
		Kashani & Ranjbari (2011) [18]
		Chaozhong & Xinping (2009) [19].
		Chong & Paprzycki (2008) [20].

Causalidad	Identificar la relación entre factores climatológicos y los accidentes de tránsito	Novkovic & Stefanovic (2017) [1]
		Xi, Zhao, Li & Wang (2016) [5]
	Identificar la relación entre eventos sociales y los accidentes de tránsito	Grimm, Fristo & Sharma (2017) [6].
		Shekhar, Setty & Mudenagudi (2016) [11]
		Cerquera Escobar (2013)[8].
		Fuentes & Hernández (2009)[21]
	Determinar la influencia del estado de los diferentes componentes partícipes en los accidentes de tránsito	Tseng Agresti (2005) [22]
		Jain, Ahuja & Mehrotra (2016) [23].

Fuente: elaboración propia con base en las referencias mencionadas en la tabla.

Para los estudios realizados en la caracterización y problemáticas de los accidentes de tránsito se han empleado diferentes algoritmos y técnicas de minería de datos. En la mayoría de los casos, de acuerdo al enfoque, los autores emplean aprendizaje supervisado y aprendizaje no supervisado. En la siguiente sección se menciona algunas de las diferentes técnicas utilizadas en los estudios realizados sobre accidentes de tránsito.

1.1.2 Técnicas de minería de datos en estudios de accidentes de tránsito

En las diferentes investigaciones (ver tabla 1), han aplicado varias técnicas de aprendizaje de máquina, por ejemplo, en el análisis para determinar la gravedad de las lesiones como consecuencia de accidentes de tránsito, Sameen y Pradhanemplearon algoritmos de redes neuronales, donde compararon los modelos red neuronal recurrente (RNN), Perceptron multicapa (MLP) y modelos Bayesian Logistic Regression (BLR). Como resultado el modelo RNN superó al MLP y BLR, con una precisión de validación del 71.77 % y con un 10 % de mayor efectividad a los demás algoritmos [15]. A pesar de los resultados el modelo RNN tenía algunas limitaciones como la dificultad con la precisión de la estimación de las probabilidades de salida, debido a la falta de factores, ya que estos son un prerrequisito de entrada del modelo propuesto. Olutayo y Eludire [16] en su estudio emplearon técnicas de redes neuronales y árboles de decisión donde obtuvieron mejores resultados con el algoritmo Id3 de árboles de decisión que con las redes neuronales, el cual superó el número de casos clasificados de manera correcta. Aunque otros investigadores enfatizan en el algoritmo artificial de árbol de decisión de red neuronal (ANN-DT), que extrae árboles

de decisión binarios de una red neuronal entrenada, como lo hacen Chong, Abraham y Paprzycki [20], para el estudio de la gravedad de las lesiones, donde este enfoque híbrido obtiene mejores resultados en comparación a las Support Vector Machine (SVM), redes neuronales y árboles de decisión de forma independiente.

Basados en otros estudios y técnicas, Zhanga, Heb, Gaod y Ming Nic [12], apuestan a técnicas de aprendizaje como una Deep Belief Network (DBN) para el análisis de datos de redes sociales en la detección de accidentes de tránsito. Estos investigadores muestran las grandes ventajas de la DBN, puesto que, en sus estudios, este supera a la red neuronal artificial (ANN) con una capa oculta, el etiquetado de secuencia con unidades de memoria largas a corto plazo LSTM y a las máquinas de vectores de soporte (SVM). Otras de las técnicas utilizadas son reglas de asociación para identificar los principales factores asociados a un accidente de tráfico el cual implementan clúster para el agrupamiento de datos y el algoritmo apriori para generar estas reglas [24]. A pesar que estas técnicas generan buenos resultados, algunas reglas generadas para los conjuntos de datos muestran que las reglas de asociación no revelan información apropiada, que puede estar relacionada con un accidente [5]. En el estudio para la correlación entre los diferentes factores climáticos y ocurrencias de accidentes de tráfico, Novkovic, Arsenovic, Sladojevic, Anderla y Stefanovic utilizaron técnicas de conjunto, árboles de decisión y SVM como Random Forest, K-nearest neighbor (KNN), AdaBoost y J48, además proponen el uso de Random Forest dada la tasa de más éxito en la predicción de los sucesos positivos [1].

Un estudio realizado en Chicago, Estados Unidos sobre accidentes de tránsito y eventos públicos, se basó en las técnicas SVM y redes neuronales para el análisis de los datos y al igual que algunos estudios ya mencionados bajo estas técnicas, los resultados no fueron los esperados, el modelo SVM era pobre predictor de incidencias de tráfico. Además, al igual que la mayoría de los casos de estudios, este precede de problemas sobre la eficiencia de los datos. Los investigadores utilizaron un programa para equilibrar el conjunto de datos de entrenamiento, supliendo la problemática de los datos, donde las redes neuronales obtuvieron mejores resultados, sin embargo, la red neuronal propuesta no fue la más efectiva [6].

Unas de las problemáticas que se han presentado en el análisis de datos, es la heterogeneidad de los datos de accidentes de tráfico. Kumar y Toshniwal [24] propusieron

que la agrupación antes del análisis es conveniente cuando se tiene heterogeneidad en los datos de accidentes. Por su parte, Ona et al. [25] utilizaron la técnica de agrupamiento de clases latentes (LCC) para eliminar la heterogeneidad de los datos en la que proporciona diferentes criterios de selección de clúster para ser utilizado en la identificación del número de grupos presentes en el conjunto de datos. En la tabla 2 se muestra un análisis comparativo de las diferentes técnicas utilizadas para el análisis de accidentes de tránsito, donde se visualiza que árboles de decisión es la técnica más utilizada, seguida de las técnicas de SVM Y ANN. Las reglas de decisión y el proceso analítico jerárquico (AHP) fueron las técnicas menos utilizadas y con menor rendimiento. Como en el caso de estudio realizado por Gupta, Mohammad y Halgamuge [26], en el que el algoritmo OneR mostró el menor número de incidencias al ser comparado con algoritmos como Naive Bayes y J48 de árboles de decisión. Entre los algoritmos más utilizados, SVM es una en las que se ha obtenido menor precisión en comparación con árboles de decisión y las redes neuronales. Ésta última supera a las demás técnicas en la predicción de los accidentes de tráfico con mayor precisión [6], [13], [16].

Tabla 1-2. Técnicas de análisis de datos utilizadas en estudio de accidentes de tránsito

Autor \ Técnica	Arboles de decisión	Arboles de clasificación	Arboles de regresión	Adaboost	Clustering	Máquinas de vector de soporte	Naive Bayes	Procesos analíticos jerárquico	Random Forest	Redes neuronales recurrentes	Redes neuronales artificiales	Reglas de asociación	Reglas de decisión
Zhang, He, Gao & Ni (2018) [12]						X				X			
Novkovic & Arsenovic M.(2017) [1]	X			X		X			X				
Grimm, ,Fristo & Amin-Naseri (2017)[6]						X					X		
Sameen & Pradhan (2017)[15]										X			
Tiwari, Prayag (2017) [13]	X				X	X	X						
Xi, Zhao, Li & Wang (2016) [5]								X				X	
Jain, Ahuja, Anuranjana, & Mehrotra (2016)[27]	X												
Gupta, Mohammad, Syed & Halgamuge (2016) [26]	X						X						X
Kumar & Toshniwal (2015) [24]					X							X	
Olutayo & Eludire (2014) [17]	X										X		
Martín, Baena, Garach, López & de Oña (2014)[11]	X											X	
T. Akomolafe & Olutayo, (2013) [27]	X												

Autor \ Técnica	Arboles de decisión	Arboles de clasificación	Arboles de regresión	Adaboost	Clustering	Máquinas de vector de soporte	Naive Bayes	Procesos analíticos jerárquico	Random Forest	Redes neuronales recurrentes	Redes neuronales artificiales	Reglas de asociación	Reglas de decisión
Shanthi & Geetha (2012)[12]	X								X		X		
Krishnaveni & Hemalatha, (2011)[28]	X			X			X		X				
Kashani, Mohaymany &,Ranjbari (2011) [18]	X		X										
Beshah & Hill, (2010) [29]	X		X										
Guan, Liu, Yin & Zhang, (2010) [30]											X		
Chaozhong, Hu, Ming, Xinping (2009) [19]									X				
Tseng, Nguyen, Liebowitz & Agresti, (2005) [22]	X									X			
Chong & Abraham, (2004)[20]	X					X					X		

Fuente: elaboración propia con base en las referencias mencionadas en la tabla.

Hasta ahora, muchos académicos analizaron y exploraron las características y causalidades de los accidentes de tráfico como se mencionó anteriormente [1], [6], [11], [27]. Sin embargo, las características no eran las mismas en las diferentes áreas, debido a la diferencia de las condiciones del tráfico y el entorno que estos se mantienen. También se encuentra la problemática de los pocos estudios [6], [8], [11] en el que contemplan los eventos públicos como un factor externo influyente en los accidentes de tránsito.

En el análisis de los datos, algunos investigadores obtuvieron información por medio de Secretarías Distritales de Movilidad o instituciones de Dirección de Seguridad Pública [7], [8]. En otras investigaciones [6], [21], [27] aprovechan los datos de Crowdsourced, como de páginas web y de redes sociales, tanto para los datos de los eventos públicos como de los accidentes de tránsito. Para el caso de páginas web, se han utilizado programas y técnicas de “Scraping Web”, que básicamente hace referencia al raspado de información de sitios web públicos.

La disparidad de los estudios nombrados en los antecedentes, parte como gran influyente el causal o la caracterización de los accidentes. A medida en la que se tiene el contexto de las investigaciones, algunas desconocen las diferentes variables que puede intervenir

parcial o globalmente en la generación de los accidentes de tránsito. Al igual los autores emplean en su mayoría métodos de machine learning para caracterizar y modelar el comportamiento de estos eventos y posibles variables influyentes. Como aporte al estudio de los antecedentes, es identifica las diferentes falencias y problemáticas como factor común. Con el objetivo de abordarlas a razón de no repetirlas y dar otras alternativas al empleo para el análisis de datos.

2. Marco teórico

En este capítulo se presenta los conceptos en los que incorpora la extracción de datos. Se describe la técnica de *Web Scraping* al igual el proceso que se lleva a cabo para las peticiones y mapeos de páginas mediante Python. También se nombra lo correspondiente a la minería de datos espaciales, sus principales tareas y componentes los cuales hacen uso para el análisis de datos espaciales, como los sistemas de información geográfica y base de datos espaciales.

2.1 Extracción de datos

2.1.1 Web Scraping

Web Scraping o raspado web es una técnica para extraer o recopilar información en grandes cantidades de un sitio web mediante programas o librerías. En el desarrollo de raspado web se transforma los datos sin estructura en datos estructurados, dado que estos se encuentran en formato HTML. Por medio de peticiones HTTP al servidor que aloja las páginas web estáticas y dinámicas del contenido de la página, se obtiene la información,

que luego es analizadopara extraer y transformar el contenido de estas páginas web en datos estructurados. Algunas aplicaciones como DataToolBar, OutWit Hub, FMiner brindan interfaces, las cuales permiten seleccionar los campos que se quieren obtener, reconociendo automáticamente la estructura de la página [31][32].

2.1.2 Web Scraping con Python

Web Scraping se utiliza para obtener grandes cantidades de información de páginas web. Algunas de las aplicaciones de raspado web:

- Comparación de precios: servicios como ParseHub utilizan *Web Scraping* para recopilar datos de compras en línea en diferentes páginas web y lo luego comparar los precios de los productos.
- Extracción de direcciones de correo electrónico: muchas empresas que utilizan el correo electrónico como medio para la comercialización, el cual obtienen correos electrónicos y luego realizar envíos de correos masivos con información comercial.
- *Web Scraping* en redes sociales: se utiliza para recopilar datos de redes sociales como Twitter para descubrir las tendencias.
- Investigación y desarrollo: se utiliza para obtener conjunto de datos (estadísticas, información general, temperatura, etc.) de las páginas web, para el análisis y realización de encuestas.
- Ofertas laborales: se extrae información de ofertas de trabajo, entrevistas de diferentes páginas web y luego se exponen en un solo sitio para que el usuario pueda acceder fácilmente.

Con Python la realización de *Web Scraping* es simple de codificar. Python trae algunas bibliotecas como *Selenium*, *BeautifulSoup* y pandas para la extracción de información en páginas web. Cuando se realiza la ejecución del código para el web Scraping, se envía una solicitud a la URL objetivo. Como respuesta a la solicitud, el servidor envía los datos y le permite leer la página HTML. Luego, el código analiza la página HTML donde encuentra los datos y los extrae. *BeautifulSoup* ofrece una serie de módulos como "*request*" y *"Beautiful Soup"* para dicho fin. El módulo `urllib.request` se utiliza para abrir

direcciones URL, mientras el paquete "*Beautiful Soup"* se utiliza para extraer datos de archivos HTML. Para obtener el contenido HTML de la página web, se crea un objeto *BeautifulSoup* desde el HTML. Este proceso se usa para analizar el HTML en un formato texto y luego dividirlo en objetos de Python [33]. Hacer Scraping de los elementos HTML usando la clase `Attribute` siendo selectivos en algunos elementos HTML basados en sus clases CSS, permite la extracción de información selectiva. El objeto *BeautifulSoup* tiene una función llamada `findAll` que filtra elementos en función de sus atributos, dado que, el objeto. *BeautifulSoup* transforma el HTML en una estructura jerárquica.

2.2 Capa de Datos

Las bases de datos pueden ser de diferentes ámbitos, una de ellas son las bases de datos espaciales, donde se almacenan objetos definidos en el espacio geográfico y representan objetos geométricos como puntos, líneas y polígonos. A comparación de las bases de datos tradicionales desarrolladas para administrar varios tipos de datos, las bases de datos adquieren una funcionalidad adicional para procesar estos tipos de datos, el cual implementan datos y características geométricas [34].

Los sistemas de bases de datos tradicionales utilizan índices para la optimización de consultas, pero a medida en que se tiene datos geográficos en las bases de datos, estos índices no son óptimos para la realización de consultas espaciales. A lo que da paso a las bases de datos espaciales que implementa índices geométricos para las operaciones y consultas espaciales. Adicional de las consultas SQL tradicionales, las bases de datos espaciales contemplan una variedad de operaciones espaciales [35].

- Mediciones espaciales: cálculos de longitudes, distancias y áreas entre puntos, líneas y geometrías.
- Funciones espaciales: operaciones entre entidades u objetos y la generación de nuevos objetos, a partir de intersecciones, diferencias, combinaciones, superposiciones, entre otras.
- Predicados espaciales: consultas sobre relaciones espaciales entre geometrías.

- Constructores de geometría: Creación de nuevas geometrías.

Las bases de datos espaciales trabajan bajo los siguientes objetos.

- Data set: especificaciones para una clase característica, catálogo ráster o una tabla de atributos.
- Dominios: características que determinan rango de valores válidos para atributos.
- Relaciones: son las relaciones explícitas entre atributos; propiedad de las bases de datos, que definen cómo las columnas se relacionan con columnas en otra tabla.
- Reglas: Indican cómo objetos se relacionan geométricamente con otros objetos.
- Capas de mapas: especificaciones que determinan la representación de los objetos datos.

2.2.1 Base de Datos PostgreSQL y PostGIS

PostGIS es una extensión de PostgreSQL, el cual contiene las características de las bases de datos empresariales, convirtiendo el sistema de administración de bases de datos PostgreSQL en una base de datos espacial. El módulo PostGIS añade soporte de objetos geográficos a la base de datos como índices y datos especiales, permitiendo realizar algunas operaciones y funciones sobre ellos. La finalidad de tener una base de datos espacial es la utilidad de un sistema de información geográfica [36], [37].

Mediante el Script "*create extension postgis*"; PostGIS instala una serie de funciones, tablas y vistas. La tabla *spatial_ref_sys* es utilizada por PostGIS para convertir entre sistemas de referencia espaciales diferentes. La tabla *spatial_ref_sys* almacena la información en los sistemas de referencia espacial válida [37].

PostGIS soporta varios tipos de datos espaciales:

- Geométrico: tipo de datos que almacena datos como vectores dibujados sobre una superficie plana.
- Geográfico: tipo de datos que almacena datos como vectores dibujados sobre una superficie esferoidal.

- Ráster: tipo de datos que almacena los datos como una matriz n-dimensional que representa cada posición (píxel) un área de espacio y cada banda (dimensión) tiene un valor para cada espacio de píxeles.

Las vistas `geometry_columns, geography_columns y raster_columns` tienen el trabajo de decirle a PostGIS qué tablas tienen geometría PostGIS, geografía y columnas de ráster. [38].

2.3 Minería de datos espaciales

La información espacial está relacionada con la cartografía, en cuanto a operaciones de gestión y análisis de datos espaciales, donde son representados en diferentes tipos de mapas y símbolos [39]. La creación de base de datos espaciales, paralelo a la minería de datos espaciales, surgen a partir de las diferentes áreas enlazadas a la información geográfica y en los avances que se han tenido en el procesamiento de información espacial, junto con sistemas y base de datos espaciales. Estos avances dan paso a los Sistemas de Información Geográfica, SIG, que constituyen una herramienta de visualización, consulta, edición y análisis espacial. Los SIG proveen sistemas para la gestión de datos georreferenciados, con una serie de componentes en el campo espacial, temática y temporal de los datos. También proveen una capa para el manejo de la geometría asociada a los datos los cuales son almacenados en una Base de Datos Espacial, SDB [40], [41].

La minería de datos espaciales SDM surge como el proceso de análisis automático. Con el objetivo de crear conocimiento inherente a la naturaleza de los datos espaciales, [42], [43]. La minería de datos espaciales se considera una rama de la minería de datos tradicional [44]. En La aplicación de métodos y algoritmos a datos geográficos, en la deducción de patrones o categorizaciones de manera asociativa, en forma de estructuras y agrupaciones de las diversas relaciones espaciales [45] o por medio de implementación de algoritmos [46].

2.3.1 Tareas en la minería de datos espaciales

La Minería de datos espaciales cubre cuatro trabajos importantes y para la ejecución de estas existen diversos métodos que vienen del área computacional, estadística y visual, también combinaciones de los métodos [44]. Algunas tareas que se hacen en la minería de datos son la clasificación, asociación y la agrupación espacial, al igual la geovisualización de estas.

2.3.1.1 Clasificación espacial y predicción

La clasificación de datos refiere a la agrupación de datos puntuales en categorías, de acuerdo con los atributos de los objetos espaciales; este tipo de clasificación también se denominan “clasificación supervisada”. Máxime cuando se necesita datos de entrenamiento para realizar el modelo de clasificación, grupo de datos para la validación y optimización de la estructura y otro grupo de datos para evaluar el desempeño del modelo del método a utilizar [47]. Dichos métodos están compuestos por árboles de decisión, redes neuronales, estimación de máxima verosimilitud análisis de discriminante lineal, máquinas de soporte vectorial, k-NN (K Nearest Neighbors) y razonamiento basado en casos entre otros [47], [48].

2.3.1.2 Reglas de asociación espacial

Comprende las relaciones que se encuentran entre los objetos y los predicados, tarea que identifica las regularidades entre los objetos espaciales. Las reglas de asociación también pueden identificar relaciones topológicas entre objetos espaciales como disyunción, intersección, adyacencia, sobreposición, vecindad e igualdad, entre otras [40]. En la **Figura 2-1.** Operaciones comunes entre entidades espaciales. se ilustra las operaciones entre dos entidades y el resultado que se obtiene.

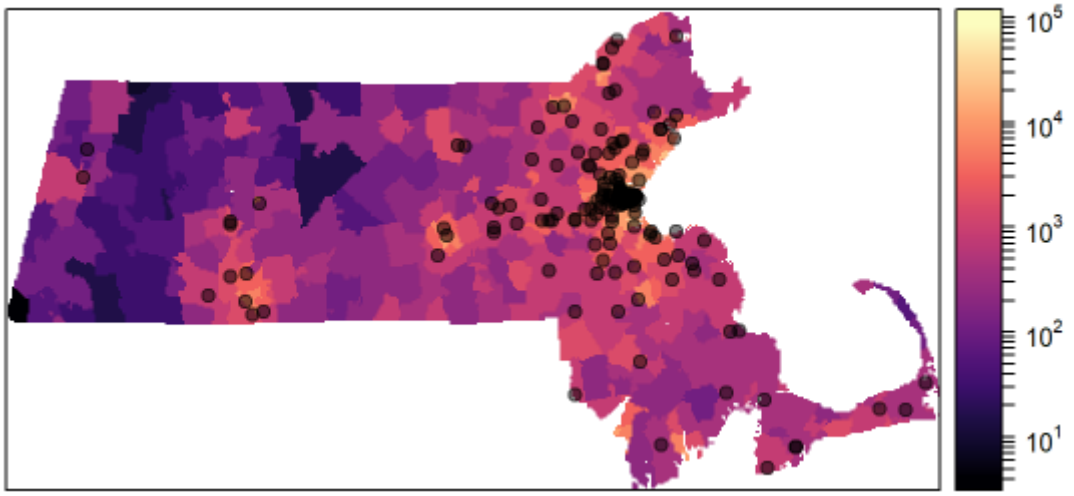
Figura 2-1. Operaciones comunes entre entidades espaciales.

Fuente: http://desktop.arcgis.com/es/arcmap/10.3/analyze/commonly-used-tools/GUID-EE844BCC9721-web.gif.

2.3.1.3 Agrupaciones espaciales y análisis de patrones de punto

Estas técnicas han sido utilizadas para el analizar un conjunto de datos y realizar el procedimiento de agrupamiento por similitud o distancia que mantienen estos objetos. Según como se necesiten para realizar su debido proceso y metodología, se tiene el agrupamiento por separación. El agrupamiento consiste en la formación de grupos o conglomerados de objetos permitiendo una autocorrelación espacial, como se muestra en la **Figura 2-2**. Estos datos son organizados según su cercanía u homogeneidad de sus atributos [43].

Figura 2-2. Presencia de eventos y áreas de concentración



Fuente: https://mgimond.github.io/Spatial/11-Point-Patterns_files/figure-html/f11-starbucks-1.png.

Por otra parte, la agrupación jerárquica realiza la agrupación de clústeres para dar paso a uno nuevo o separar uno ya existente, generando una secuencia de particiones, dando paso al concepto de regionalización, que es una manera de organizar los objetos espaciales en conjuntos de objetos cercanos o contiguos [46].

El análisis de patrones de punto o de punto caliente (hot spot), se enfatiza en las concentraciones de objetos en un área determinada, dando a conocer la dispersión o superposición de puntos geográficos en el espacio de estudio. Esto indica el punto más alto donde se concentra de diferentes acciones, por ejemplo, accidentes de tránsito. El objetivo del análisis de hot spot es destacar los puntos en que se presentan más estas situaciones [45].

2.3.1.4 Geovisualización

La visualización es el primer paso para cualquier análisis de datos referenciados espacialmente. Existen muchas decisiones a tomar cuando se visualizan los valores ya que todo varía según la interpretación del usuario y las formas de que el análisis impacte en el resultado. Los patrones de puntos son a menudo utilizados para representar

localizaciones de eventos y elementos concentrados. En los casos de objetos con atributos categóricos (e.g. la población en zonas urbanas o zonas rurales), la visualización se basará en una selección de colores o sombras para representar cada categoría. En ese caso, la selección de colores o sombras posibilitan la diferenciación entre categorías importantes o de relevancia para el análisis [39].

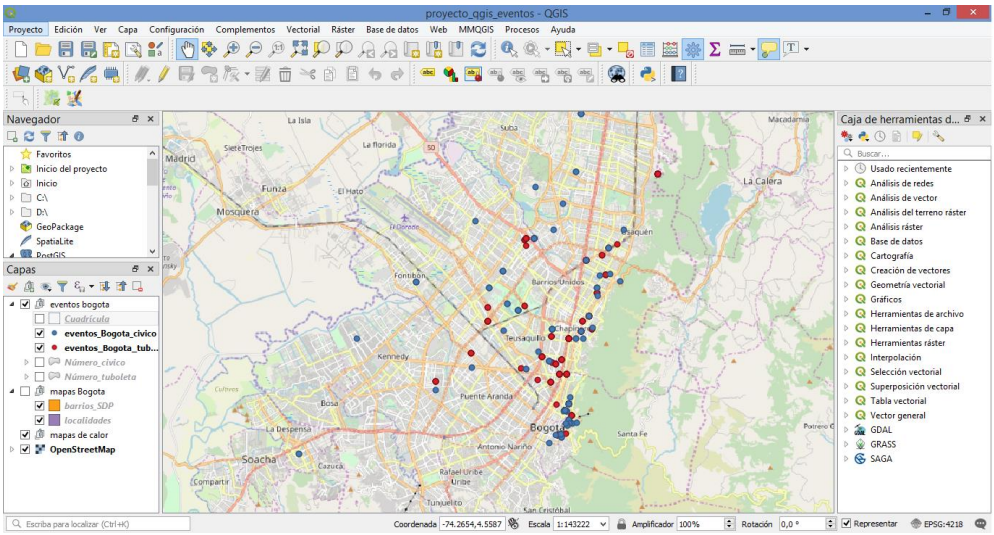
2.4 Sistemas de Información Geográfica

Los sistemas de información geográfica o GIS por sus siglas en ingles son Herramientas que permiten visualizar, manipular e interpretar características para entender las relaciones y patrones que existen entre elementos de una base de datos. Todas estas características están referenciadas espacialmente y son tabuladas para administrar y resolver distintas tareas.

2.4.1 QGIS

QGIS es un Sistema de Información Geográfica (GIS por sus siglas en ingles) de Código Abierto licenciado bajo GNU - General Public License. QGIS es un proyecto oficial de Open Source Geospatial Foundation (OSGeo). Este programa permite visualizar, gestionar, editar y analizar datos, además de diseñar mapas imprimibles. QGIS proporciona una gama creciente de herramientas por el motivo de ser código abierto, proporcionando así la capacidad de que el usuario cree sus propios complementos.

Figura 2-3. Ventana de inicio del QGIS



Fuente: elaboración propia.

QGIS se compone por una serie de elementos como el Panel de exploración, Barras de herramientas, Lienzo del mapa, Barra de estado y una barra de herramientas lateral como se identifica en la **Figura 2-3**. En la lista de capas puede ver todas las capas que están disponibles. El explorador de QGIS es un panel que le permite navegar fácilmente por la base de datos espaciales como PostGIS, Oracle, Spatialite, MYSQL Spatial y conexiones WMS/WFS. La barra de herramientas muestra los accesos básicos. El lienzo y barra de estado del mapa muestra la información y el mapa actual.

3. Web Scraping en eventos públicos

En este capítulo se procede a describir el proceso y métodos para la realización de *Web Scraping* de la información de los eventos públicos. Se especifica las fuentes de información el cual se realiza el Scraping de los datos, las variables a extraer y los tipos de datos. Al igual se explica la técnica creada para la extracción de datos, haciendo relación de los pasos y módulos que este comprende. Finalmente se presenta el proceso de creación de la base de datos con los datos extraídos de los eventos públicos y los de accidentes de tránsito los cuales ya se cuenta.

Para la extracción de la información se seleccionaron las páginas:

- Cívico (https://www.civico.com/bogota).
- Ticket Express (https://ticketexpress.com.co/).
- TuBoleta (https://vive.tuboleta.com/),

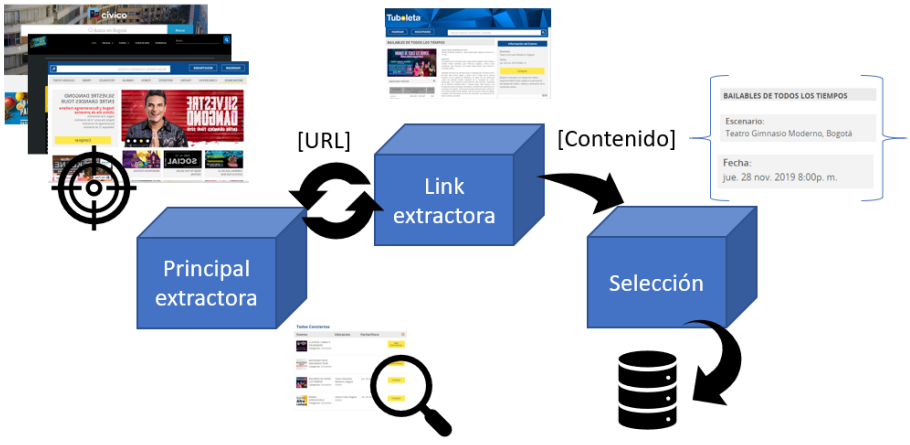
Como parte del proceso de selección, se tuvo en cuenta la información que se tiene en estas páginas web y en principio por publicar información sobre eventos públicos mayormente de la ciudad de Bogotá. Cada Evento contiene características como el lugar, nombre, fecha, hora y dirección.

Figura 3-1. Campos de extracción de datos

Fuente: https://vive.tuboleta.com/.

El método diseñado para la extracción de información web está dividido en tres módulos principales, principal extractora, link extractora y selección. Los módulos contemplan el procesamiento de la página principal y subpáginas para la extracción de información de eventos públicos. El enfoque de filtrado de enlaces y extracción de información se ilustra en la **Figura 3-2**.

Figura 3-2. Módulos del método de Web Scraping.



Fuente: elaboración propia.

Las partes principales del método de *Web Scraping* son:

`Principal extractora`. Este módulo está dedicado a encontrar, dada una página principal, páginas nombradas anteriormente, las páginas segundarias, adecuadas para ser los enlaces entrantes al siguiente módulo. Es decir, todas las páginas que enlazan con la página principal. Link extractora recoge los enlaces entrantes de la página principal, para la extracción de los enlaces de cada uno de los eventos que estas contienen. Las subpáginas normalmente agrupan los eventos por cierta categoría.

`Link extractora` da como salida la lista de las páginas de los eventos extraídos que serán procesados por el módulo de selección para la extracción de información relevante de los eventos públicos.

El módulo de `selección` está dirigido a la extracción de los campos y etiquetas de las páginas de los eventos públicos de forma individual. Proceso el cual se depende del *Scraping Web*, este hace referencia a un conjunto de técnicas utilizadas para obtener automáticamente información de una página web y no de forma manual. En particular, se adopta la técnica de Scraping para el acceso a información pertenecientes a secciones u objetos; buscar nombres de etiquetas, contenido o atributos que coincidan con criterios de selección; y acceder a los atributos por medio de id de referencia. Raspado se realiza mediante el uso de bibliotecas de Python proporcionados por HTMLParser y Beautiful Soap.

El módulo `selección` da como salida como un conjunto de datos ordenados, perteneciente a los atributos de los eventos extraídos junto con la URL correspondiente y sus descripciones. Este conjunto de datos será almacenado en una base de datos espacial para el procesamiento, limpieza y estandarización de datos.

3.1 Proceso de extracción.

Con base en los módulos que conforman la arquitectura del método de *Scraping Web*, la arquitectura comprende una serie de pasos, que llevan a la extracción de información de eventos públicos. El proceso de extracción inicia con la petición a la página principal y finaliza con la descarga de la información de cada uno de los eventos encontrados en las subpáginas de la página web principal.

- Petición a URL de la página web de los eventos públicos.
- Descarga del código fuente de la página web.
- Mapeo del código fuente a código HTML.
- Recorrido de código y obtención URL de las diferentes páginas por categorías de los eventos públicos.
- Petición y descarga de código fuente de cada una de las URLs obtenidas de las categorías.
- Mapeo del código fuente a código HTML de cada URL.
- Obtener valores de los atributos (Nombre, categoría, lugar, fecha, hora) de cada evento público presente en las categorías.
- Filtrado de la información, eliminación de espacios y caracteres especiales.
- Obtención de la dirección, latitud y longitud por medio del lugar de cada evento.
- Descarga de archivo en formato de archivo plano (CSV) con los datos obtenidos.

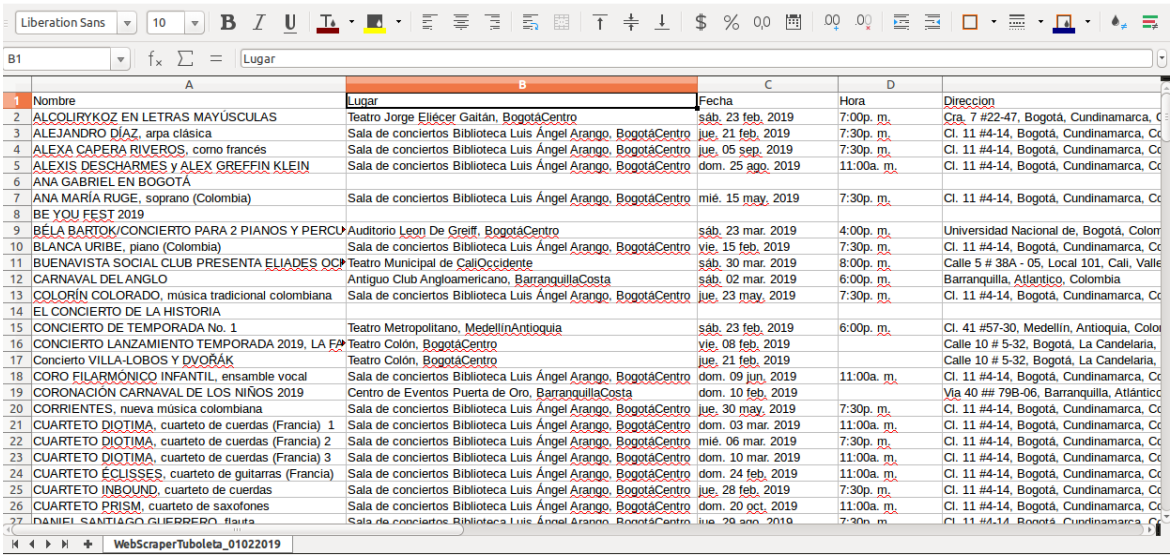
La ejecución de proceso de extracción de información se realiza diariamente, con el fin de capturar los eventos con múltiples presentaciones, abarcando eventos públicos con presentaciones diarias y con ciertos períodos característicos.

Alineado con el proceso de extracción de datos y los módulos propuestos, se integra con la consulta de las ubicaciones geográficas de los eventos. El método contempla la consulta y extracción de las longitudes y latitudes de los eventos públicos, según la dirección capturada de cada evento. En el **Anexo 1** la lógica con la cual es desarrollado el método.

3.1.1 Descarga de información de eventos públicos.

Posterior al proceso de extracción de la información de los eventos públicos, los datos son guardados en archivos con formato CSV, en las que las columnas se separan por comas y las filas por salto de línea. Al realizarse la extracción diaria los archivos se denominan con el nombre de la página de extracción más la fecha del día. Como consiguiente a la descarga de los archivos, se almacena la información en una base de datos espacial PostgreSQL. En la **Figura 3-3**, se ilustra una sección de datos extraídos de la página de TuBoleta.

Figura 3-3. Datos Extraídos de la página web TuBoleta.



	A	B	C	D	
1	Nombre	Lugar	Fecha	Hora	Direccion
2	ALCOLIRYKOZ EN LETRAS MAYÚSCULAS	Teatro Jorge Eliécer Gaitán, BogotáCentro	sáb. 23 feb. 2019	7:00p. m.	Cra. 7 #22-47, Bogotá, Cundinamarca, C
3	ALEJANDRO DÍAZ, arpa clásica	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 21 feb. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
4	ALEXA CAPERA RIVEROS, corno francés	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 05 sep. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
5	ALEXIS DESCHARMES y ALEX GREFFIN KLEIN	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 25 ago. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
6	ANA GABRIEL EN BOGOTÁ				
7	ANA MARÍA RUGE, soprano (Colombia)	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	mié. 15 may. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
8	BE YOU FEST 2019				
9	BÉLA BARTOK/CONCIERTO PARA 2 PIANOS Y PERCU	Auditorio Leon De Greiff, BogotáCentro	sáb. 23 mar. 2019	4:00p. m.	Universidad Nacional de, Bogotá, Colom
10	BLANCA URIBE, piano (Colombia)	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	vie. 15 feb. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
11	BUENAVISTA SOCIAL CLUB PRESENTA ELIADES OCH	Teatro Municipal de CaliOccidente	sáb. 30 mar. 2019	8:00p. m.	Calle 5 # 38A - 05, Local 101, Cali, Valle
12	CARNAVAL DEL ANGLO	Antiguo Club Angloamericano, BarranquillaCosta	sáb. 02 mar. 2019	6:00p. m.	Barranquilla, Atlantico, Colombia
13	COLORÍN COLORADO, música tradicional colombiana	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 23 may. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
14	EL CONCIERTO DE LA HISTORIA				
15	CONCIERTO DE TEMPORADA No. 1	Teatro Metropolitano, MedellínAntioquia	sáb. 23 feb. 2019	6:00p. m.	Cl. 41 #57-30, Medellín, Antioquia, Colo
16	CONCIERTO LANZAMIENTO TEMPORADA 2019, LA FA	Teatro Colón, BogotáCentro	vie. 08 feb. 2019		Calle 10 # 5-32, Bogotá, La Candelaria,
17	Concierto VILLA-LOBOS Y DVOŘÁK	Teatro Colón, BogotáCentro	jue. 21 feb. 2019		Calle 10 # 5-32, Bogotá, La Candelaria,
18	CORO FILARMÓNICO INFANTIL, ensamble vocal	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 09 jun. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
19	CORONACIÓN CARNAVAL DE LOS NIÑOS 2019	Centro de Eventos Puerta de Oro, BarranquillaCosta	dom. 10 feb. 2019		Via 40 ## 79B-06, Barranquilla, Atlántico
20	CORRIENTES, nueva música colombiana	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 30 may. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
21	CUARTETO DIOTIMA, cuarteto de cuerdas (Francia) 1	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 03 mar. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
22	CUARTETO DIOTIMA, cuarteto de cuerdas (Francia) 2	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	mié. 06 mar. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
23	CUARTETO DIOTIMA, cuarteto de cuerdas (Francia) 3	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 10 mar. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
24	CUARTETO ÉCLISSES, cuarteto de guitarras (Francia)	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 24 feb. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
25	CUARTETO INBOUND, cuarteto de cuerdas	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 28 feb. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
26	CUARTETO PRISM, cuarteto de saxofones	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	dom. 20 oct. 2019	11:00a. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co
27	DANIEL SANTIAGO GUERRERO, flauta	Sala de conciertos Biblioteca Luis Ángel Arango, BogotáCentro	jue. 29 ago. 2019	7:30p. m.	Cl. 11 #4-14, Bogotá, Cundinamarca, Co

WebScraperTuboleta_01022019

Fuente: elaboración propia con base a los datos extraídos.

3.2 Creación de una base de datos espacial

3.2.1 Creación de tablas

Partiendo de una base de datos para mantener el conjunto de datos. Se crea una tabla basada en el esquema de la información. Posterior a la creación y carga de datos, se convertirá en una tabla con datos espaciales. Para el estudio se crean las tablas EVENTOS_PUBLICOS, ACCIDENTES_TRANSITO y LUGARES, esta última hace

referencia a la ubicación de los lugares o establecimientos donde se presenta los eventos públicos. Por medio de la Siguiente sentencia se realiza la creación de la tabla de eventos públicos.

```
CREATE TABLE EVENTOS_PUBLICOS
    (
        NOMBRE character varying(100),
        CATEGORIA character varying(60),
        LUGAR character varying(100),
        FECHA character varying(30),
        HORA character varying(15),
        LATITUD numeric,
        LONGITUD numeric,
        DIRECCION character varying(100),
        GEOM geometry(POINT,4218)
    );
```

PostGIS utiliza un campo de geometría especial en el conjunto de datos que codifica la geometría en una cadena de caracteres variable grande. Estos son interpretados por la base de datos como la geometría. Antes de visualizar el conjunto de datos, se debe actualizar el campo Geom utilizando una declaración PostGIS que toma las columnas de latitud y longitud y las interpreta en una cadena codificada en la columna de Geom. Para el tipo de datos, se declara de tipo de geometría (PUNTO, 4218). Los argumentos son el tipo de geometría (PUNTO) y el sistema de coordenadas (código EPSG , que para WGS84 es 4218).

```
UPDATE eventos_publicos SET geom = ST_SetSRID
        (ST_MakePoint(longitud,latitud),4218);
```

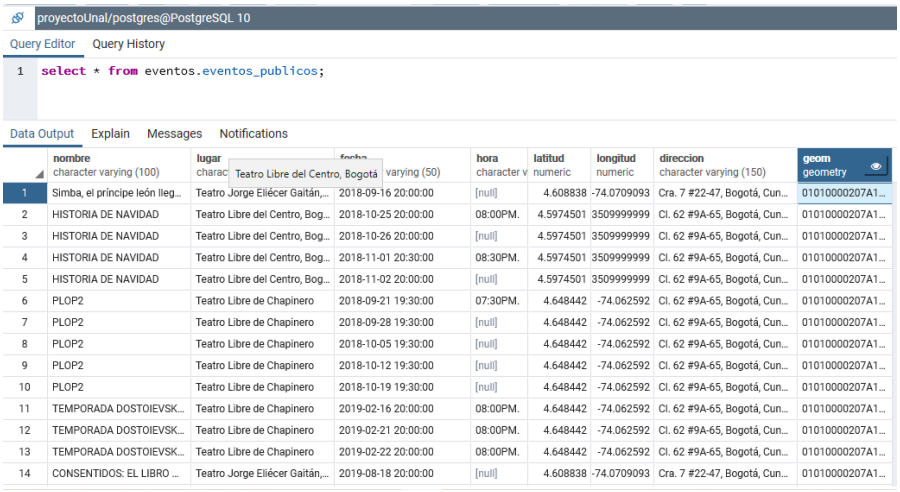
3.2.2 Estandarización de datos

En la importación y fusión de la información de las diferentes paginas extraídas, los datos no tienen el mismo formato. Para las fechas, es necesario aplicar un formato estándar, como requisito en el análisis espacio temporal. El proceso se realiza en SQL por medio de funciones básicas de extracción y remplazo de caracteres. En primera instancia se da formato al campo hora con el sistema horario de 12 horas (hh:mm AM|PM) ver **Anexo 3.** Script de estandarización de hora.

En el campo fecha se remplaza valores alfabéticos por numéricos, para el caso de los meses y días de la semana. Adicional se obtiene la hora anteriormente para estandarizar la fecha con un formato (AAAA-MM-DD HH24:MM:SS) ver **Anexo 2.** Script de estandarización de fecha.

Como afinamiento de los datos, donde también se realiza la actualización de los campos de lugar, dirección, latitud, longitud y dirección del evento para mantener la homogeneidad en los valores. Es decir, al tener diferentes nombres que hacían referencia a un mismo lugar, se realizaba la actualización de los campos dejando el mismo valor para todos. En la **Figura 3-4** se muestra un conjunto de datos procesados en base de datos.

Figura 3-4. Datos de eventos públicos procesados y almacenados en base de datos



	nombre character varying (100)	lugar	fecha varying (50)	hora	latitud numeric	longitud numeric	direccion character varying (150)	geom geometry
1	Simba, el príncipe león lleg...	Teatro Jorge Eliécer Gaitán,...	2018-09-16 20:00:00	[null]	4.608838	-74.0709093	Cra. 7 #22-47, Bogotá, Cun...	01010000207A1...
2	HISTORIA DE NAVIDAD	Teatro Libre del Centro, Bog...	2018-10-25 20:00:00	08:00PM.	4.5974501	3509999999	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
3	HISTORIA DE NAVIDAD	Teatro Libre del Centro, Bog...	2018-10-26 20:00:00	[null]	4.5974501	3509999999	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
4	HISTORIA DE NAVIDAD	Teatro Libre del Centro, Bog...	2018-11-01 20:30:00	08:30PM.	4.5974501	3509999999	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
5	HISTORIA DE NAVIDAD	Teatro Libre del Centro, Bog...	2018-11-02 20:00:00	[null]	4.5974501	3509999999	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
6	PLOP2	Teatro Libre de Chapinero	2018-09-21 19:30:00	07:30PM.	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
7	PLOP2	Teatro Libre de Chapinero	2018-09-28 19:30:00	[null]	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
8	PLOP2	Teatro Libre de Chapinero	2018-10-05 19:30:00	[null]	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
9	PLOP2	Teatro Libre de Chapinero	2018-10-12 19:30:00	[null]	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
10	PLOP2	Teatro Libre de Chapinero	2018-10-19 19:30:00	[null]	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
11	TEMPORADA DOSTOIEVSK...	Teatro Libre de Chapinero	2019-02-16 20:00:00	08:00PM.	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
12	TEMPORADA DOSTOIEVSK...	Teatro Libre de Chapinero	2019-02-21 20:00:00	08:00PM.	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
13	TEMPORADA DOSTOIEVSK...	Teatro Libre de Chapinero	2019-02-22 20:00:00	08:00PM.	4.648442	-74.062592	Cl. 62 #9A-65, Bogotá, Cun...	01010000207A1...
14	CONSENTIDOS: EL LIBRO ...	Teatro Jorge Eliécer Gaitán,...	2019-08-18 20:00:00	[null]	4.608838	-74.0709093	Cra. 7 #22-47, Bogotá, Cun...	01010000207A1...

Fuente: elaboración propia.

A partir del procesamiento y limpieza de los datos, se procede a la filtración y eliminación de duplicados, Para tener como resultado 3988 eventos extraídos durante el periodo de septiembre 2018 y julio de 2019. Se resalta que para el caso de los accidentes vehiculares solo fue necesario importarlos a la base de datos.

4. Análisis espacio-temporal

Este capítulo aborda los procesos y técnicas utilizadas en el análisis de datos a nivel espacial y temporal. Esto con el objetivo de determinar el grado de asociación entre la ocurrencia de eventos públicos y accidentes de tránsito en la ciudad de Bogotá. Con la información recogida de los eventos públicos se identificaron 40 lugares de ocurrencia. Los eventos son abarcados en lugares como centros comerciales, teatros, bares, restaurantes, parques y algunos espacios de gran aforo para presentaciones y festivales. En la **Tabla 4-1**, se muestran los lugares obtenidos de los eventos públicos con su respectiva dirección y aforo.

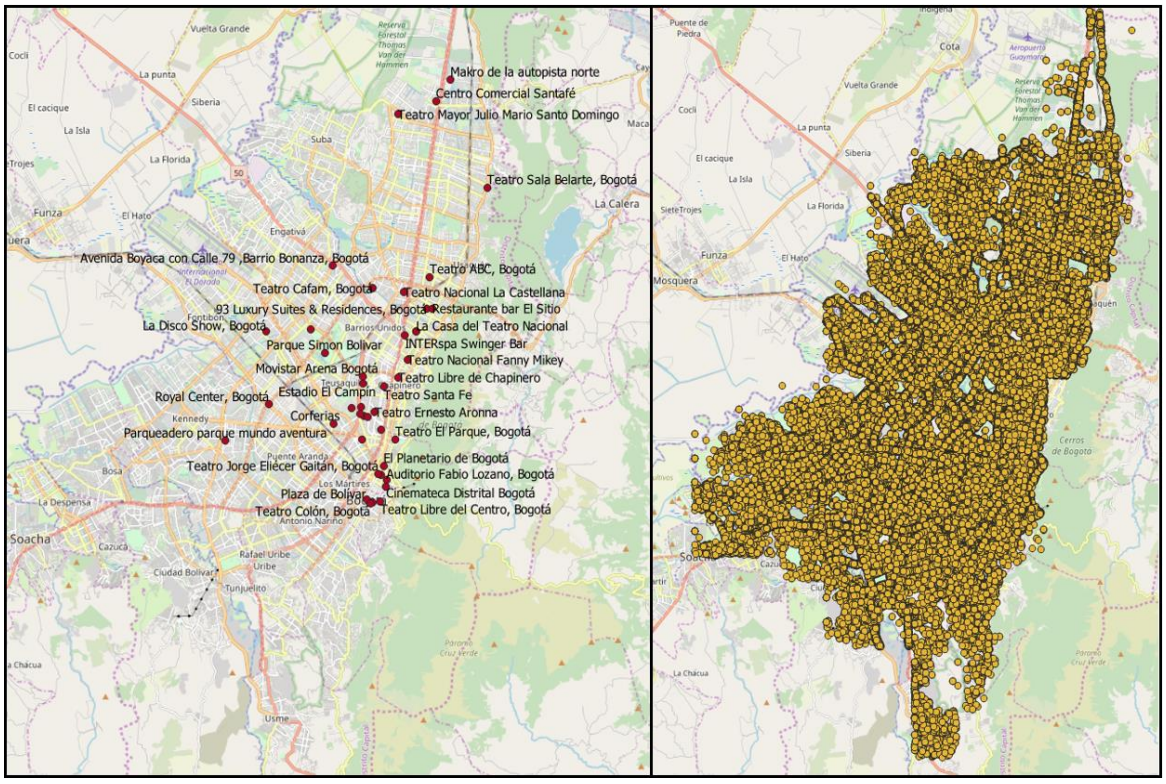
Tabla 4-1. Lugares de ocurrencia de eventos públicos.

Lugar	*Dirección*	*Aforo*
93 Luxury Suites & Residences	Cra. 13a #93-51, Bogotá	200
Auditorio Fabio Lozano	Cra. 4 a 22-98, Bogotá	592
Auditorio León de Greiff	Universidad Nacional de, Bogotá	1619
Auditorio Mayor Cun	Cl. 23 #6 - 19, Bogotá	1320
Avenida Boyacá Calle 79	Cl. 79 & Av. Boyacá, Bogotá	1200
Casa E	Ak. 24 # 41-69, Bogotá	400
Centro Comercial Santafé	Avenida Carrera 45 #185, Bogotá	800
Cinemateca Distrital	Cra. 3 #19-10, Bogotá	272
Corferias	Cra. 37 #24-67, Bogotá	6000
El Planetario de Bogotá	Cl. 26b #5-93, Bogotá	376
Estadio El Campín	Carrera 30 y Calle 57, Bogotá	36343
INTERspa Swinger Bar	Cra. 18 #78 50, Bogotá	300
Jardín Botánico José Celestino Mutis	Cl. 63 #6895, Bogotá	2700
La Casa del Teatro Nacional	Cra. 20 #37-54, Bogotá	160
La Casa en el Aire	Cra. 13 #82-39, Bogotá	100
La Disco Show	Cl. 24c ##75 - 43, Bogotá	300
Makro de la autopista norte	Carrera 45 # 192 - 18 Bogotá	1500

Movistar Arena Bogotá	Dg. 61c #26-36, Bogotá	14000
Parque Simón Bolívar	Av. Calle 53 y Av., Bogotá	140000
Plaza de Bolívar	Cra. 7 #11-10, Bogotá	55600
Restaurante bar El Sitio	Cra. 11a #93b12, Bogotá	400
Royal Center	Cl. 13 #66-80, Bogotá	4500
Sala de conciertos Biblioteca Luis Ángel Arango	Cl. 11 #4-14, Bogotá	367
Teatro ABC	Cl. 104 #1722, Bogotá	967
Teatro Bernardo Romero	Cl. 46 #28-30, Bogotá	200
Teatro Cafam	Ak 68 #9088, Bogotá	2487
Teatro Colón	Calle 10 # 5-32 La Candelaria, Bogotá	1000
Teatro El Parque	Cra. 5 # 36-05, Bogotá	170
Teatro Ernesto Aronna	# 45a, Cra. 20 #59, Bogotá	110
Teatro Jorge Eliécer Gaitán	Cra. 7 #22-47, Bogotá	1685
Teatro Libre de Chapinero	Cl. 62 #9A-65, Bogotá	204
Teatro Libre del Centro	Cl. 62 #9A-65, Bogotá	200
Teatro Mayor Julio Mario Santo Domingo	Cl. 170 #67-51, Bogotá	1320
Teatro Nacional Fanny Mikey	Cl. 71 #10-25, Bogotá	351
Teatro Nacional La Castellana	Cl. 95 #47-15, Bogotá	714
Teatro Petra	# 39, Cra. 15 Bis ##39 - 39, Bogotá	171
Teatro Sala Belarte	Cra. 7 #152- 54, Bogotá	239
Teatro Santa Fe	Cl. 57 #17-13, Bogotá	400
Vinyl Box Music Center	Cra. 28 #42-55, Bogotá	100

La distribución espacial de los diferentes lugares de los eventos públicos se concentra mayormente en la zona centro occidente y en la zona norte de la ciudad. A diferencia de los accidentes de tránsito se tiene una muestra que cubre todas las zonas de la ciudad como se ilustra en la **Figura 4-1**. Utilizando distintos métodos de geovisualización se puede construir diversos modelos.

Figura 4-1. Visualización geoespacial de los eventos públicos (izquierda) y accidentes de tránsito (Derecha), entre el periodo de septiembre 2018 y julio de 2019.



Fuente: elaboración propia.

4.1 Creación de Buffers

El buffer es una técnica de análisis espacial que se utiliza normalmente en los SIG para las operaciones entre capas. Por medio de zonas buffer se denota las áreas de influencia de los eventos. Con el objetivo de demarcar la región de los accidentes de tránsito que pueden tener relación con la presencia de eventos públicos. El buffer genera dos áreas, una que se encuentra dentro de una distancia especificada a un objeto espacial y otra área que está fuera. El área interna demarcada por la distancia especificada es nombrada como la zona buffer [49].

En los sistemas de información geográficos las zonas buffer están representadas como polígonos vectoriales, rodeando a un punto, línea u otro polígono. Un objeto puede también tener más de una zona buffer [50]. Teniendo en cuenta que en la visualización de la

distribución espacial de puntos se tiene la ubicación del evento; en este no se contempla la magnitud de establecimiento o área en la cual se presenta el evento. Por esta razón se crea la necesidad de generar zonas buffer que cubran tanto el área del lugar del evento público, como la de los accidentes de tránsito alrededor de este. Contemplando el aforo que los lugares presentan, el cual la movilidad puede verse afectada, sea por una reducción por la velocidad a la que se maneja o un aumento por la afluencia de vehículos en la zona. Con la finalidad de abarcar las zonas de influencia/afluencia donde corren el riesgo de congestión o de aumento de accidentalidad.

4.1.1 Zonas buffer múltiples

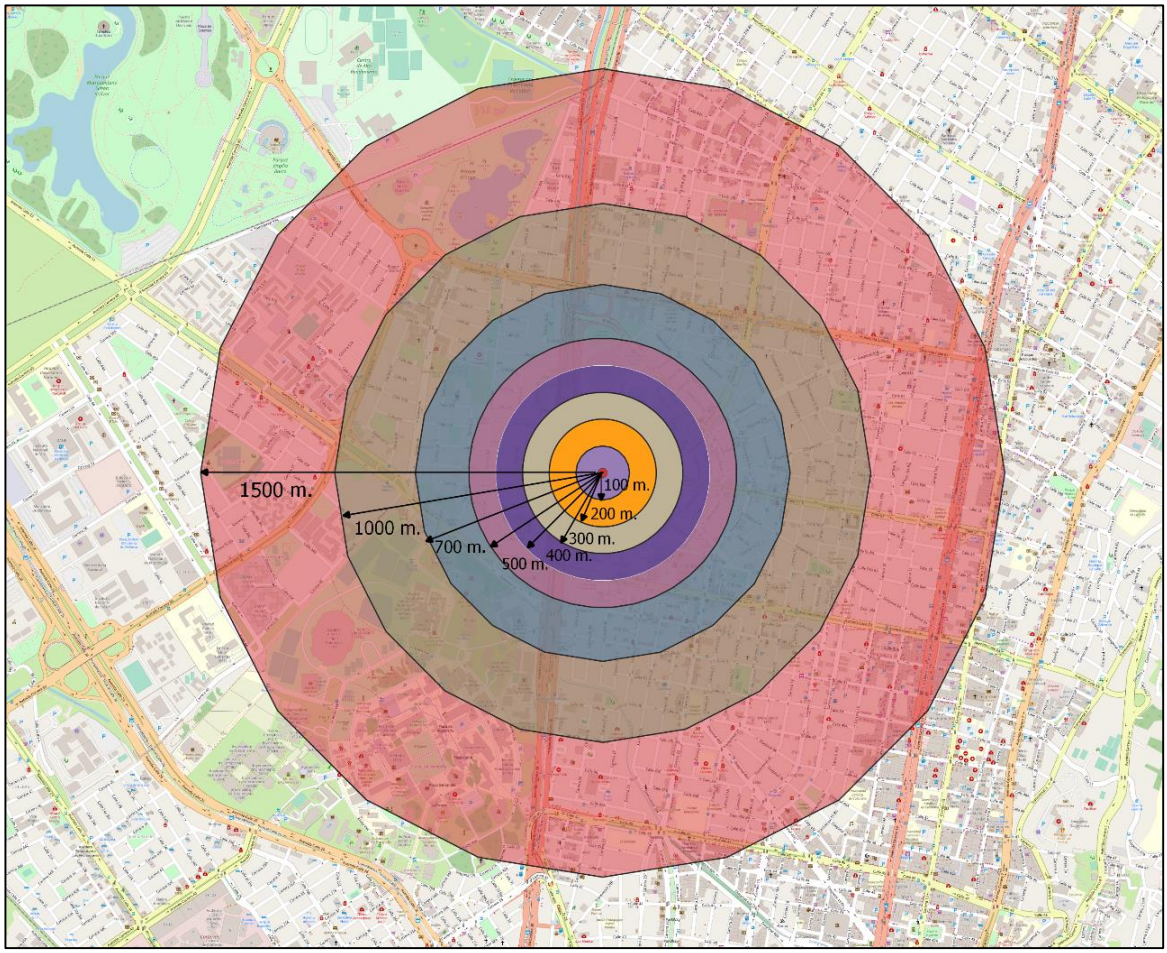
Hay varias alternativas a la hora de generar un buffer. La distancia de buffer o tamaño de buffer puede variar conforme a valores de atributos de la capa vectorial para cada entidad. Los valores numéricos tienen que ser definidos en unidades del mapa conforme al Sistema de Referencia de Coordenadas (SRC) utilizado con los datos.

Tabla 4-2. Equivalencias de grados con metros de las zonas buffer.

Grados	***Metros***
0.000898	100
0.00179	200
0.00269	300
0.00359	400
0.00449	500
0.00628	700
0.00898	1000
0.01347	1500

Para la generación de buffers se crearon zonas con diferentes distancias, comprendidas entre los 100 y 1500 metros a partir del centro o ubicación del lugar. En la **Tabla 4-2**, se nombra las diferentes distancias de los buffers creados y su equivalencia en grados para su generación conforme al SRC. En la **Figura 4-2**, se ilustra las zonas buffers creados alrededor de los eventos públicos como área de influencia en los accidentes vehiculares.

Figura 4-2. Buffers múltiples alrededor de los lugares de los eventos públicos



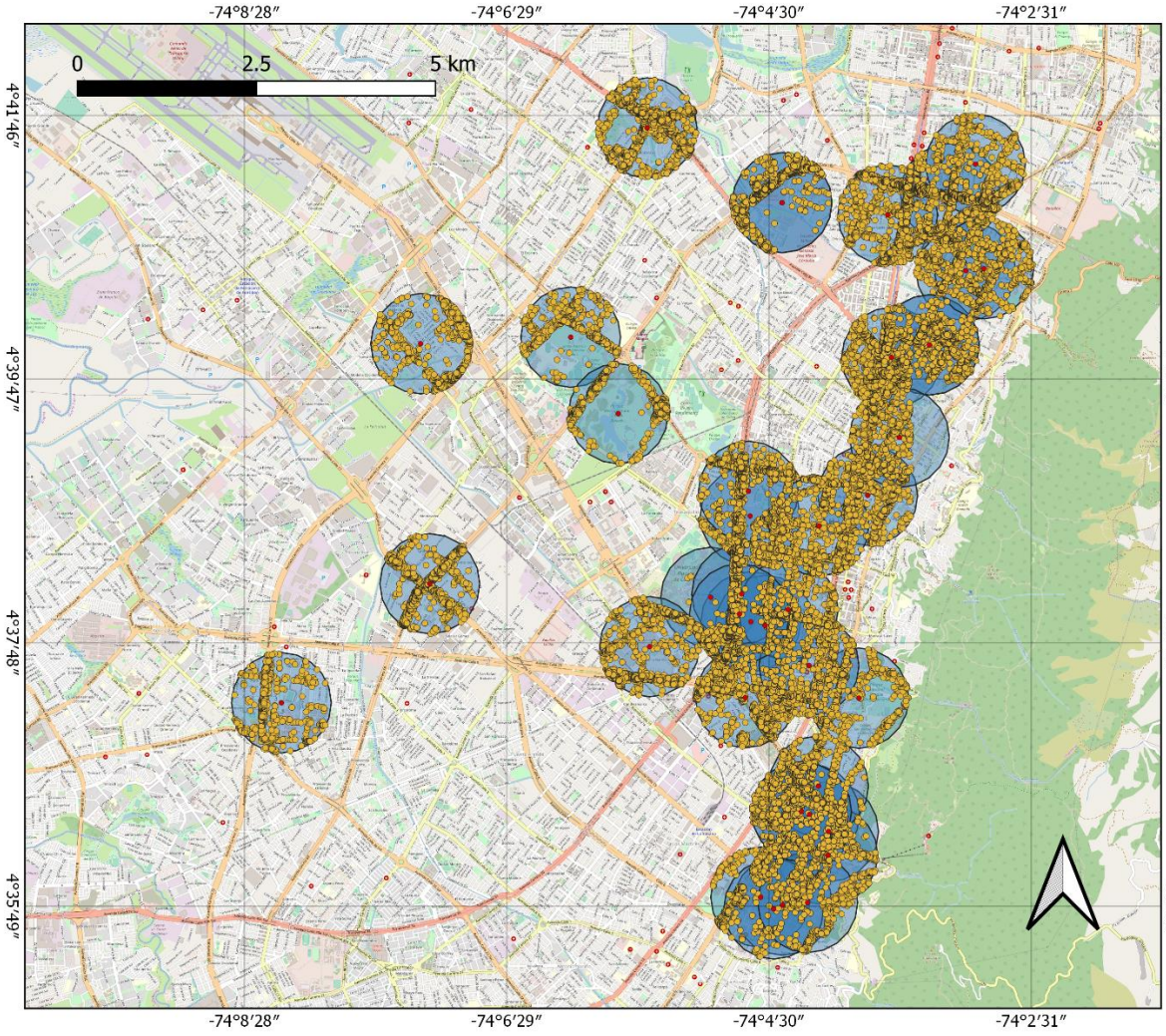
Fuente: elaboración propia.

4.1.2 Intersección de capas y objetos

La Superposición espacial es un proceso que permite identificar relaciones entre dos capas con objetos vectoriales, polígonos o puntos, el cual comparten parte o una misma superficie. La capa de salida es una combinación de características y atributos de las entidades de entrada [51]. Para el análisis espacial se utiliza el proceso de intersección

donde se tiene una capa de salida que contiene todas las áreas y objetos donde ambas capas se interceptan. Con base a esta técnica se obtiene una capa con los accidentes de tránsito que se encuentra dentro de la zona buffer de los eventos públicos, como se muestra en la **Figura 4-3**.

Figura 4-3. Intersección entre Buffers de eventos públicos y accidentes vehiculares.

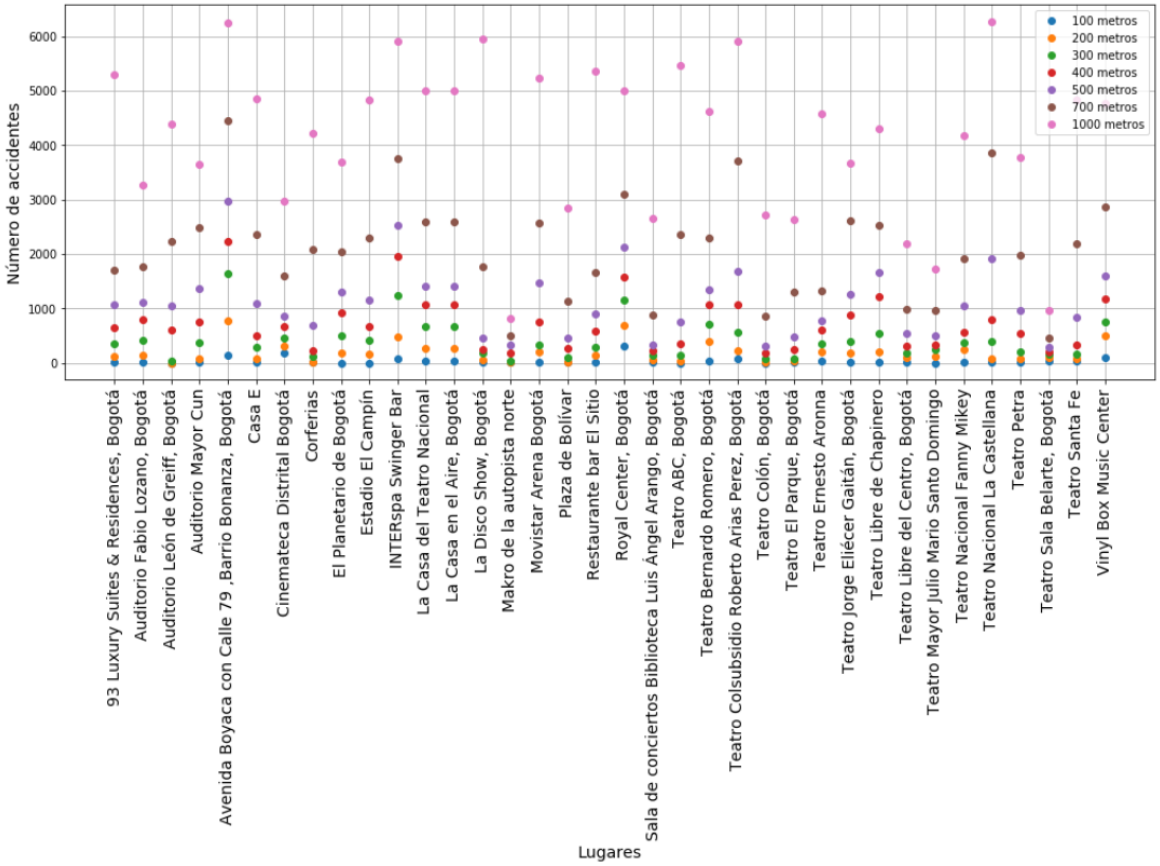


Fuente: elaboración propia.

Con base a los datos obtenidos en el análisis espacial, se obtiene que los accidentes pueden estar relacionados con los eventos públicos. Se procede a realizar un análisis temporal. En un primer paso se tiene los eventos públicos con el número de accidentes de tránsito por cada una de las áreas de influencia (buffers) creados anteriormente. En la

Figura 4-4, se muestra el número de accidentes que ocurrieron dentro de las distintas distancias de los lugares de ocurrencia de eventos. Donde se resalta los lugares con mayor número de accidentes y el comportamiento a una mayor distancia.

Figura 4-4. Número de accidentes por ubicación de los eventos públicos en comparación con la distancia.



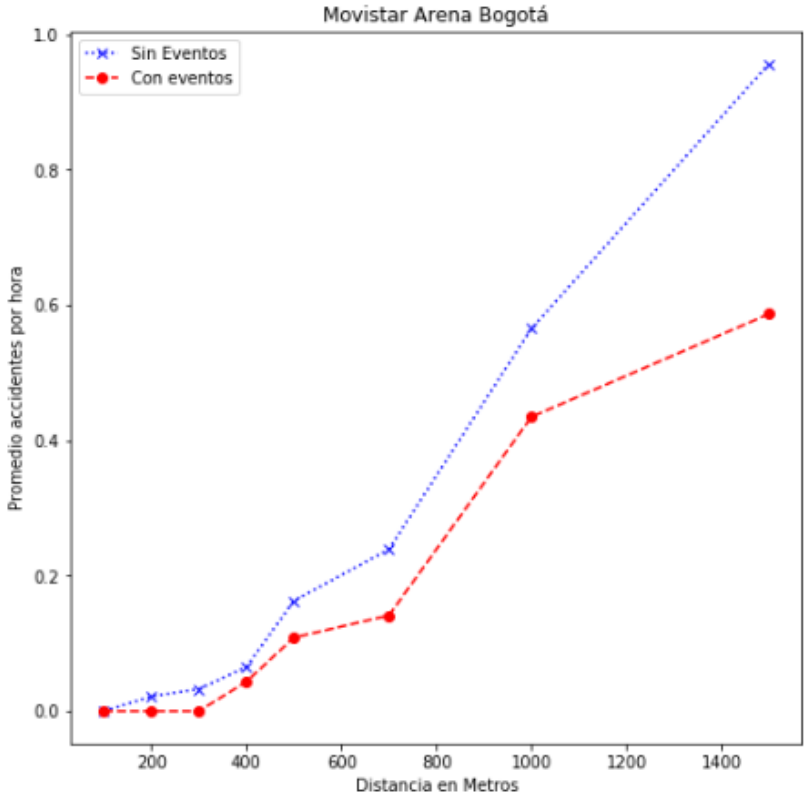
Fuente: elaboración propia.

4.2 Análisis por zona de influencia

Con el fin de observar el comportamiento de la ocurrencia de los accidentes, se realiza un análisis temporal en las zonas de influencia, cuando hay y no presencia de eventos. Teniendo en cuenta los eventos capturados y los lugares de ocurrencia, se seleccionaron cinco lugares a partir de la accidentalidad (**Figura 4-4**), el aforo y tipo de evento. Con base a lo anterior se seleccionaron los siguientes lugares: El Movistar Arena, estadio el Campín,

teatro Nacional la Castellana, parque Simón Bolívar y el teatro Jorge Eliecer Gaitán. Donde se tienen lugares de eventos deportivos, festivales, presentación de obras de teatros y conciertos, al igual algunos de los presentan los más grandes aforos. Para el respectivo análisis se procedió a contar el número de accidentes por hora, para cada uno de los conjuntos de datos de las áreas de influencia. Posterior al conteo, se procedió a calcular el promedio de accidentes por hora cuando hay presencia o ausencia de eventos públicos. En la **Figura 4-5**, se muestra el comportamiento del promedio de accidentalidad por hora cuando la distancia de área de influencia aumenta con respecto al lugar de ocurrencia. El análisis se lleva a cabo en cada uno de los lugares.

Figura 4-5. Promedio de accidentes por hora de acuerdo con la distancia de influencia en el Movistar Arena.

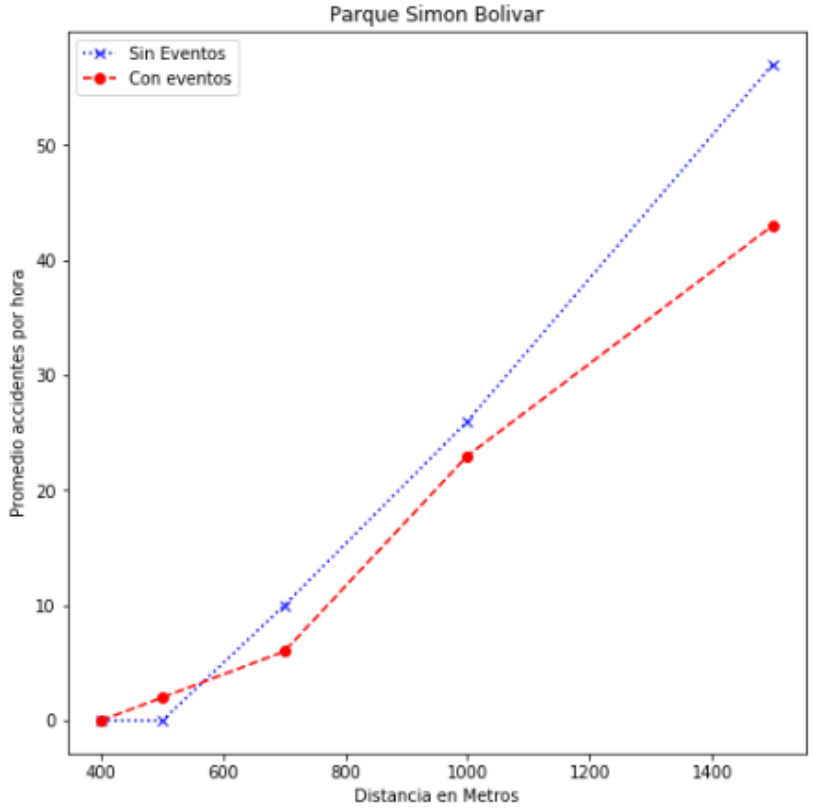


Fuente: elaboración propia.

Inicialmente entre los primeros 300 del centro del lugar no ocurren accidentes cuando hay presencia de eventos públicos, teniendo en cuenta que el diámetro del lugar es de unos

400 metros aproximadamente. Conforme la distancia aumenta se observa que el número de accidentes es mayor cuando no hay eventos. A pesar de los promedios mantienen una cercanía para ambas muestras, en los 1500 metros se presenta una considerable diferencia respecto a las demás distancias.

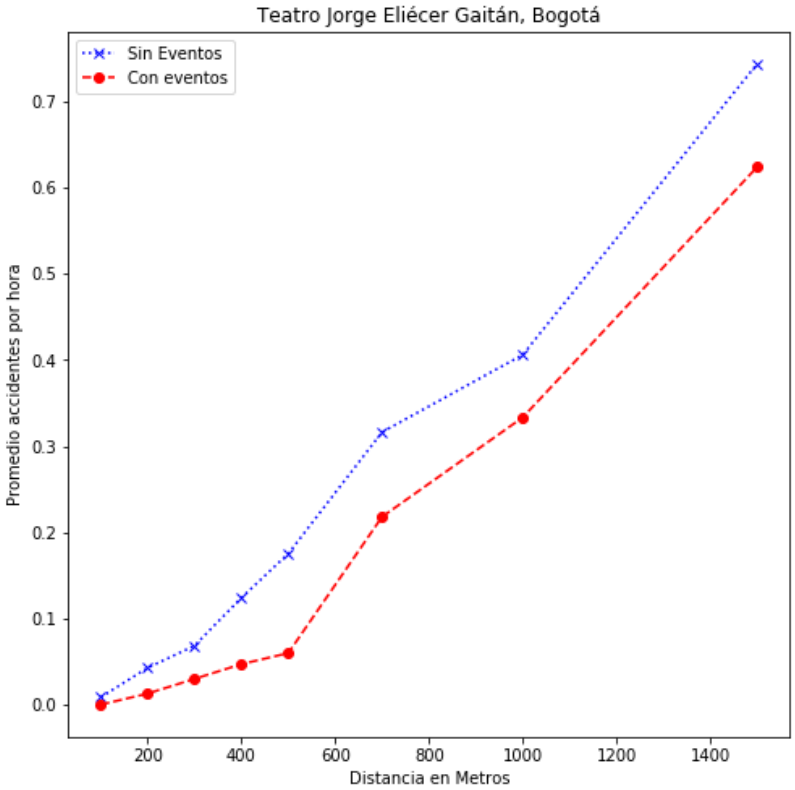
Figura 4-6. Promedio de accidentes por hora de acuerdo con la distancia de influencia en el Parque Simón Bolívar.



Fuente: elaboración propia.

Teniendo en cuenta que el diámetro de que el parque Simón Bolívar es de unos 900 metros aproximadamente, se presentan accidentes en un radio mayor a los 400 metros (**Figura 4-6**). En este ocurren más accidentes cuando hay presencia de eventos públicos en un radio de 500 metros. Conforme la distancia aumenta se observa que el número de accidentes es mayor cuando no hay eventos. A pesar de los promedios mantienen una cercanía para ambas muestras, en los 1500 metros se presenta una considerable diferencia respecto a las demás distancias.

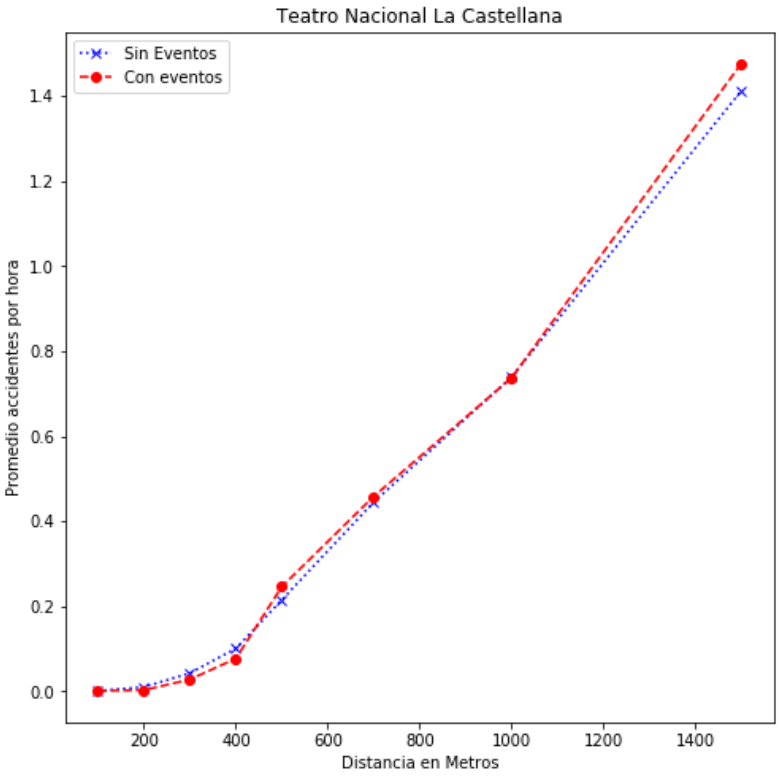
Figura 4-7. Promedio de accidentes por hora de acuerdo con la distancia de influencia en el Teatro Jorge Eliécer Gaitán.



Fuente: elaboración propia.

Siendo el teatro Jorge Eliecer Gaitán un lugar pequeño se tiene accidentes a partir de los 100 metros como se ilustra en la **Figura 4-7**. Para este caso el promedio de los accidentes es mayor cuando no hay presencia de eventos públicos y tienen un comportamiento de ocurrencia similar tanto cuando hay y no presencia de eventos.

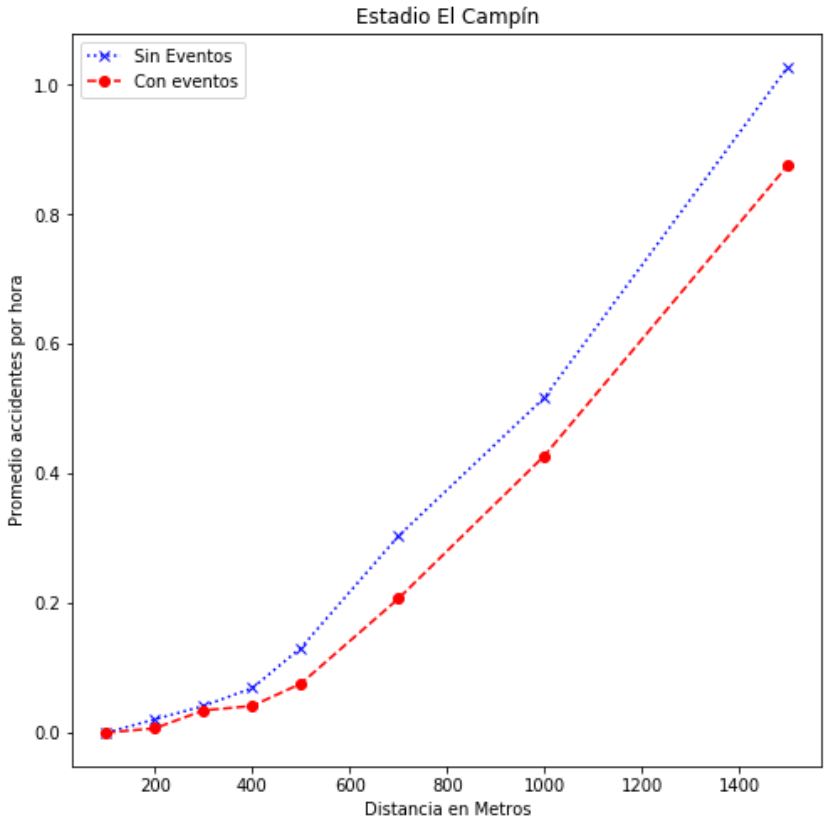
Figura 4-8. Promedio de accidentes por hora de acuerdo con la distancia de influencia en el Teatro Nacional la Castellana.



Fuente: elaboración propia.

A diferencia de los demás lugares en el teatro nacional la castellana la ocurrencia de los accidentes es similar cuando hay y no presencia de eventos públicos (**Figura 4-8**). En las diferentes distancias los accidentes no presentan diferencias significativas entre las dos muestras.

Figura 4-9. Promedio de accidentes por hora de acuerdo con la distancia de influencia en el Estadio El Campín.



Fuente: elaboración propia.

En la **Figura 4-9**, se muestra la ocurrencia de accidentes a partir de los 200 metros. El cual de observa que el promedio de ocurrencia de accidentes es mayor cuando no hay presencia de eventos públicos. En las diferentes distancias el comportamiento de la ocurrencia de los accidentes es similar para ambos casos.

4.3 Análisis por Franja Horaria

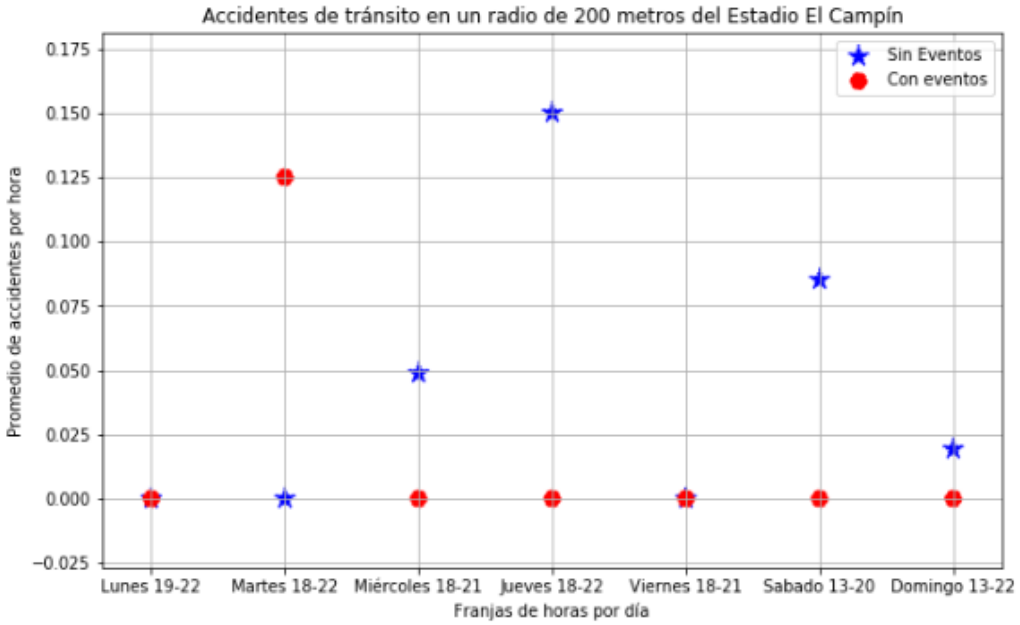
Otro punto de vista en el análisis temporal se basa en la periodicidad de ocurrencia de los eventos. Hay lugares con ocurrencia de eventos en ciertos días de la semana; algunos eventos se realizan entre semanas y otros los fines de semana. Por ejemplo, la mayoría de los teatros, los eventos ocurren con una periodicidad diaria de lunes a viernes; mientras

otros lugares como el Movistar Arena, el estadio el Campín y los bares, se realizan los fines de semana incluyendo los días viernes en algunos casos. A partir de lo anterior se realiza un análisis de forma individual y en un área de influencia específica para los lugares nombrados anteriormente. El proceso da a conocer el promedio de accidentes de tránsito por día de la semana con eventos y sin eventos.

El proceso consta de la selección de los accidentes que se presentaron alrededor del lugar de eventos, a una distancia de 100, 200, 300, 400, 500, 700, 1000 y 1500 metros, junto con los eventos de dicho luchar y dentro de un periodo de tiempo de septiembre 2018 a julio 2019. Posteriormente se cuentan los eventos y accidentes que se presentan por hora, para el caso de los eventos se extiende el valor a una hora anterior y una hora después de la hora de inicio del evento, con el fin de abarcar el tiempo en el que transcurre el evento. Luego de tener el número de accidentes y eventos que ocurrieron por cada hora dentro de septiembre a julio, se toman los registros entre las franjas horarias donde mayormente ocurren los eventos. Finalmente se procede a la agrupación de los accidentes por día de la semana cuando hay y no presencia de eventos. En los diferentes lugares se procesan el mismo número de horas para ambos casos (hay ocurrencia de evento y no hay ocurrencia de evento).

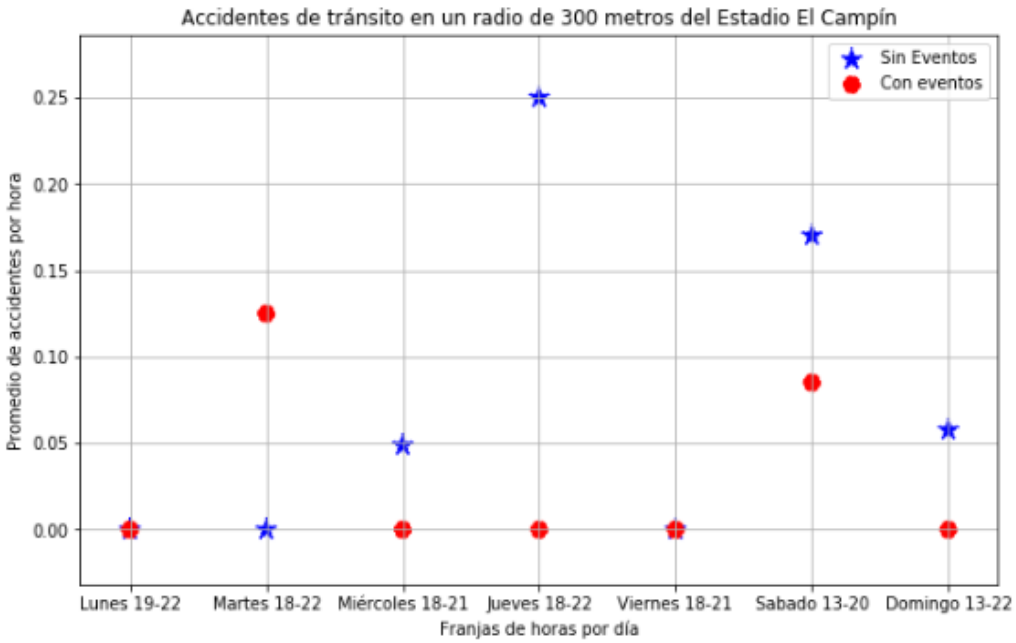
En el análisis realizado en el estadio el Campín (**Figura 4-10**), se tiene en una duración de cuatro horas de ocurrencia por cada evento, el cual se obtuvieron 181 horas de ocurrencias de eventos en una franja horaria entre las 12 y 23 horas.

Figura 4-10. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 200 metros.



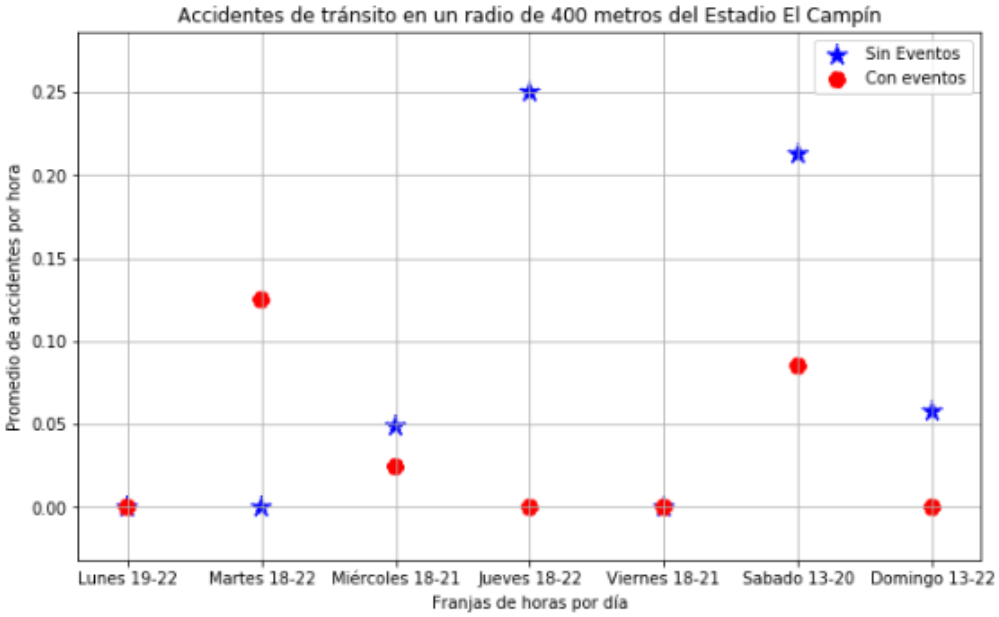
Fuente: elaboración propia.

Figura 4-11. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 300 metros.



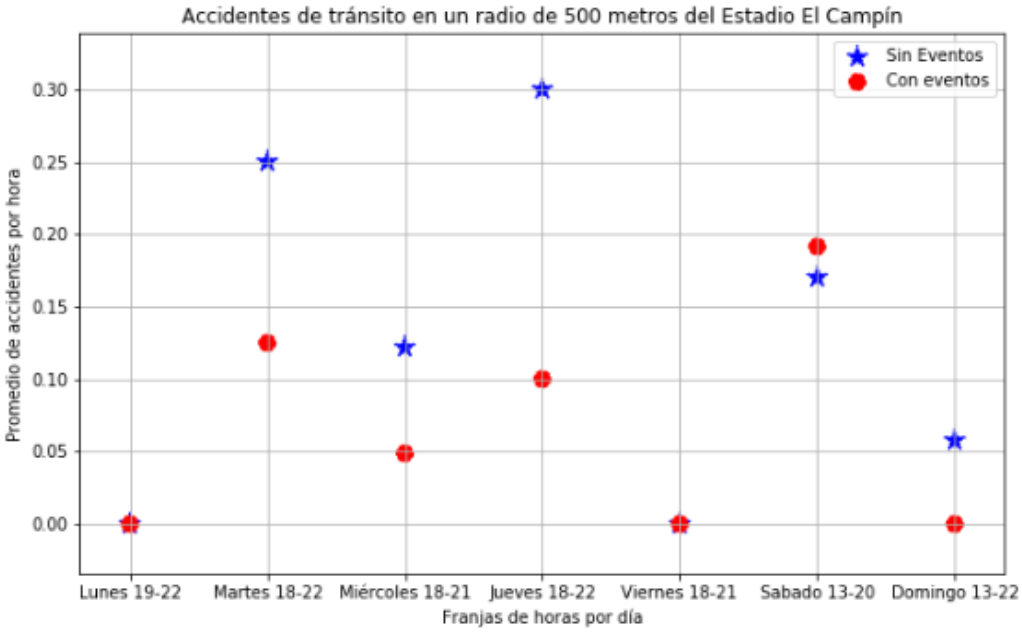
Fuente: elaboración propia.

Figura 4-12. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 400 metros.



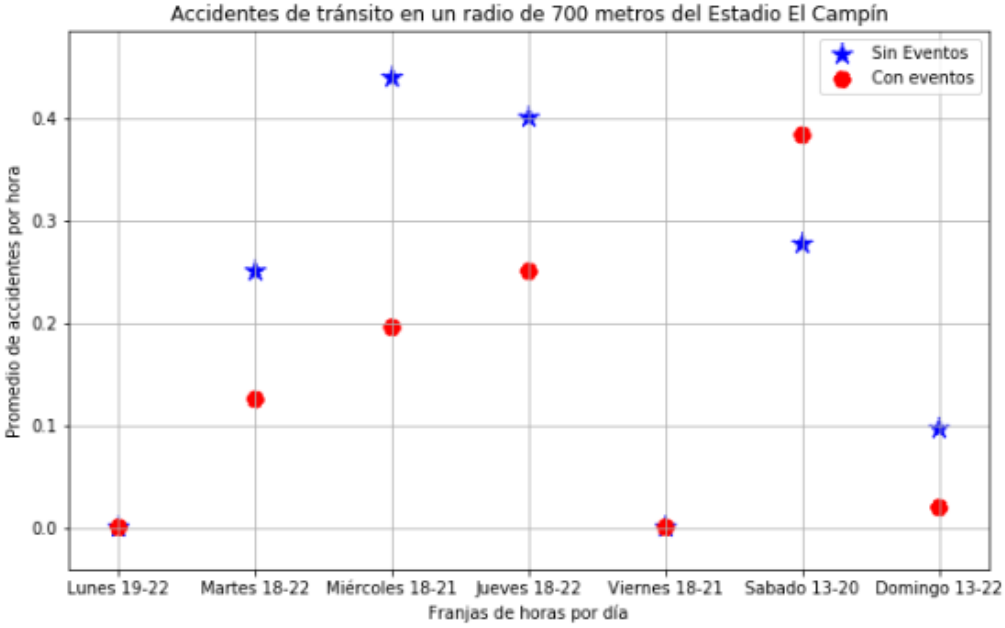
Fuente: elaboración propia.

Figura 4-13. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 500 metros.



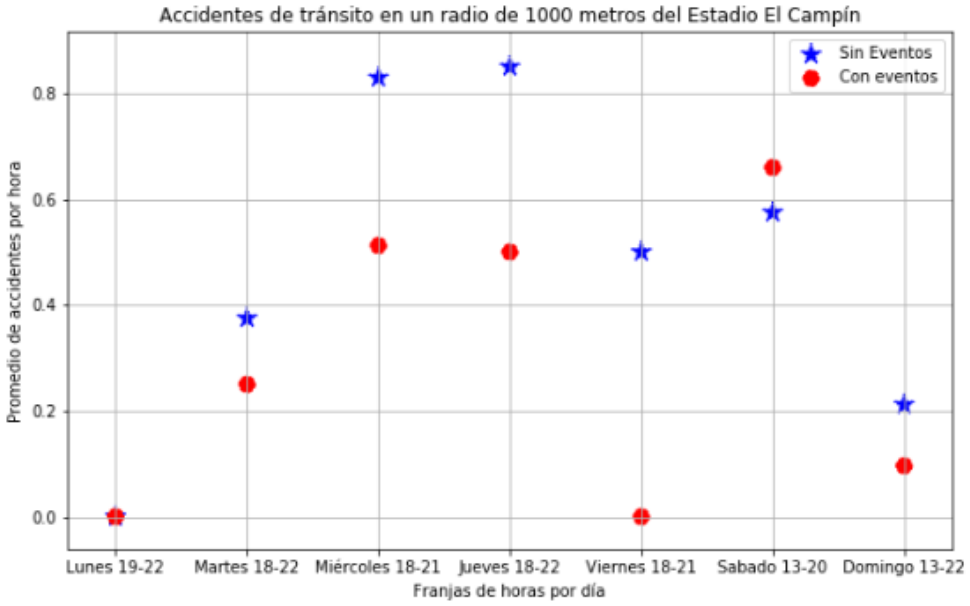
Fuente: elaboración propia.

Figura 4-14. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 700 metros.



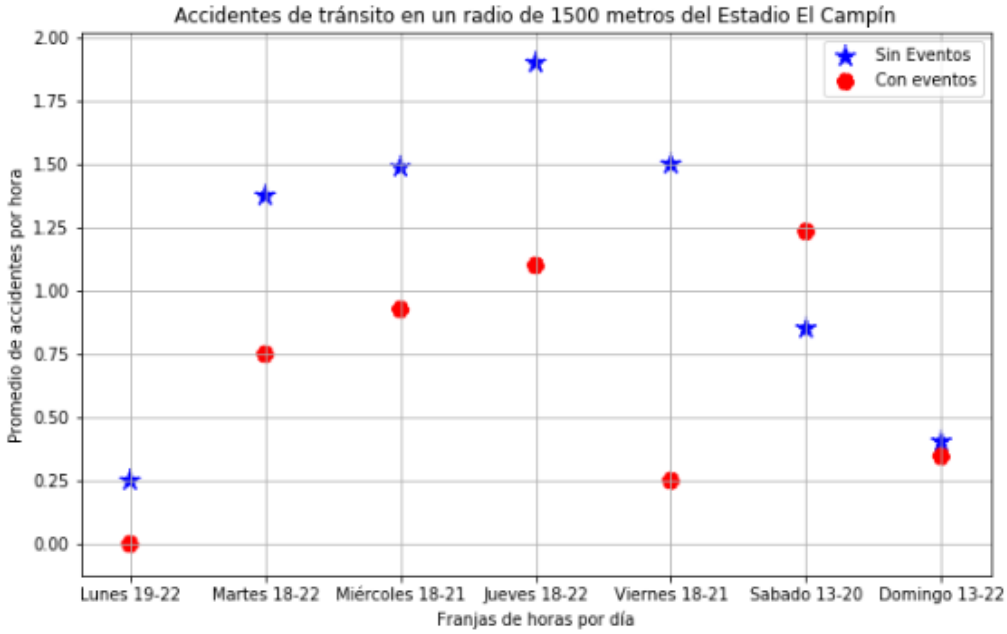
Fuente: elaboración propia.

Figura 4-15. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 1000 metros.



Fuente: elaboración propia.

Figura 4-16. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el estadio El Campín en un radio de 1500 metros.

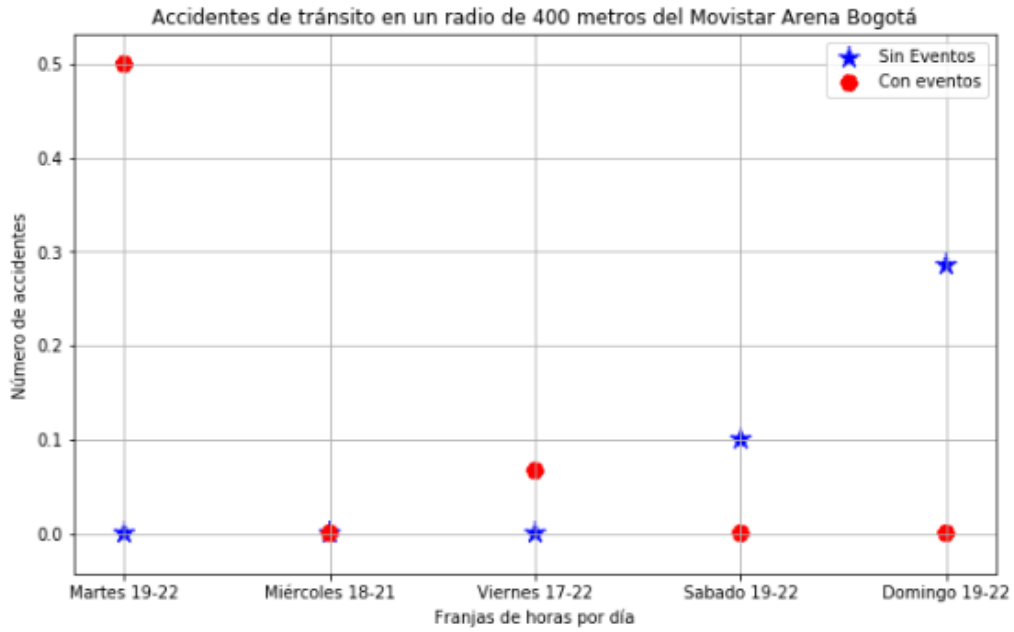


Fuente: elaboración propia.

En el estadio el Campín la ocurrencia de accidentes para los diferentes días de la semana es mayor cuando no hay eventos, a excepción de los días martes y sábados. Para los primeros 400 metros los días martes se tiene una mayor ocurrencia cuando hay presencia de eventos públicos, a partir de los 500 metros los días domingos los accidentes es mayor cuando hay ocurrencia de eventos públicos. El promedio de los accidentes tiene un comportamiento aleatorio para cada una de las diferentes distancias para cada una de las franjas horarias en los siete días de la semana. Dado que a medida que la distancia aumenta el promedio de los accidentes varia disminuyendo o aumentando.

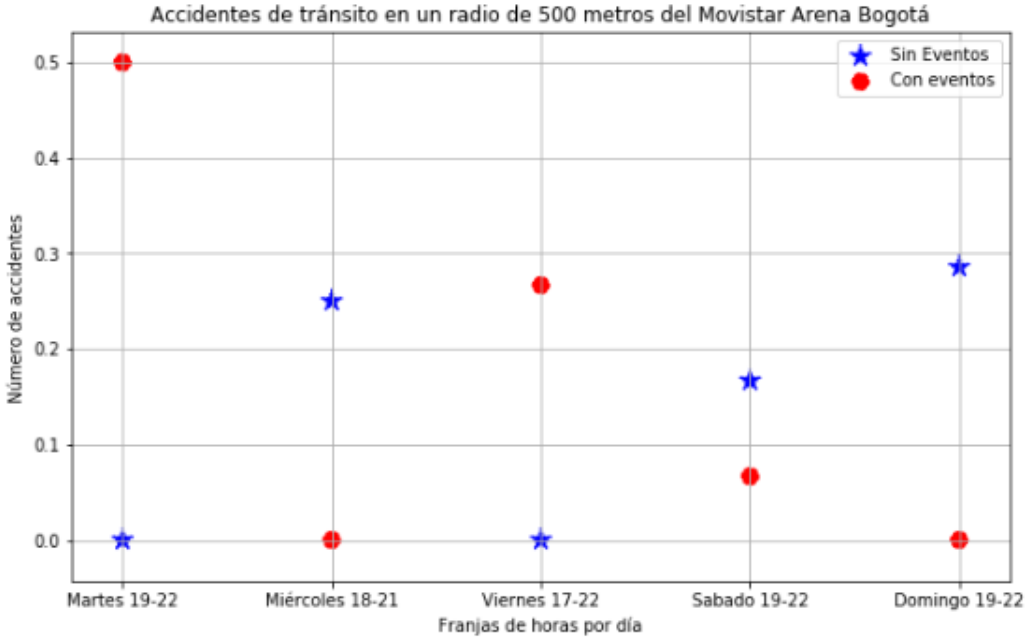
En el análisis realizado en el Movistar Arena (**Figura 4-17**), se tiene en una duración de cuatro horas de ocurrencia por cada evento, el cual se obtuvieron 65 horas de ocurrencias de eventos en una franja horaria entre las 18 y 23 horas.

Figura 4-17. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Movistar Arena en un radio de 400 metros.



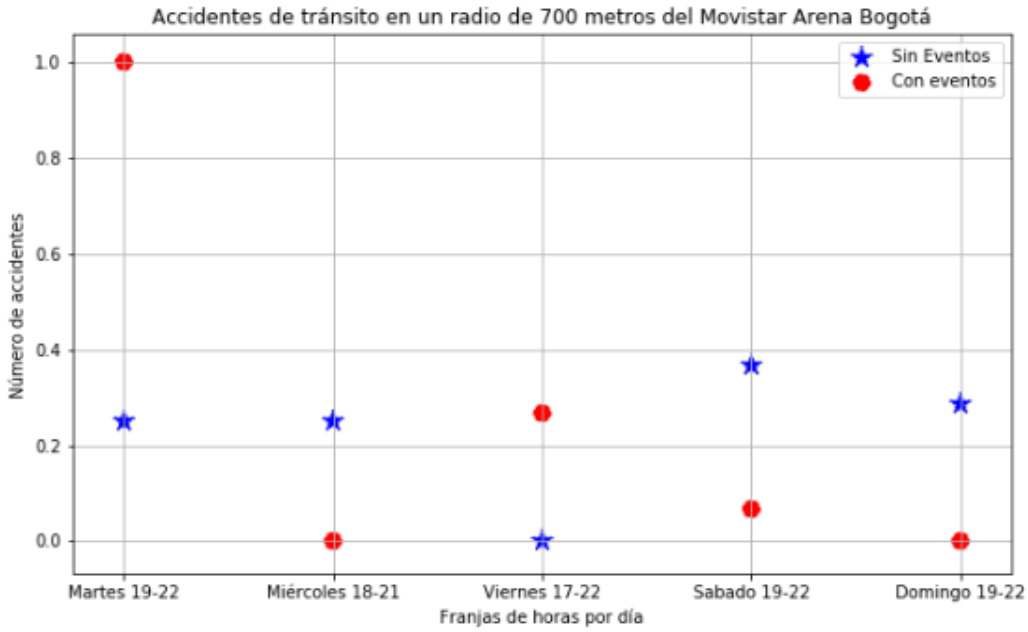
Fuente: elaboración propia.

Figura 4-18. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Movistar Arena en un radio de 500 metros.



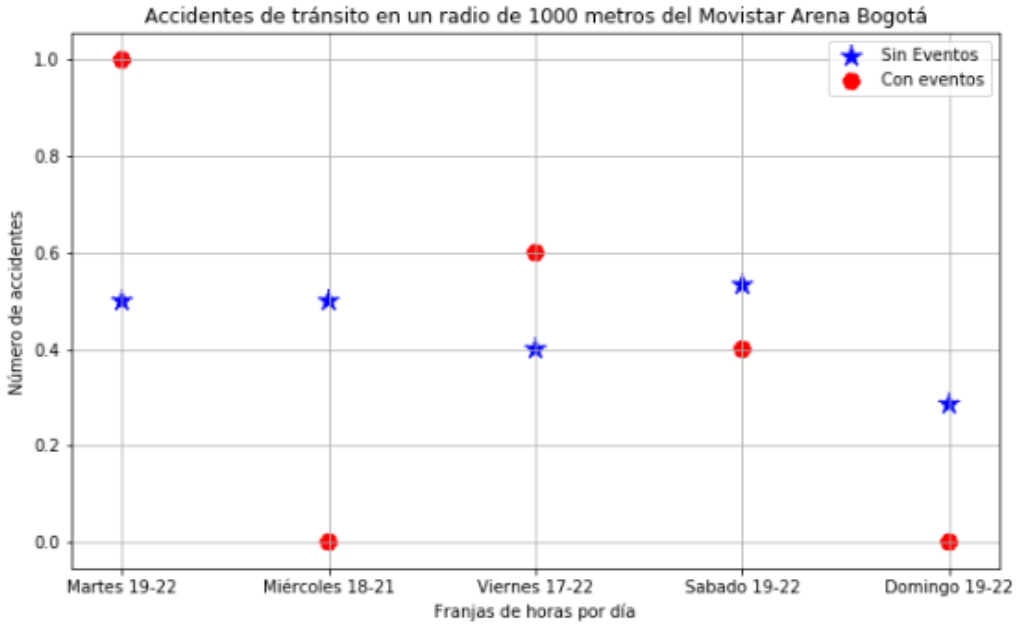
Fuente: elaboración propia.

Figura 4-19. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Movistar Arena en un radio de 700 metros.



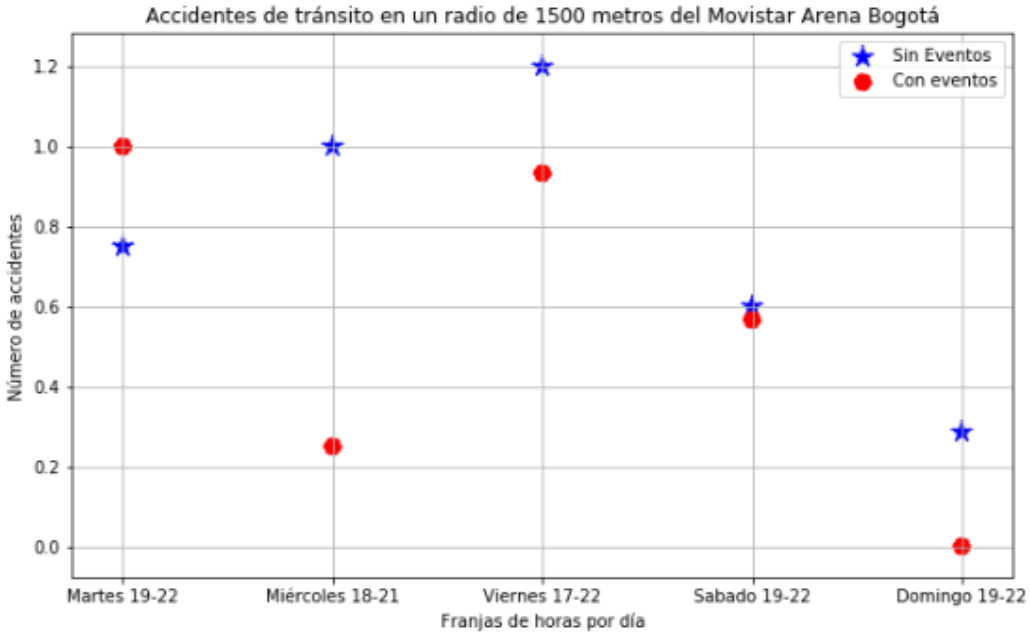
Fuente: elaboración propia.

Figura 4-20. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Movistar Arena en un radio de 1000 metros.



Fuente: elaboración propia.

Figura 4-21. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Movistar Arena en un radio de 1500 metros.

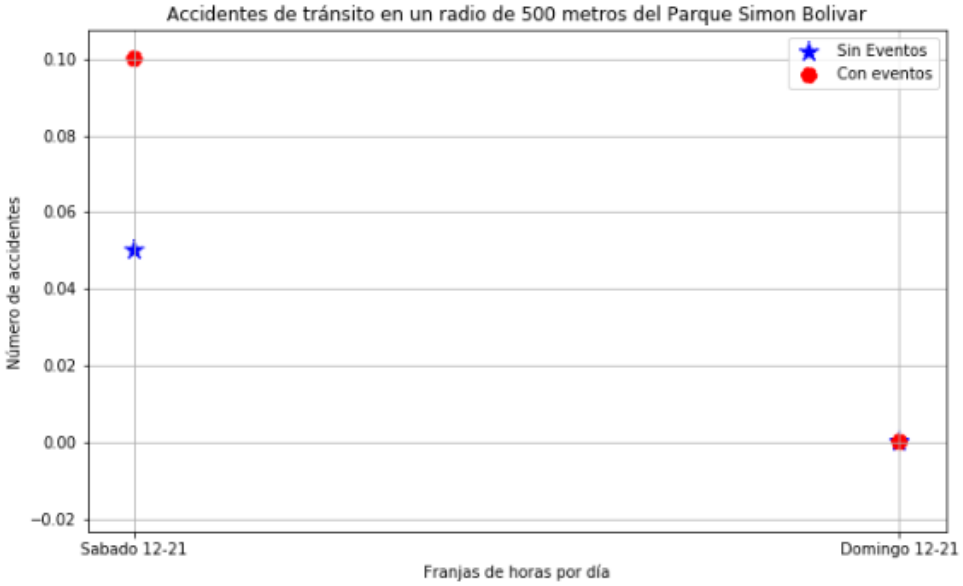


Fuente: elaboración propia.

Dado la ocurrencia de los eventos en el Movistar Arena, solo se tiene ocurrencia de estos en los días martes, miércoles, viernes, sábados y domingos. En cual los días martes y viernes ocurren más accidentes cuando hay presencia de eventos públicos. Al igual también se tiene el mayor promedio de accidentalidad para estos dos días en la mayoría de las distancias.

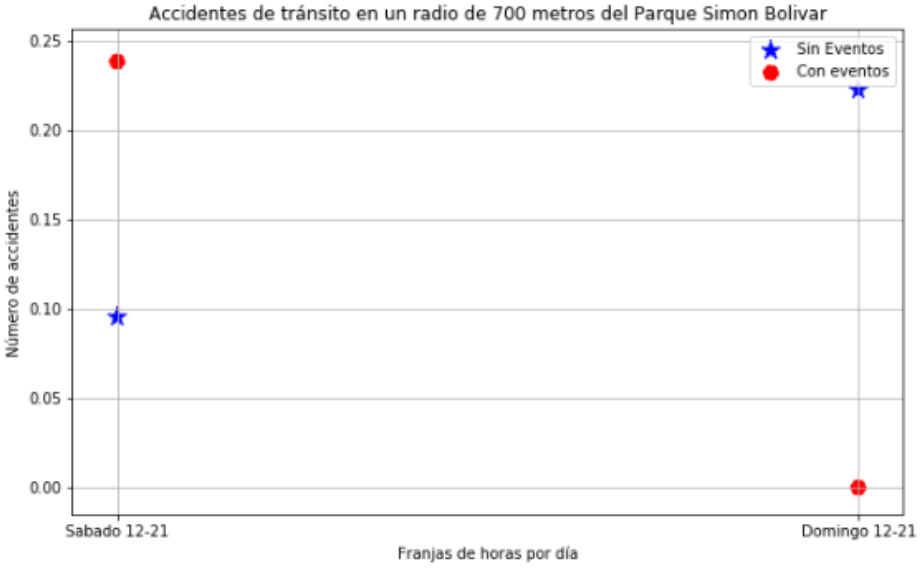
En el análisis realizado en el Parque Simón Bolívar (**Figura 4-22**), se tiene en una duración de cuatro horas de ocurrencia por cada evento, el cual se obtuvieron 39 horas de ocurrencias de eventos en una franja horaria entre las 12 y 22 horas.

Figura 4-22. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Parque Simón Bolívar en un radio de 500 metros.



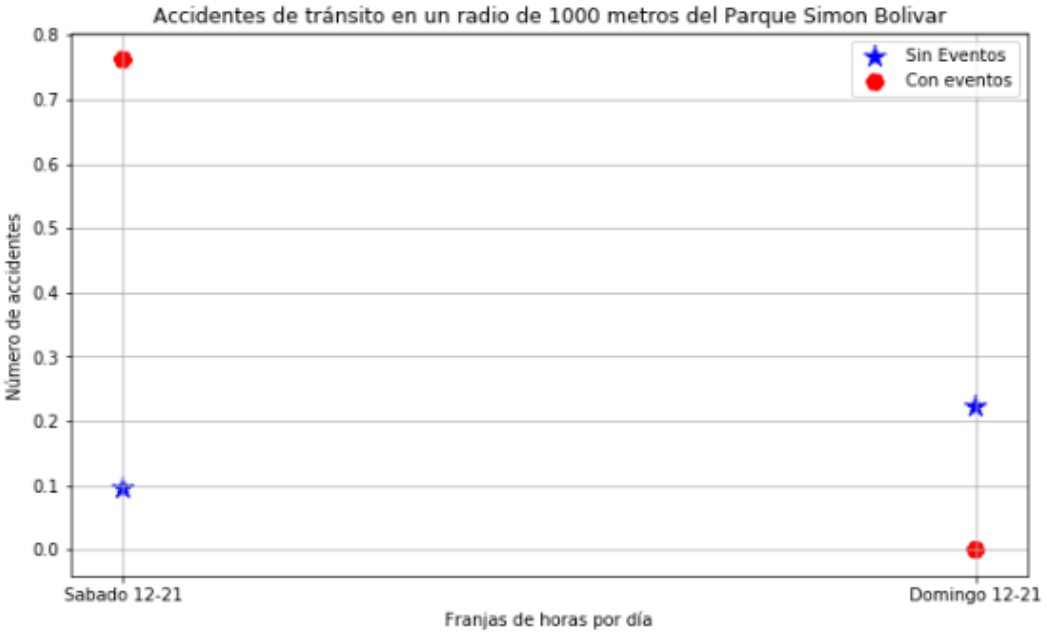
Fuente: elaboración propia.

Figura 4-23. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Parque Simón Bolívar en un radio de 700 metros.



Fuente: elaboración propia.

Figura 4-24. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Parque Simón Bolívar en un radio de 1000 metros.



Fuente: elaboración propia.

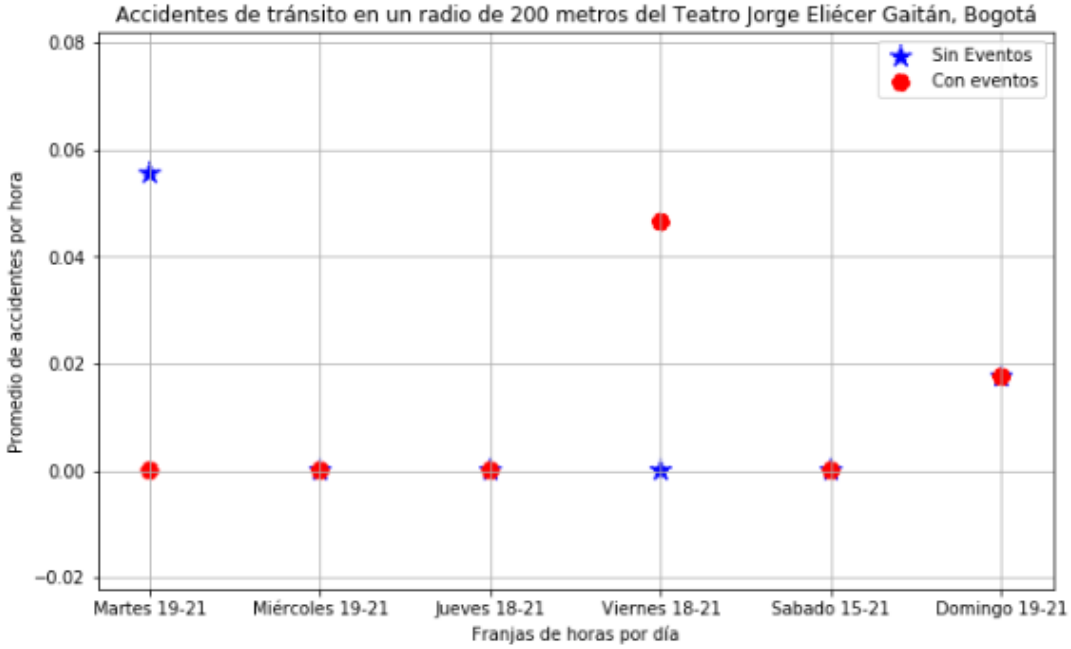
Figura 4-25. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Parque Simón Bolívar en un radio de 1500 metros.

Fuente: elaboración propia.

Las diferentes eventos y festivales ocurridos en Parque Simón Bolívar, solo se llevaron a cabo los días sábados y domingos, dando a lugar un análisis para estos dos días. Como resultado se da que para los días sábados la ocurrencia de accidentes es mayor para cuando hay presencia de eventos públicos a diferencia de los días domingos que es mayor cuando no hay presencia de eventos.

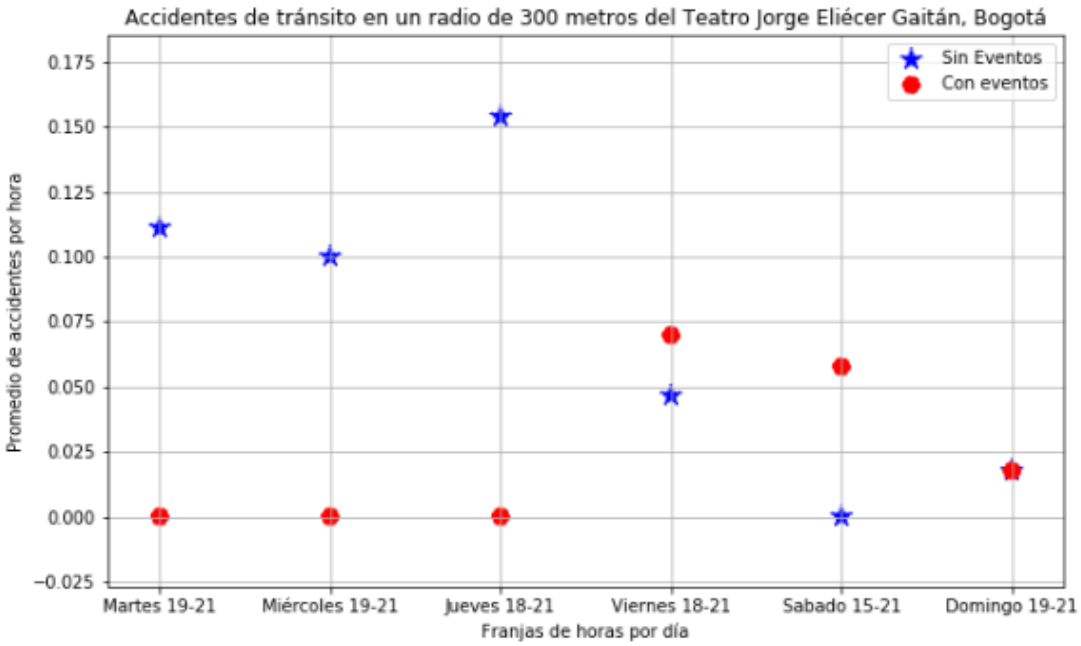
En el análisis realizado en el Teatro Jorge Eliecer Gaitán (**Figura 4-26**), se tiene en una duración de tres horas de ocurrencia por cada evento, el cual se obtuvieron 221 horas de ocurrencias de eventos en una franja horaria entre las 16 y 21 horas.

Figura 4-26. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 200 metros.



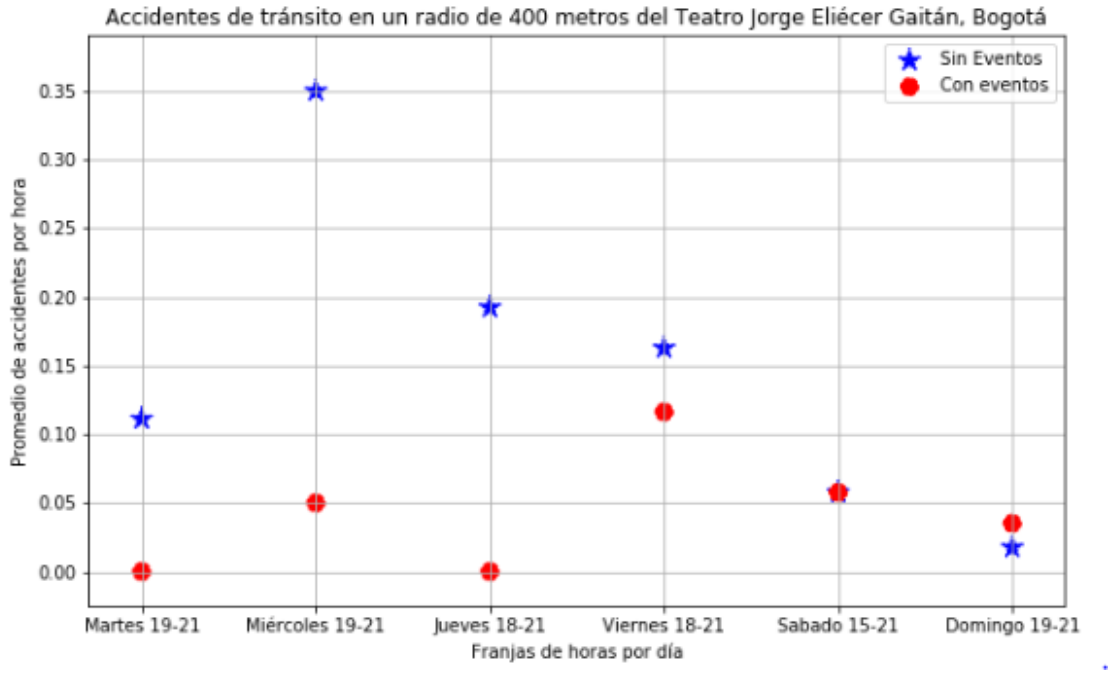
Fuente: elaboración propia.

Figura 4-27. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 300 metros.



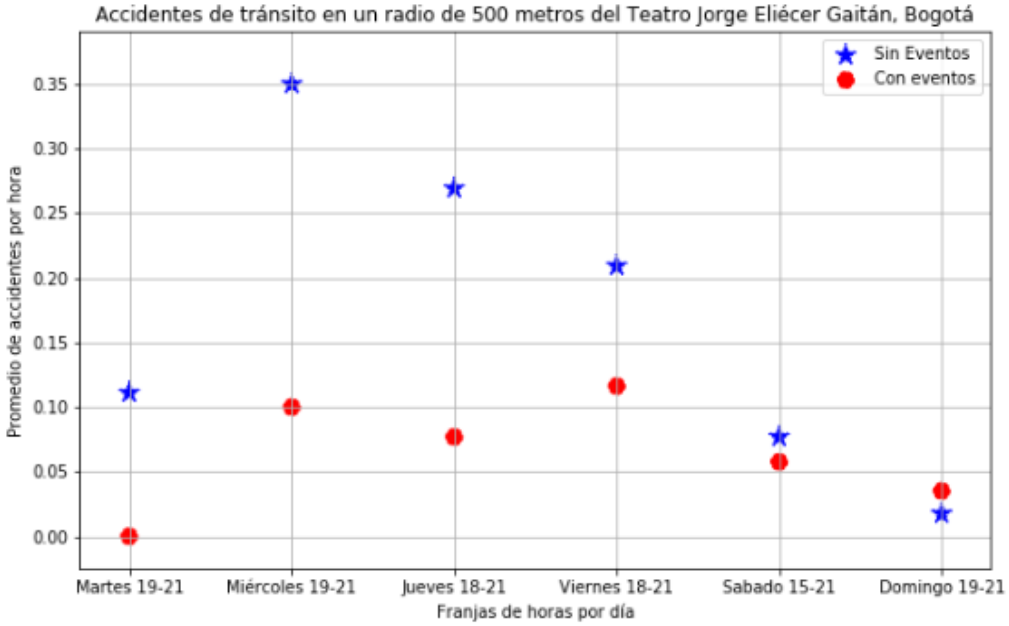
Fuente: elaboración propia.

Figura 4-28. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 400 metros.



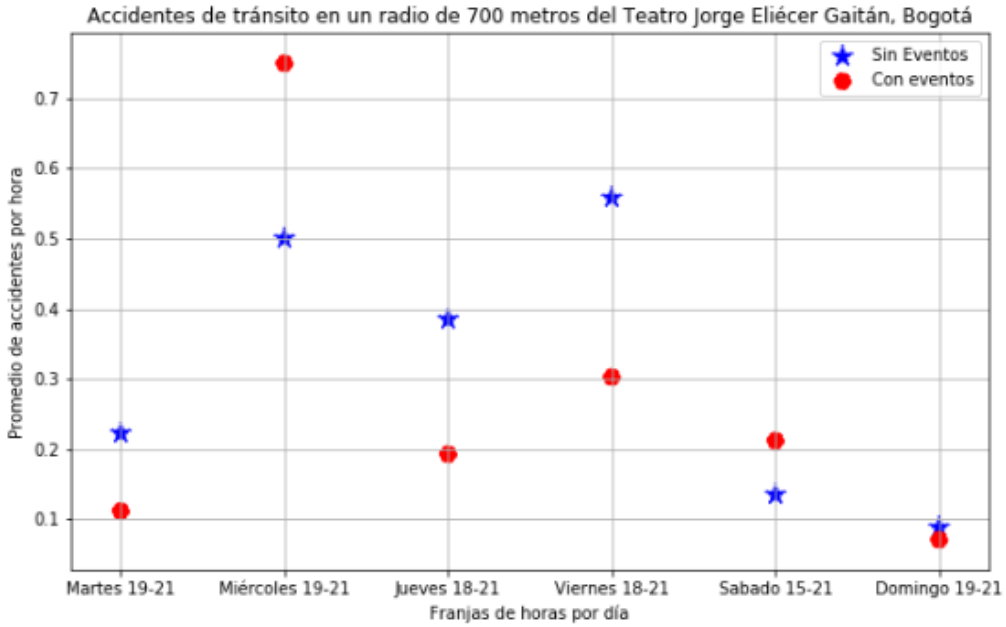
Fuente: elaboración propia.

Figura 4-29. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 500 metros.



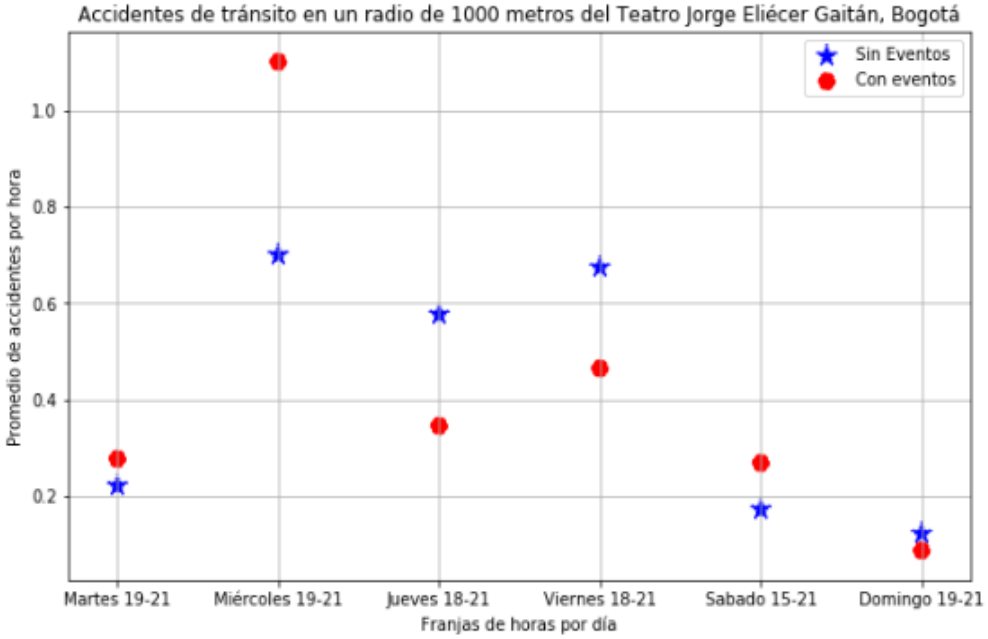
Fuente: elaboración propia.

Figura 4-30. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 700 metros.



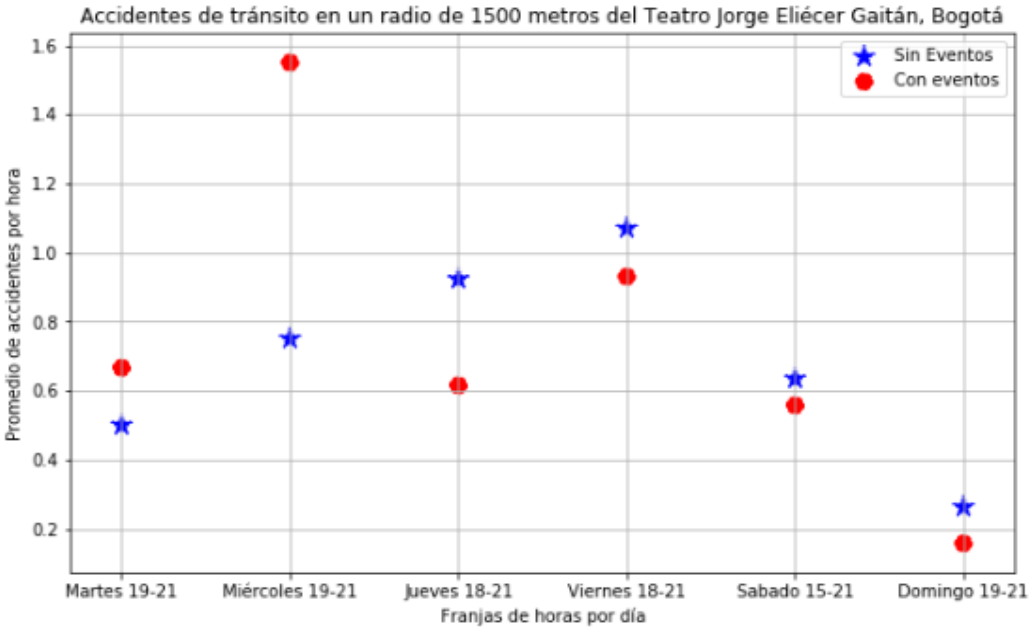
Fuente: elaboración propia.

Figura 4-31. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 1000 metros.



Fuente: elaboración propia.

Figura 4-32. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Jorge Eliecer Gaitán en un radio de 1500 metros.

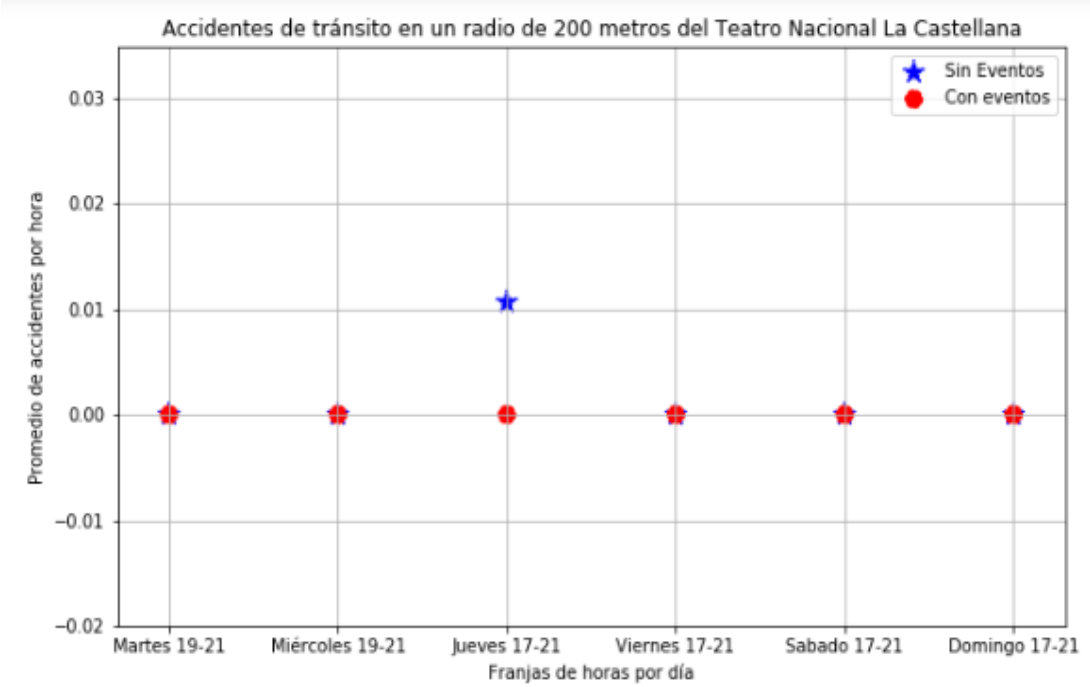


Fuente: elaboración propia.

En las diferentes distancias los accidentes alrededor del teatro Jorge Eliecer Gaitán no se tiene el mismo comportamiento. En el sentido de que a medida que va aumentando la distancia, el promedio de accidentes presentan una aleatoriedad donde para algunos casos los accidentes es mayor par cuando hay presencia de eventos públicos y en una distancia mayor para la misma franja la accidentalidad en mayor en ausencia de eventos públicos.

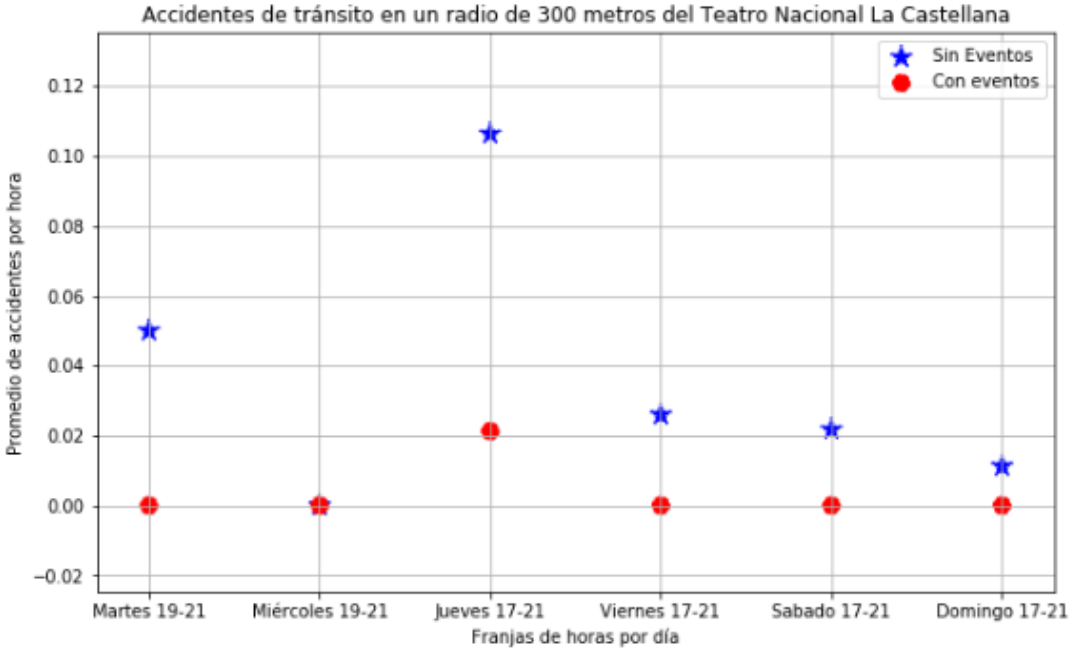
En el análisis realizado en el Teatro Nacional la Castellana (**Figura 4-33**), se tiene en una duración de tres horas de ocurrencia por cada evento, el cual se obtuvieron 562 horas de ocurrencias de eventos en una franja horaria entre las 17 y 21 horas.

Figura 4-33. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 200 metros.



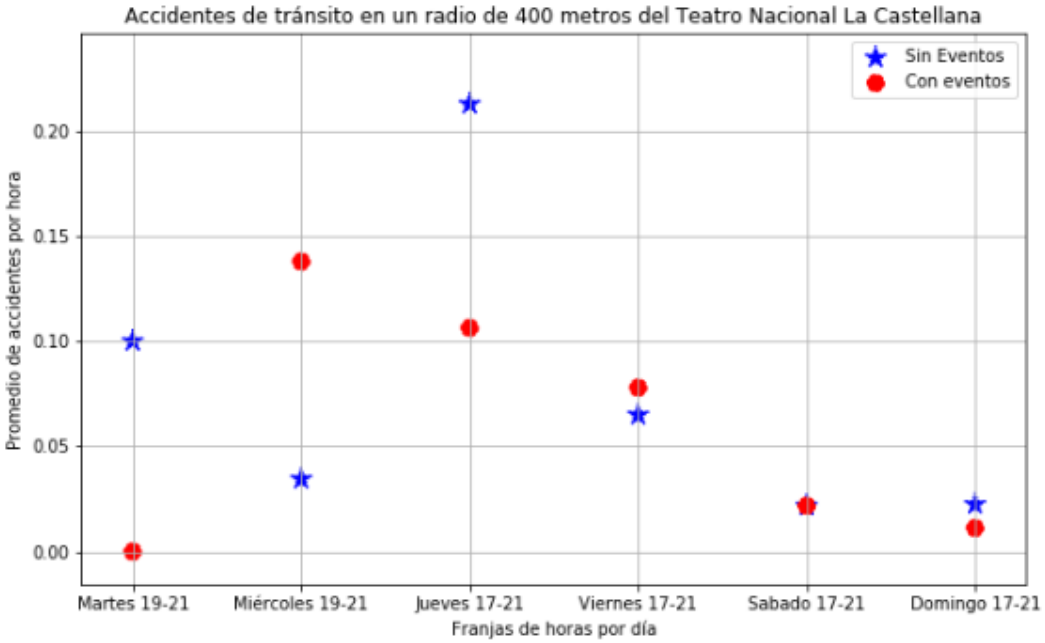
Fuente: elaboración propia.

Figura 4-34. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 300 metros.



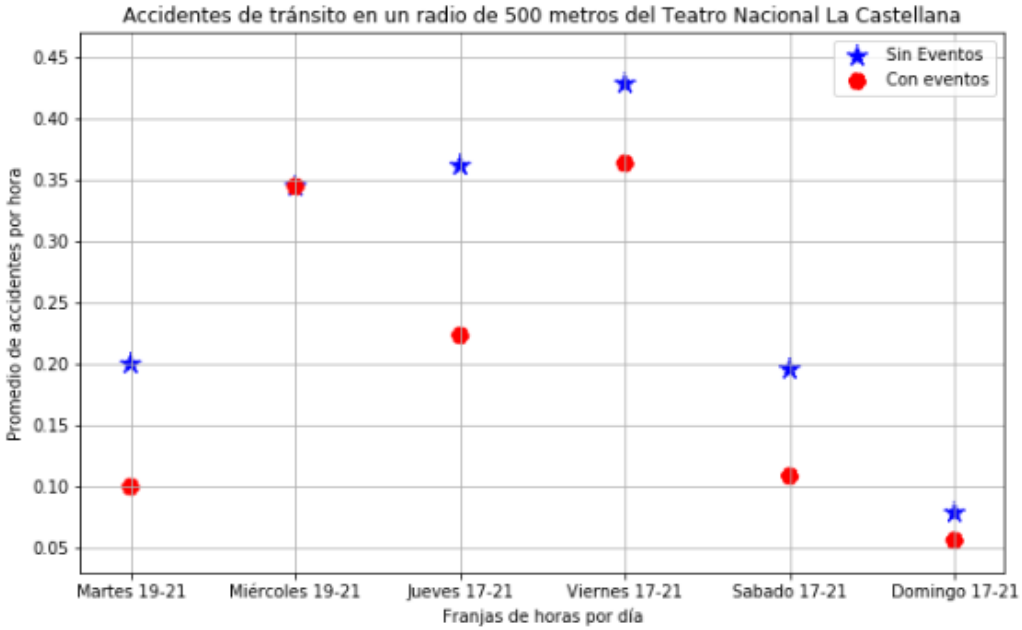
Fuente: elaboración propia.

Figura 4-35. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 400 metros.



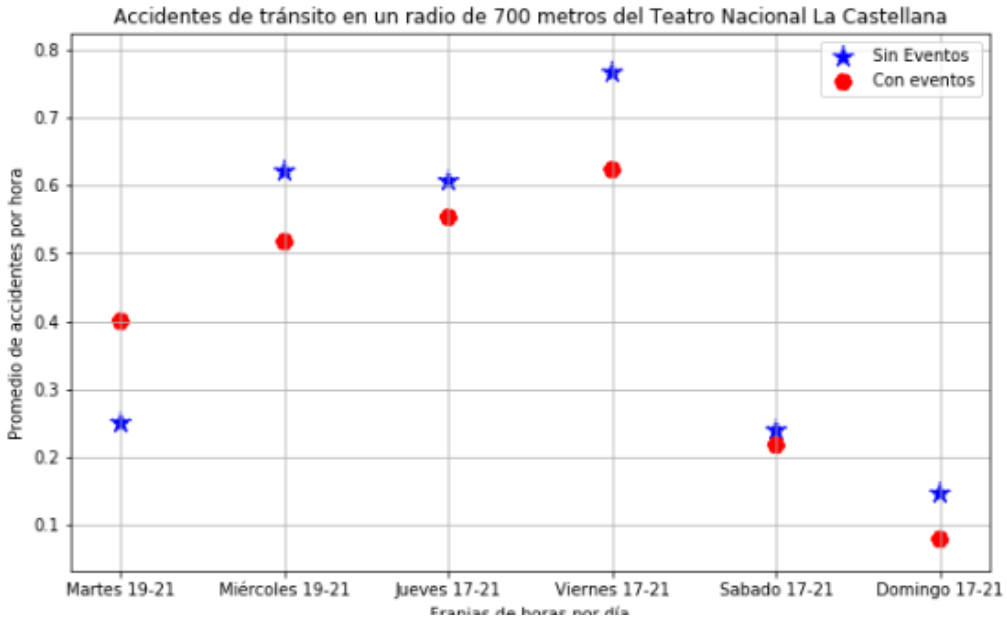
Fuente: elaboración propia.

Figura 4-36. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 500 metros.



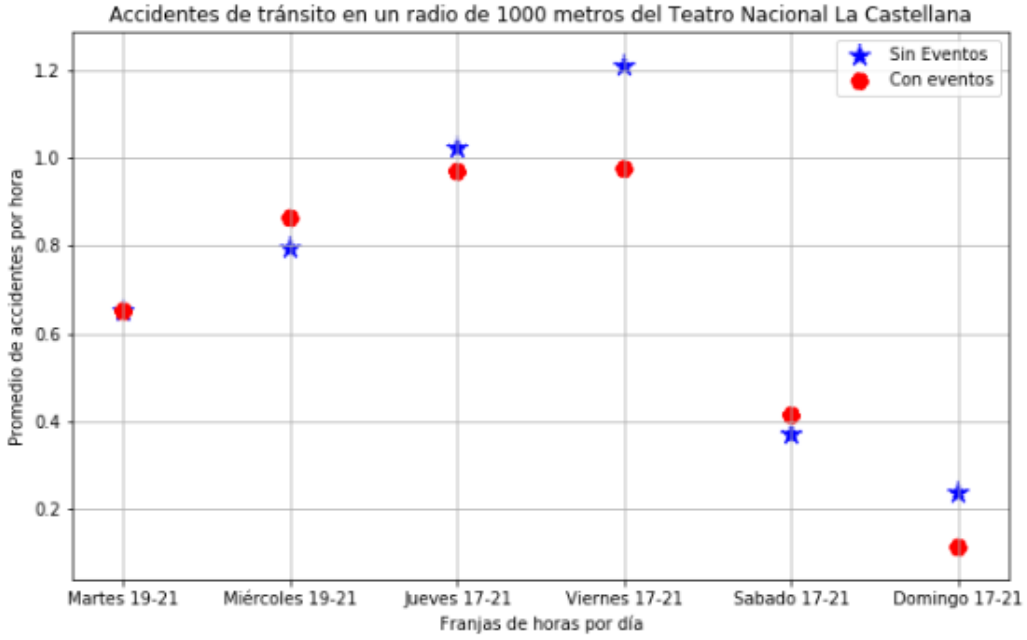
Fuente: elaboración propia.

Figura 4-37. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 700 metros.



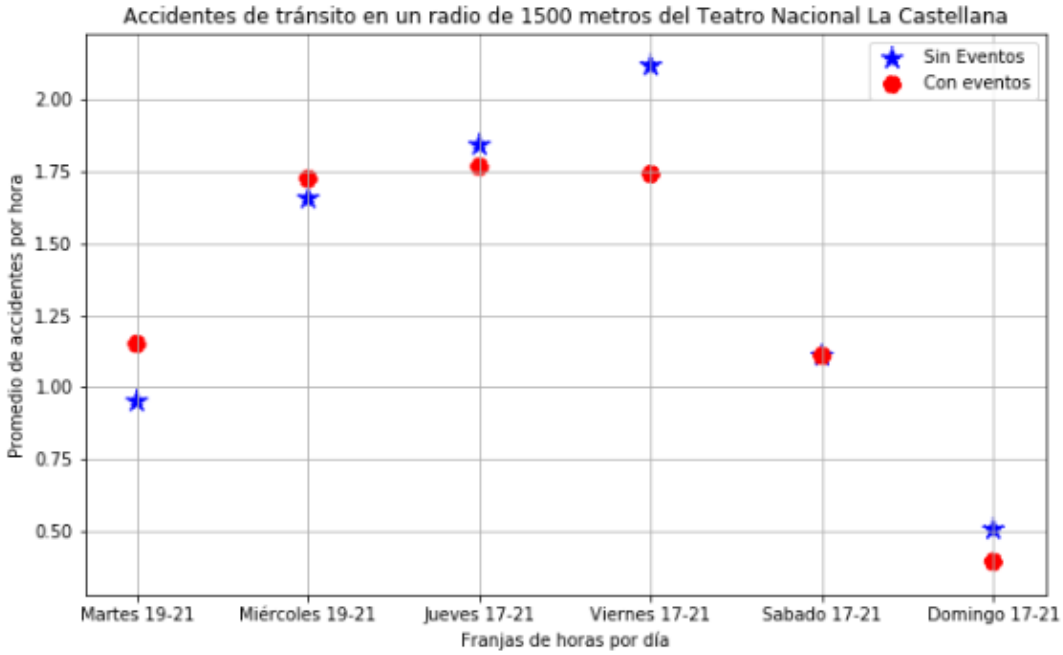
Fuente: elaboración propia.

Figura 4-38. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 1000 metros.



Fuente: elaboración propia.

Figura 4-39. Promedio de accidentes por día de la semana con y sin presencia de eventos públicos en el Teatro Nacional la Castellana en un radio de 1500 metros.



Fuente: elaboración propia.

Teniendo en cuenta que en el teatro nacional la Castellana no presenta eventos los días lunes, los días martes y miércoles son los días donde hay una variabilidad del promedio de los accidentes para cuando hay y no presencia de eventos públicos. Para los demás días la accidentalidad es mayor cuando hay ausencia de eventos públicos en las diferencias distancias.

Cabe aclarar que para algunos lugares no se presentaron gráficos del promedio de los accidentes desde la distancia de cien metros, dado que las primeras distancias cubren la zona u espacio del lugar de ocurrencia de los eventos. Las distancias son a partir del punto de referencia,

4.4 Análisis estadístico

En esta sección se realiza la descripción de los métodos estadísticos empleados para el análisis de la correlación entre los eventos públicos y los accidentes de tránsito. Al igual los resultados que se obtuvieron. Entre los métodos aplicados se encuentra la prueba de hipótesis, coeficiente de correlación de Tau de Kendall y las pruebas de rangos con signo de Wilcoxon expresando los resultados por medio de gráficos y tablas [52].

4.4.1 Descripción de métodos estadísticos

4.4.1.1 Prueba de hipótesis

Un método estadístico utilizado para comparar dos grupos de datos experimentales se denomina pruebas de hipótesis. Esta parte de la suposición sobre un parámetro de población (media y/o varianza) generando así la hipótesis que se pone a prueba (puede ser o no ser cierta) [53].

La prueba contiene un sistema de hipótesis que incluye la hipótesis nula y la alternativa. La hipótesis nula es una afirmación estadística que supone que dichas poblaciones que se están comparando son igual según el parámetro poblacional de referencia; la hipótesis alternativa responde a contradecir la hipótesis nula. En la formula (4-1) se muestra el sistema de hipótesis cuando se toma la media como referencia el parámetro poblacional.

$$\text{Hipótesis nula- es: } H_0 : \mu_A = \mu_B \qquad \text{Hipótesis alternativa es: } H_1 : \mu_A \neq \mu_B \tag{4-1}$$

Donde A y B hace referencia a las dos poblaciones de interés.

En el caso que se trabaje con un único conjunto de datos. μ_B se puede reemplazar por un valor especifico cuando lo que se busca no es comparar dos poblaciones sino una población frente a un valor dado por experto. Es decir, la hipótesis nula hace referencia a que el parámetro es igual a un valor dado, en consecuencia, la hipótesis alternativa hace referencia a que el parámetro poblacional es diferente al valor dado.

Continuando con el caso de comparación de dos poblaciones, que es el que aplica en este caso, la hipótesis nula tal como se especifica en la ecuación (4-1), lo que resume es que no hay diferencia entre las dos medias de población y las observaciones de ambos conjuntos de datos resultan puramente del azar. La hipótesis alternativa se refiere a que la media de las observaciones de cada uno de los conjuntos de datos de interés son el resultado de un efecto real, siendo influenciadas por la variable que especifica la población A de la B.

Ahora bien, la posibilidad de aceptar o rechazar una hipótesis no comprende el 100 % de precisión a lo que da a lugar al nivel de significación; el cual refiere a la probabilidad de rechazar una hipótesis nula por la prueba cuando es realmente cierta, que se denota como α y técnicamente se denomina error tipo I. El método también contempla el error tipo II (β.) que es el error de aceptar la hipótesis nula cuando en realidad es falsa, sin embargo, para este análisis se contempla a prueba que maneja el error tipo I.
Lo anterior implica que el valor p (p-value) de cada prueba se va a comparar con un nivel de significancia tolerable para la investigación y de ser menor, se rechaza la hipótesis nula. La significancia tolerable para este caso es del 10%.

Es de aclarar que, para la aplicación de las pruebas de hipótesis, se requiere cumplir el supuesto de normalidad de los datos, por lo que en este caso en particular se aplicó la prueba de Kolmogorov Smirnov para confirmar el supuesto y poder analizar los resultados arrojados por las pruebas de hipótesis.

4.4.1.2 Coeficiente de correlación de Tau de kendall

La correlación de rango de Kendall es una prueba no paramétrica que mide la fuerza de dependencia entre dos variables, evaluando las asociaciones estadísticas basadas en los rangos de los datos. El coeficiente de correlación de Kendall usa pares de observaciones y determina la fuerza de asociación basada en el patrón de concordancia y discordancia entre los pares. Siendo insensible al error, los valores de P son más precisos con tamaños de muestra más pequeños [52]. La siguiente fórmula se utiliza para calcular el valor de la correlación de rango de Kendall:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4-2}$$

Donde $n_c$es el número de concordantes y n_d el número de discordantes. Definiendo concordante como un par de observaciones $(x_2 - x_1)$ y $(y_2 - y_1)$ que tienen el mismo signo y discordante cuando tienen signos opuestos.

4.4.1.3 Pruebas de rangos con signo Wilcoxon

Es otra prueba no paramétrica de comparación de dos muestras relacionadas (en este caso son dos poblaciones que coinciden en el lugar de ocurrencia y se diferencian por el tema de eventos públicos). Esta permite probar la aleatoriedad de una secuencia de datos. La prueba compara dos medidas de centro (medianas) y determina que la diferencia no se deba al azar (que la diferencia sea estadísticamente significativa). El modelo supone que los datos que provienen de la misma población a través del tiempo o lugar. Debido a que la prueba es de índole no paramétrica, no requiere una distribución de probabilidad particular de la variable dependiente [54].

La prueba esencialmente calcula la diferencia entre cada conjunto de datos y analiza estas diferencias. La hipótesis nula de interés es:

H_0 : la diferencia entre las medianas de ambos conjuntos de datos es igual a cero.
H_1 : la diferencia entre las medianas de ambos conjuntos de datos no es igual a cero.

Bajo la hipótesis alternativa, las diferencias tenderán a ser positivas o negativas. Si H_0 es verdadero, tendería a que la mitad de las diferencias sean positivas y aproximadamente la mitad de las diferencias sean negativas. Dando a lugar, a que la suma de los rangos positivos sería aproximadamente igual a la suma de los rangos negativos: $(T+) \approx (T-) = \frac{1}{2} * n(n+1)/2$ y con esto no se rechaza la hipótesis nula planteada por el método.

En este mismo caso, la decisión se toma con base en el valor p (p-value) de cada prueba que se comparar con un nivel de significancia tolerable para la investigación y de ser menor, se rechaza la hipótesis nula. La significancia tolerable para este caso es del 10%

4.4.2 Descripción de los resultados

Teniendo clara la intención de cada una de las metodologías estadísticas más pertinentes para dar respuesta al objetivo No 2 de la investigación. Dicho en otras palabras, son metodologías que tienen como fin entregar resultados sobre la relación que puede tener la existencia de eventos públicos y la accidentalidad vial, a continuación, se presentan los resultados arrojados para cada uno de los sitios de interés.

Es de aclarar que el conjunto de datos para cada uno de los sitios responde a las observaciones hechas entre el 01/09/2018 y el 29/06/2019; los registros se dividen en dos subpoblaciones de interés que son:

- Registros que se toman en momentos en los cuales no hay presencia de evento público.

- Registros que se toman previo, durante o posterior al desarrollo de uno o más eventos públicos.

El tiempo que transcurre entre la previa, el desarrollo y los momentos de evacuación de los asistentes de cualquier evento, es de cuatro (3) horas para el teatro nacional la Castellana y el teatro Jorge Eliecer Gaitán, para el parque Simón Bolívar, el Movistar Arena y el estadio el Campín es de (4) horas.

Los resultados se describen por cada uno de los sitios, toda vez que las particularidades de ellos a nivel de ubicación, aforo, tipos de eventos, entre otros, no hace viable una comparación de resultados entre los lugares de interés.

Por cada sitio, se hace una descripción general del conjunto de datos trabajados, posteriormente se aplica el coeficiente de concordancia Tau-Kendall para tener en primera instancia una idea sobre el comportamiento de la accidentalidad vial reportada en el conjunto de datos cuando no hay presencia de eventos públicos y en el conjunto de datos cuando si hay un evento o más.

Dado que existen otras metodologías para analizar el comportamiento de la accidentalidad vial en cada una de las subpoblaciones de interés, se aplica la prueba de normalidad de Kolmogorov-Smirnov para confirmar este supuesto y tener viabilidad para la aplicación de pruebas de hipótesis y otros coeficientes de correlación paramétricos como el de Pearson. Para los casos en los que no se cumpla el supuesto de normalidad en los datos, se aplicará una metodología de pruebas de hipótesis no paramétrica (no requiere supuestos de distribución en los datos). Lo anterior busca tener soporte estadístico para analizar la diferencia entre las medianas de cada subpoblación; la mediana es una estadística de centro más robusta y que no se deja influenciar por los valores extremos.

Es importante aclarar que el conjunto de datos extraído inicialmente incluye registros sin eventos públicos por el simple hecho que son horas del día en los que difícilmente se van a desarrollar eventos de esta índole, lo que implica que se está contemplando accidentalidad que no tendrán punto de comparación en un escenario con evento público, como lo es la franja horaria entre 01:00 y 09:00. Esto quiere decir que, dentro del análisis el conjunto de datos debe reducirse a registros en los que independientemente se haya presentado un evento púbico o no, sea un momento en el que sea viable el desarrollo de

un evento público; desde el enfoque estadístico es importante que la comparación de variables busque la medición de estas en escenarios sin sesgo o con el menor sesgo posible, para que los resultados sean producto de la variabilidad propia de la variable y no respondan a comportamientos predecibles por la misma naturaleza en la que se generaron los datos.

Lo anterior implica trabajar con los siguientes conjuntos de datos:

Tabla 4-3. Conjunto de datos para el análisis estadístico

CONJUNTO DE DATOS	FINALIDAD
1. Registros inicialmente extraídos	Trabajar con los datos inicialmente extraídos y sobre los cuales se han desarrollado los análisis espaciales y temporales. *Se denomina conjunto de datos original.*
2. Registros solo generados en momentos en los que se desarrollan eventos públicos	Trabajar con datos en los que se pueda analizar el comportamiento de la accidentalidad vial en el escenario en el que más se concentra el interés de la investigación: presencia de eventos públicos (con mínimo un evento público en desarrollo). Lo que implica tener los insumos más específicos para poder responder a la siguiente inquietud: ¿qué ocurre con la accidentalidad en presencia de eventos públicos?, ¿aumenta?, ¿disminuye?, ¿no hay relación?, ¿entre más eventos públicos, más accidentalidad vial?, ¿menos?, ¿no hay comportamiento especifico? *Se denomina conjunto de datos ácido por ser el conjunto con los registros mínimos que se deben tener para responder a la pregunta de esta investigación.*
3. Registros agregados por día.	Registros a nivel de día en donde se pueden tener días sin evento público o con un evento público o más. Esto tiene como fin mitigar el riesgo de analizar datos sesgados. Por un lado, un conjunto de datos que solo cuenta con registros en un solo escenario: solo bajo la presencia de eventos públicos, (conjunto de datos No. 2). Por el otro, datos en los que hay un aumento de registros sin evento público por el hecho de ser horas inviables para que se desarrolle eventos de esta índole. *Se denomina conjunto de datos a nivel día.*

Las herramientas estadísticas antes descritas fueron aplicadas para los tres conjuntos de datos, en la descripción de éstos, se resaltan los resultados más relevantes y/o concluyentes.

4.4.2.1 Estadio Nemesio Camacho el Campín

Tal como se ha mencionado en secciones anteriores, el total de observaciones que se tienen para este lugar son 7.235, de las cuales el 2,5% son registros con evento y el 97,5% restante hacer referencia a observaciones en donde no se identificó ningún evento público.

Tabla 4-4. Resumen de los datos originales estadio el Campín.

		Cantidad de accidentes ocurridos						
		Recuento	***Mín.***	***%***	***Media***	***Mediana***	***Máx.***	***Desviación estándar***
Existencia de eventos	***No hubo eventos***	7054	0	97,5%	0,4716472	0	11	1
	Si hubo eventos	181	0	2,5%	0,3812154	0	4	1
	Total	7235	0	100%	0,4693849	0	11	0,99180836

Fuente: elaboración propia.

De la **Tabla 4-4** se concluye que hay un desequilibrio de información ya que la cantidad de reportes de momentos en los que no hay evento es considerablemente superior a la cantidad de reportes en momentos previos. Al analizar solo las medidas de centro dentro de cada conjunto se tiene que, por cada hora observada se presenta entre cero (0) y máximo un (1) accidente vial en promedio, haya o no haya presencia de evento público; la mediana en ambas subpoblaciones es cero (0) lo que significa que mínimo el 50% de los datos reporta cero (accidentes).

Teniendo en cuenta el resultado de la desviación estándar, esta referencia se puede ampliar a máximo dos (2) accidentes por hora en ambos casos.
Es de resaltar que en la hora en la que más se presentaron accidentes viales (11 accidentes), fue en un momento del día en que no se reporta ningún evento público; el

número máximo de accidentes viales que se presentaron cuando hubo evento público fueron cuatro (4).

Lo anterior entrega una conclusión preliminar sobre dos conjuntos de datos que no tienen una marcada diferencia en cuanto a la accidentalidad vial. Adicionalmente, justifica el hecho de trabajar con otros conjuntos de datos que hagan más justa la comparación.

Tabla 4-5. Prueba de correlación por Tau de Kendall en el estadio el Campín.

		Cantidad de eventos ocurridos	**Cantidad de accidentes ocurridos**
Cantidad de eventos ocurridos	***Coeficiente de correlación***	1,000	-0,006
	Sig. (bilateral)	.	0,593
	N	7235	7235
Cantidad de accidentes ocurridos	***Coeficiente de correlación***	-0,006	1,000
	Sig. (bilateral)	0,593	.
	N	7235	7235

Fuente: elaboración propia.

Ahora bien, aplicando un coeficiente de correlación no paramétrico que nos permita identificar si los datos de cantidad de accidentes viales y la cantidad de eventos ocurridos; se espera que la cantidad de accidentes sea similar, por un lado, en el conjunto de datos cuando no hay evento, y por el otro, en el conjunto de datos cuando si hay evento, pero que la cantidad de accidentes difiera entre una subpoblación y otra.

Al tener que el coeficiente es de -0.006, se concluye que no hay ninguna concordancia entre los accidentes ocurridos y los eventos públicos desarrollados. Este resultado no se puede comparar con los análisis hechos anteriormente ya que lo que se pretende con el Tau de Kendall es identificar si hay concordancia entre el número de accidentes viales y el número de eventos. Los análisis anteriores se centraban en verificar el comportamiento de la accidentalidad vial según la cercanía al sitio de influencia (**Figura 4-9**), y la accidentalidad vial en un día y un horario particular en el que se pudiese ver la diferencia con y sin evento público (**Figura 4-10**); la variable eventos públicos en los análisis anteriores se trabajó como una variable nominal (hay o no hay evento), en el tau de Kendall se toma como una variable discreta (hay un evento, dos, tres, los que se registren).

El coeficiente de Tau de Kendall se aplicó en el conjunto de datos que solo contiene registros bajo el desarrollo de eventos públicos (conjunto de datos ácido), y la conclusión no difiere a la ya obtenida al tener un coeficiente de correlación de 0.012. Dado lo anterior,

se insiste en aplicar las otras metodologías que se tienen para indagar sobre el comportamiento de la accidentalidad en ambos conjuntos de datos.

Se aplica la prueba de Kolmogorov que pretende confirmar la hipótesis nula sobre la distribución de los datos trabajados y su correspondencia con la distribución normal. No obstante, según la **Tabla 4-6**, no hay significancia estadística para confirmar esta prueba de hipótesis en el conjunto de datos ácido y por ende la conclusión es rechazarla; el valor p arrojado por la prueba (0) es menor que el umbral definido por el investigador (10% o 0.1) por lo cual se rechaza la hipótesis nula sobre la normalidad en la distribución de los datos de accidentalidad vial en el Estadio Nemesio Camacho el Campín

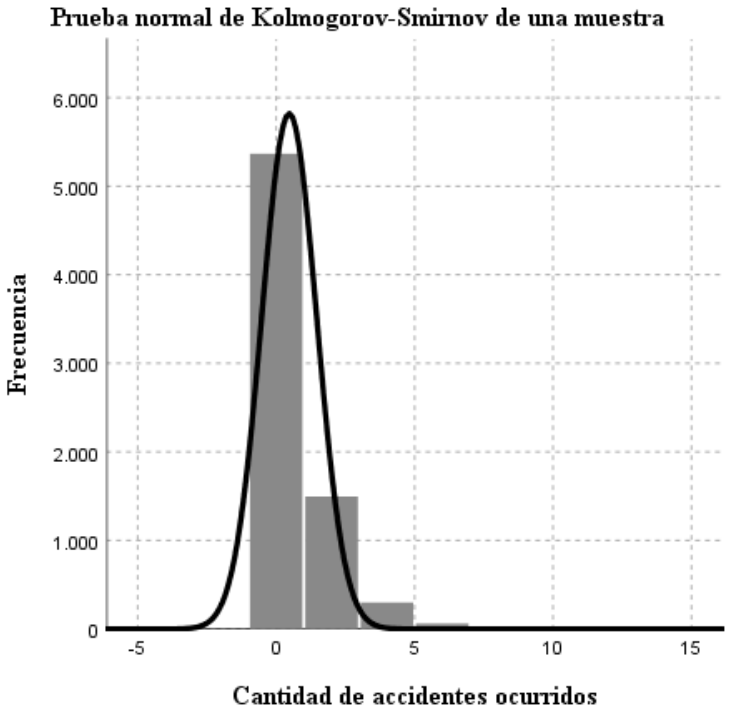
Tabla 4-6. Resultados del contraste de hipótesis frente a distribución de normalidad en los datos registrados en el estadio el Campín.

Hipótesis nula	***Prueba***	***Sig.***	***Decisión***
La distribución de Cantidad de accidentes ocurridos es normal con la media 0 y la desviación estándar ,992.	Prueba de Kolmogorov-Smirnov para una muestra	0	Rechace la hipótesis nula.

Fuente: elaboración propia.

Haciendo un análisis gráfico de los datos, se tiene que efectivamente no hay similitud con una distribución normal, ya que hay concentración de datos en los valores más bajos de accidentalidad; hay un sesgo a cantidades de accidentes viales altas. Se recuerda la que distribución normal es aquella en donde los valores más bajos ni los más altos de la variable son los de mayor frecuencia, la concentración de los datos se da en los valores promedio.

Figura 4-40. Análisis gráfico sobre la Prueba de normalidad de Kolmogorov-Smirnov de los accidentes alrededor del estadio el Campín.



Fuente: elaboración propia.

La prueba de normalidad también se aplicó al conjunto de datos original y dada la conclusión para el conjunto de datos ácido, allí también se rechaza distribución gaussiana en los datos.

Ahora bien, habiendo rechazado el supuesto de normalidad, estadísticamente es inviable aplicar pruebas de hipótesis paramétricas que pongan a discusión la igualdad de parámetros como la media y/o la varianza; también se hace inviable calcular coeficientes de correlación como el de Pearson. La última opción que se tiene, es aplicar una prueba no paramétrica que compare las medianas y que permita saber si la diferencia entre dichas medianas es cero o no; en el caso que sea cero implica que el comportamiento de los accidentes viales es similar haya o no haya presencia de eventos públicos. En este caso se aplica para el conjunto de datos agregados por día para eliminar el sesgo de registros sin evento público por ser inviable el desarrollo de este tipo de eventos en horas no apropiadas.

Tabla 4-7. Prueba de suma Wilcoxon Rank de correlación de continuidad en el estadio el Campín.

W	*p-value*	*Hipótesis alternativa*
6551,5	0.01968	Decisión: rechazar el hecho que las medianas de ambos conjuntos de datos son iguales.

Fuente: elaboración propia.

La **Tabla 4-4**, reporta que el p valor es menor que 0.1, lo que significa que hay significancia estadística para asumir que la mediana de accidentes en el conjunto de datos cuando no hay eventos públicos es diferente a la mediana de accidentes que se dan en el conjunto de datos cuando hay eventos públicos.
Lo que significa que si bien no se puede concluir que a medida que hay más eventos públicos más (o menos) accidentalidad vial se genera. En el Campín se puede evidenciar una leve diferencia entre la accidentalidad que representa el conjunto de registros cuando no hay evento que la que representa el conjunto de registros que se tienen cuando si hay eventos.

Se dice que es leve por las conclusiones preliminares que se tuvieron al analizar los resultados desde un enfoque descriptivo, sin embargo, en los resultados de algunos buffers (300, 400, 500 y 1000 metros de buffer) sí se evidenció que el comportamiento de los accidentes era totalmente contrario enfrentando ambos escenarios, haciendo esto que el parámetro central difiera.
No se aplica la prueba de Wilcoxon para los otros dos conjuntos de datos porque la predominancia de registros con cero eventos en el caso de los datos originales implica sesgo en la prueba; para el caso del conjunto de datos ácido la inexistencia de registros con eventos cero, genera inconsistencia en la prueba.

4.4.2.2 Movistar Arena

La cantidad de observaciones registradas para este lugar son 7.235, las cuales están distribuidas en un 99% de observaciones las cuales no se desarrollaron eventos públicos y un 1% en las que sí.

La cantidad de accidentes observados en las dos subpoblaciones de interés oscila entre cero (0) y máximo dos (2) accidentes, lo que implica que en esta primera instancia exploratoria no hay evidencia que el comportamiento de los accidentes es diferente en los momentos cuando no se desarrollan eventos públicos que cuando sí se desarrollan.

Tabla 4-8. Resumen descriptivo de los datos originales registrados para el Movistar Arena.

		Cantidad de accidentes ocurridos						
		Recuento	***%***	***Mín.***	***Media***	***Mediana***	***Máx.***	***Desviación estándar***
Existencia de eventos	***No hubo eventos***	7166	99 %	0	0,5122802	0	12	1,057
	Si hubo eventos	69	1 %	0	0,4492753	0	4	0,900
	Total	7235	100%	0	0,5116793	0	12	1,056

Fuente: elaboración propia.

La mediana en ambas subpoblaciones es cero (0) lo que significa que mínimo el 50% de los datos reporta cero (accidentes). Lo anterior justifica el hecho de trabajar con otros conjuntos de datos que hagan más justa la comparación. Ahora bien, dando un paso adicional para identificar el comportamiento de los accidentes viales y la cantidad de evento públicos, se aplica el coeficiente de correlación de Tau de Kendall. El resultado es de -0.001, lo que significa que no hay soporte para afirmar que hay concordancia entre los valores de la cantidad de eventos ocurridos y los accidentes viales.

Tabla 4-9. Coeficiente de correlación Tau de Kendall en el Movistar Arena.

		Cantidad de eventos ocurridos	**Cantidad de accidentes ocurridos**
Cantidad de eventos ocurridos	***Coeficiente de correlación***	1	-0,001
	Sig. (bilateral)	0	0,947
	N	7235	7235
Cantidad de accidentes ocurridos	***Coeficiente de correlación***	-0,001	1
	Sig. (bilateral)	0,947	0
	N	7235	7235

Fuente: elaboración propia.

El coeficiente de Tau de Kendall también se aplicó al conjunto de datos ácido y la conclusión no difiere.

Otra metodología estadística relacionada con el análisis de una variable en dos conjuntos de datos generados por otra variable temáticamente relacionada (presencia de eventos públicos), son las pruebas de hipótesis; la más usada es la prueba que compara la media entre las dos subpoblaciones de interés.

No obstante, y como ya se había mencionado anteriormente, esta metodología parte del supuesto de una distribución normal entre los datos, por lo cual, debe aplicarse una prueba para garantizar este supuesto. La prueba de Kolmogorov-Smirnov para el caso del conjunto de datos ácido, rechaza la distribución normal, lo que implica que no hay viabilidad para utilizar metodologías paramétricas; el valor p de la prueba resulta menor que el nivel de significancia del 10% tolerable por la investigación.

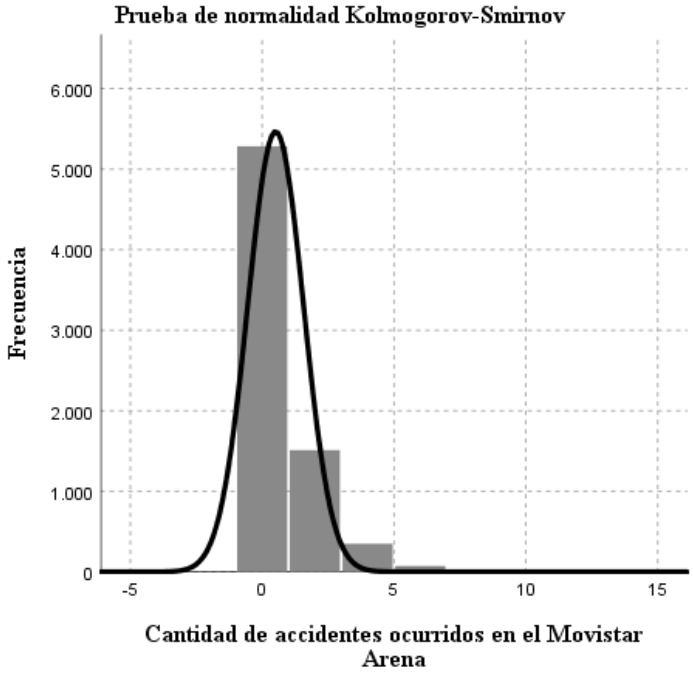
Tabla 4-10. Resultados del contraste de hipótesis frente a distribución de normalidad en los datos registrados en el Movistar Arena.

Hipótesis nula	***Prueba***	***Valor p***	***Decisión***
La distribución de Cantidad de accidentes ocurridos es normal con la media 0 y la desviación estándar 0,823.	Prueba de Kolmogorov-Smirnov para una muestra	0	Rechace la hipótesis nula.

Fuente: elaboración propia.

Acudiendo a un análisis gráfico de la distribución del conjunto de observaciones extraído para la zona del Movistar Arena, se confirma el rechazo de la normalidad de la variable cantidad de accidentes viales, ya que hay una concentración de éstos en los datos más pequeños y un sesgo en los valores más altos. Estando en contra de la distribución gaussiana; esta distribución teórica implica una baja cantidad de datos en los valores extremos y una concentración en los valores centrales.

Figura 4-41. Análisis gráfico sobre la Prueba de normalidad de Kolmogorov-Smirnov de los accidentes alrededor del Movistar Arena.



Fuente: elaboración propia.

La prueba de normalidad también se aplicó al conjunto de datos original y dada la conclusión para el conjunto de datos ácido, también se rechaza distribución gaussiana en los datos.

Para finalizar, se aplica una prueba de hipótesis no paramétrica denominada prueba de rangos Wilcoxon, la cual pretende verificar si la diferencia entre la mediana del conjunto de datos cuando se desarrollan eventos públicos y el conjunto de datos cuando no, es cero, implicando esto que no hay diferencia entre la cantidad de accidentes viales que puede representar a ambas subpoblaciones.

Los resultados de la prueba aplicados en el conjunto de datos agregados por día, concluyen que no hay significancia estadística para rechazar la diferencia entre medianas de las subpoblaciones de interés; el valor p arrojado por la prueba no es menor a la significancia tolerable (10%).

Tabla 4-11. Prueba de los rangos con signo de Wilcoxon en el Movistar Arena.

W	*p-value*	*Hipótesis alternativa*
2586	0.3771	Decisión: las medianas son iguales

Fuente: elaboración propia

Es de aclarar que no se aplica la prueba de Wilcoxon para los otros dos conjuntos de datos porque la predominancia de registros con cero eventos en el caso de los datos originales implica sesgo en la prueba; para el caso del conjunto de datos ácido la inexistencia de registros con eventos cero, genera inconsistencia en la prueba.

4.4.2.3 Teatro Nacional La Castellana

Para este sitio de concentración de público se tiene que, del total de observaciones un 8.5% hacen referencia a registros bajo el desarrollo de eventos públicos. Adicionalmente, se tiene que en las observaciones realizadas con o sin presencia de eventos públicos la cantidad de accidentes ocurridos esta entre cero (0) y máximo de dos (2) accidentes viales aproximadamente; la mediana en ambas subpoblaciones es cero (0) lo que significa que mínimo el 50% de los datos reporta cero (accidentes).
En este sitio también se justifica trabajar con otros conjuntos de datos con menor sesgo en sus observaciones.

Tabla 4-12. Resumen descriptivo de los datos originales registrados para el teatro Nacional la Castellana

		Cantidad de accidentes ocurridos						
		Recuento	***%***	***Min.***	***Media***	***Mediana***	***Máx.***	***Desviación estándar***
Existencia de eventos	***No hubo eventos***	6634	91,5%	0	0,35649683	0	7	0,857
	Si hubo eventos	615	8.5%	0	0,45853658	0	5	0,802
	Total	7249	100%	0	0,36515381	0	7	0,861

Fuente: elaboración propia.

Analizando la relación entre los accidentes viales y el desarrollo de eventos públicos, se aplica el coeficiente de Tau de Kendall, el cual desde un enfoque no paramétrico (no requiere que los datos analizados cuenten con una distribución en específico), aporta

evidencia para concluir sobre concordancia entre los datos, que para este caso son los que arroja la variable de cantidad de eventos públicos y los que arroja la variable de cantidad de accidentes viales en el conjunto de datos original. Como resultado se tiene que el coeficiente es de 0.045, lo que significa que no hay soporte estadístico para aducir a una relación entre las variables de interés.

Tabla 4-13. Coeficiente de correlación Tau de Kendall en el teatro Nacional la Castellana.

		Cantidad de eventos ocurridos	**Cantidad de accidentes ocurridos**
Cantidad de eventos ocurridos	***Coeficiente de correlación***	1	0,045
	Sig. (bilateral)	.	0,001
	N	7249	7249
Cantidad de accidentes ocurridos	***Coeficiente de correlación***	0,045	1
	Sig. (bilateral)	0,001	.
	N	7249	7249

Fuente: elaboración propia.

Se calcula el coeficiente de Tau de Kendall en el conjunto de datos ácido y la conclusión no difiere, al tener un valor de 0.038. Ahora bien, dado que existen otras pruebas para confirmar la relación, se hace necesario aplicar una prueba de normalidad para saber si es viable la aplicación de pruebas paramétricas. En este caso se aplica la prueba de Kolmogorov-Smirnov en el conjunto de datos ácido y se tiene que no hay significancia estadística para afirmar normalidad en los datos de análisis, por lo cual se genera imposibilidad de aplicar pruebas de hipótesis paramétricas, así como el cálculo de coeficientes de correlación de Pearson.

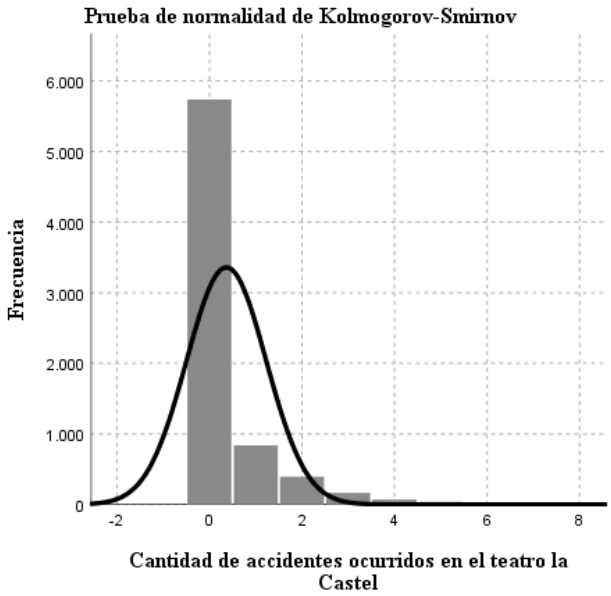
Tabla 4-14. Resultados del contraste de hipótesis frente a distribución de normalidad en los datos registrados en el teatro Nacional la Castellana.

Hipótesis nula	**Prueba**	**Valor p**	**Decisión**
La distribución de Cantidad de accidentes ocurridos es normal con la media 0 y la desviación estándar ,702.	Prueba de Kolmogorov-Smirnov para una muestra	,000[a]	Rechace la hipótesis nula.

Fuente: elaboración propia.

El rechazo de la hipótesis de normalidad en los datos es evidente al analizar gráficamente la distribución de los datos: hay una concentración de los valores en el extremo inferior estando en contravía de lo que implica tener una distribución Gaussiana, que es que no hay concentración de los datos en los valores más grandes y los más pequeños de las variables y sí en los valores centrales.

Figura 4-42. Análisis gráfico sobre la Prueba de normalidad de Kolmogorov-Smirnov de los accidentes alrededor del teatro Nacional la Castellana.



Fuente: elaboración propia.

Aun cuando se rechaza la prueba de normalidad en el conjunto de datos más exigente, se aplica para el conjunto de datos originales y se reitera que no hay sustento estadístico para confirmar la distribución gaussiana en este caso.

Dado lo anterior, resta aplicar una prueba de hipótesis en la cual ponga a juzgar la afirmación de diferencia entre medianas de accidentes viales con y sin presencia de evento público. Para este caso el umbral con el que se rechaza la hipótesis 0,1 (p-valor < significancia teórica 0,1) está muy cercano al valor arrojado por la prueba 0,1034, por lo cual no se puede afirmar de manera estricta que la diferencia de las medianas no está influenciada por el conjunto de datos en el que se encuentre (con evento o sin evento público).

Tabla 4-15. Prueba de los rangos con signo de Wilcoxon en el teatro Nacional la Castellana.

W	*p-value*	*Hipótesis alternativa*
10131	0.1034	Decisión: las medianas son iguales

Fuente: elaboración propia.

En este caso lo que se debe hacer es seguir indagando con otros mecanismos más complejos para confirmar el comportamiento diferencial de los accidentes viales. Dado el alcance del estudio, no se profundiza al respecto, sin embargo, el método de extracción descrito en este trabajo de investigación será una herramienta clave para continuar con la profundización del comportamiento de estas variables en este lugar.

Se reitera que no se aplica la prueba de Wilcoxon para los otros dos conjuntos de datos porque la predominancia de registros con cero eventos en el caso de los datos originales implica que prueba se sesgue; para el caso del conjunto de datos ácido la inexistencia de registros con eventos cero, genera inconsistencia en la prueba.

4.4.2.4 Parque Metropolitano Simón Bolívar

En cuanto a las observaciones extraídas de los alrededores del parque Simón Bolívar, se tiene que en el 0,6% se desarrollaron eventos públicos. Calculando un intervalo en el que oscilan los accidentes viales en ambas subpoblaciones, se tiene que en promedio se registran entre cero (0) y tres (3) accidentes cuando hay eventos públicos y entre cero (0) y dos (2) accidentes cuando no hay eventos públicos.

Adicionalmente, se resalta que la mayor cantidad de accidentes se da en una observación que no reporta eventos públicos.

Tabla 4-16. Resumen descriptivo de los datos originales registrados para el Parque Simón Bolívar.

		Cantidad de accidentes ocurridos						
		Recuento	***%***	***Mín.***	***Media***	***Mediana***	***Máx.***	***Desviación estándar***
Existencia de eventos	***No hubo eventos***	7215	99,4%	0	0,7259	0	11	1,317
	Si hubo eventos	47	0,6%	0	0,7021	0	9	1,614
	Total	762	100%	0	0, 7258	0	11	1,319

Fuente: elaboración propia.

Lo anterior evidencia una leva diferencia entre la cantidad de accidentes entre una subpoblación y otra. Sin embargo, estos datos son a nivel exploratorio y desde lo estadístico requieren un soporte mayor para confirmar el hecho.

Dado lo anterior, se aplica un coeficiente de correlación no paramétrico para seguir indagando sobre el comportamiento de la accidentalidad vial y el desarrollo de eventos públicos. Tau de Kendal arroja un valor de -0.007, lo cual implica que no hay soporte estadístico para confirmar el hecho de una concordancia de los datos de eventos públicos y cantidad de accidentes viales en la zona de influencia del lugar del evento.

Tabla 4-17. Coeficiente de correlación Tau de Kendall en el Parque Metropolitano Simón Bolivar.

		Cantidad de eventos ocurridos	**Cantidad de accidentes ocurridos**
Cantidad de eventos ocurridos	***Coeficiente de correlación***	1	-0,007
	Sig. (bilateral)	.	0,521
	N	7262	7262
Cantidad de accidentes ocurridos	***Coeficiente de correlación***	-0,007	1
	Sig. (bilateral)	0,521	.
	N	7262	7262

Fuente: elaboración propia.

Se calcula el coeficiente de Tau de Kendall en el conjunto de datos ácido y la conclusión no difiere. Dado que hay otras pruebas más robustas por basarse en desarrollos teóricos de tradición, se aplica la prueba de normalidad para dar viabilidad a metodologías paramétricas que también son pertinentes para perseguir el objetivo planteado. La prueba aplicada se denomina Kolmogorov-Smirnov, pero para este caso se rechaza la hipótesis que hace referencia a que los datos de análisis se distribuyen normalmente, lo cual imposibilita la aplicación de pruebas de hipótesis paramétricas y el cálculo de coeficientes de correlación de Pearson, entre otras.

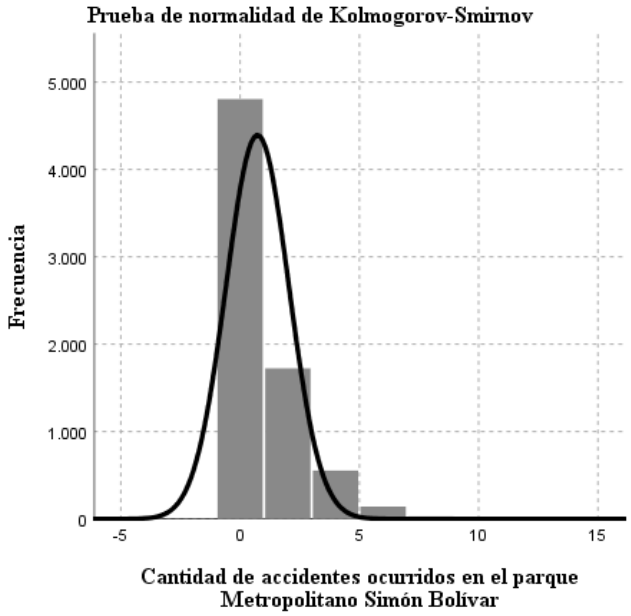
Tabla 4-18. Resultados del contraste de hipótesis frente a distribución de normalidad en los datos registrados en el parque Simón Bolívar.

Hipótesis nula	Prueba	Valor p	Decisión
La distribución de Cantidad de accidentes ocurridos es normal con la media 1 y la desviación estándar 1,779.	Prueba de Kolmogorov-Smirnov para una muestra	,000[a]	Rechace la hipótesis nula.

Fuente: elaboración propia.

Los resultados de la prueba se confirman desde un análisis gráfico toda vez que la concentración de los datos no está en los valores centrales sino en los valores más bajos de la variable. Lo cual va en contradicción de la justificación de una distribución Gaussiana.

Figura 4-43. Análisis gráfico sobre la Prueba de normalidad de Kolmogorov-Smirnov de los accidentes alrededor del Parque Simón Bolívar.



Fuente: elaboración propia.

Lo anterior implica contemplar pruebas no paramétricas como lo son la prueba de Rangos de Wilcoxon. En este caso se rechaza la hipótesis de que la diferencia entre las medianas de las dos subpoblaciones de interés es cero (0). Lo que implica que no hay significancia estadística para afirmar que la accidentalidad vial tiene comportamiento diferencial en el conjunto de datos con desarrollo de eventos públicos y el que no contempla ningún evento de esta índole.

Tabla 4-19. Prueba de los rangos con signo de Wilcoxon en el Parque Metropolitano Simón Bolívar.

W	***p-value***	***Hipótesis alternativa***
1246,5	0.3575	Decisión: las medianas son iguales

Fuente: elaboración propia.

Lo anterior, tiene concordancia con un resultado que se generó desde el enfoque exploratorio en el que se observaba que el 50% de los datos tanto en presencia de evento público como sin ella, reportan cero (0) accidentes viales.

4.4.2.5 Teatro Jorge Eliécer Gaitán

Para este sitio de eventos de entretenimiento se tiene que el 3,2% de las observaciones extraídas reportan eventos públicos desarrollados. Se tiene que el escenario en el que se dieron la mayor cantidad de accidentes responde a un momento en el que no se estaba desarrollando ningún evento.

Para dar una referencia del comportamiento de los accidentes incluyendo el parámetro de centralidad y de variabilidad, se tiene que se generan entre cero (0) y un accidente (1) vial sin evidencia que esta situación sea diferente cuando si se presenten eventos públicos.

Tabla 4-20. Resumen descriptivo de los datos originales registrados para el teatro Jorge Eliécer Gaitán.

		Cantidad de accidentes ocurridos						
		Recuento	***%***	***Min.***	***Media***	***Mediana***	***Máx.***	***Desviación estándar***
Existencia de eventos	***No hubo eventos***	6991	96,8%	0	0,264482	0	8	0,710
	Si hubo eventos	228	3,2%	0	0,223684	0	4	0,635
	Total	7219	100%	0	0,263194	0	8	0,707

Fuente: elaboración propia.

Con el fin de ampliar la indagación sobre la relación entre el desarrollo de eventos públicos y la accidentalidad vial, se aplica un coeficiente de correlación no paramétrico que da cuenta sobre la concordancia entre la cantidad de accidentes viales y la cantidad de eventos públicos desarrollados. Bajo el coeficiente de correlación de -0.008 muestra que no hay suficiencia estadística para soportar que hay una concordancia entre la cantidad de eventos y accidentes ocurridos.

Tabla 4-21. Coeficiente de correlación Tau de Kendall en el teatro Jorge Eliécer Gaitán.

		Cantidad de eventos ocurridos	Cantidad de accidentes ocurridos
Cantidad de eventos ocurridos	***Coeficiente de correlación***	1	-0,008
	Sig. (bilateral)	.	0,481
	N	7219	7219
Cantidad de accidentes ocurridos	***Coeficiente de correlación***	-0,008	1
	Sig. (bilateral)	0,481	.
	N	7219	7219

Fuente: elaboración propia.

Dado que hay otras pruebas que permiten confirmar si hay relación entre dos variables, así como tener insumos para analizar el comportamiento de una sola variable en dos escenarios distintos, se requiere confirmar la distribución de los datos de análisis. Para ello se aplica la prueba de normalidad Kolmogorov-Smirnov. En este caso, los resultados arrojan que no hay significancia estadística para afirmar la normalidad de los datos analizados.

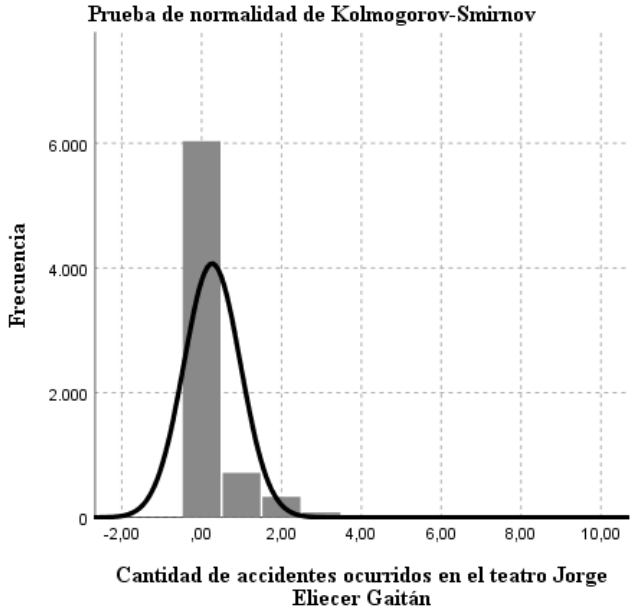
Tabla 4-22. Resultados del contraste de hipótesis frente a distribución de normalidad en los datos registrados en el teatro Jorge Eliécer Gaitán.

Hipótesis nula	*Prueba*	*Valor p*	*Decisión*
La distribución de Cantidad de accidentes ocurridos es normal con la media 1 y la desviación estándar 1,168.	Prueba de Kolmogorov-Smirnov para una muestra	,000[a]	Rechace la hipótesis nula.

Fuente: elaboración propia.

Al igual que en los lugares antes analizados, en las observaciones alrededor del Teatro Jorge Eliecer Gaitán se tiene que la cantidad de accidentes viales se concentra en los valores menores de la variable y no en los valores centrales de la misma, estando en contravía de lo que significa una distribución de datos Gaussiana.

Figura 4-44. Análisis gráfico sobre la Prueba de normalidad de Kolmogorov-Smirnov de los accidentes alrededor del teatro Jorge Eliécer Gaitán.



Fuente: elaboración propia.

Dado lo anterior, se aplica una prueba de hipótesis no paramétrica que busca significancia estadística para afirmar que la diferencia entre medianas de dos subpoblaciones no es cero (0), es decir, que dos subpoblaciones son diferentes en cuanto al valor representante de los datos. No obstante, para este caso la prueba no tiene significancia estadística para rechazar la hipótesis, lo que implica que no hay diferencia entre el comportamiento de la cantidad de accidentes viales en la ocurrencia o no de eventos públicos.

Tabla 4-23. Prueba de los rangos con signo de Wilcoxon en el Jorge Eliecer Gaitán

W	***p-value***	***Hipótesis alternativa***
8893	0.2489	Decisión: las medianas son iguales

Fuente: elaboración propia.

5. Análisis conglomerados

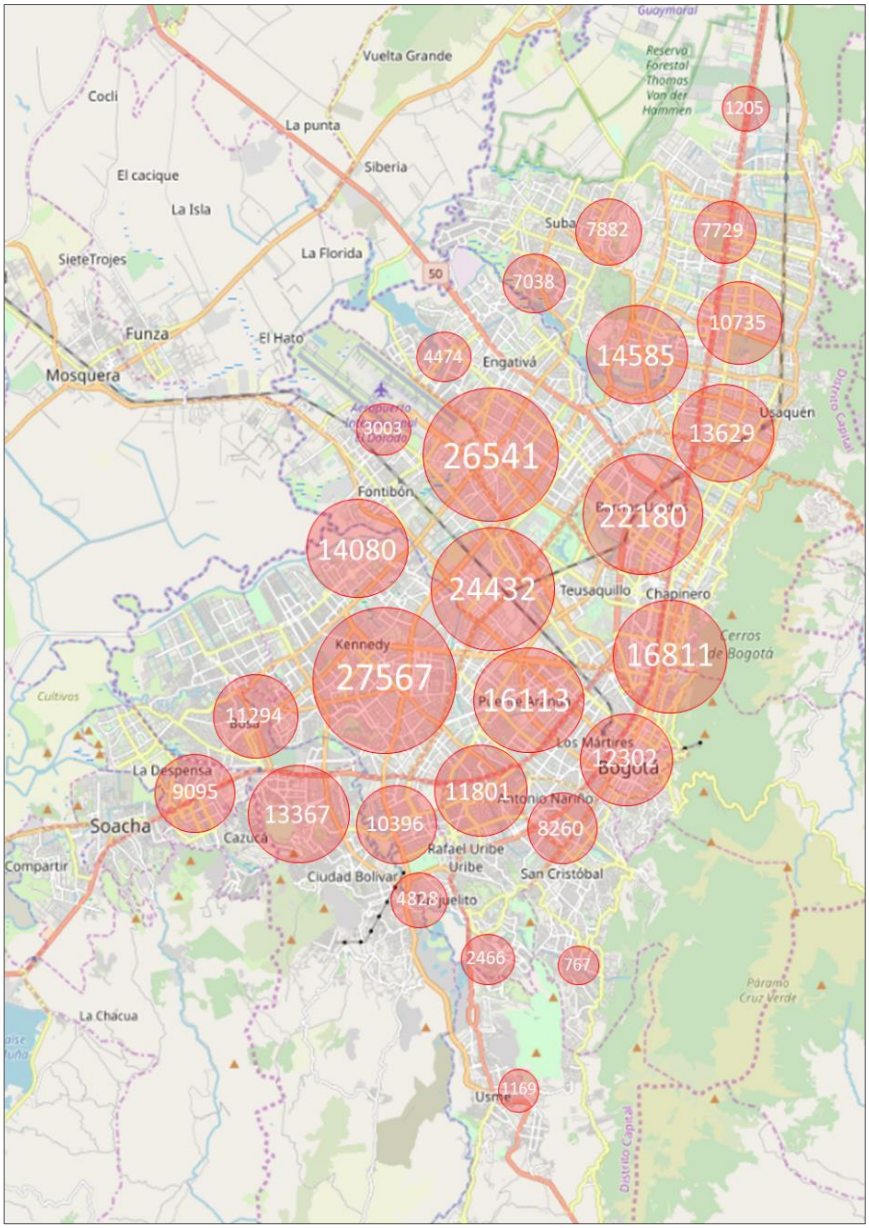
En este capítulo se realiza un análisis espacial geoestadístico de los eventos públicos y los accidentes de tránsito. Se emplea técnicas de agrupación, al igual conteo de puntos en polígonos para el análisis de patrones y índices de autocorrelación espacial. Con el objetivo de Identificar patrones de agrupación o dispersión de los eventos accidentales y públicos en las diferentes zonas y puntos críticos de la ciudad de Bogotá.

5.1 Agrupación de puntos espaciales

Cuando se tiene una capa vectorial en la que se tiene miles de puntos es útil utilizar marcadores que agrupen los puntos, el cual permita visualizar elementos de interés. En este caso de estudio ver que los puntos se agrupan en conglomerados en zonas concretas del espacio. El análisis de conglomerados de los puntos espaciales (accidentes de tránsito) de agrupamiento jerárquico se basa en la distancia existente entre ellos. En la **Figura 5-1**, se muestra los 27 clúster obtenidos por medio del algoritmo *DBSCAN Clustering,* con el respectivo número de accidentes de cada clúster durante el periodo de septiembre de 2018 a julio de 2019. El algoritmo *DBSCAN clustering* agrupa las entidades punto en cierto número de clústeres. Cada entidad punto se asigna al clúster cuyo centroide le sea más cercano.

la identificación de clústeres no se basa en un rango de distancias o de proximidad espacial, sino en función de su *proximidad estadística*, el cual permite identificar clústeres con una determinada significación estadística.

Figura 5-1. Agrupación de accidentes de tránsito

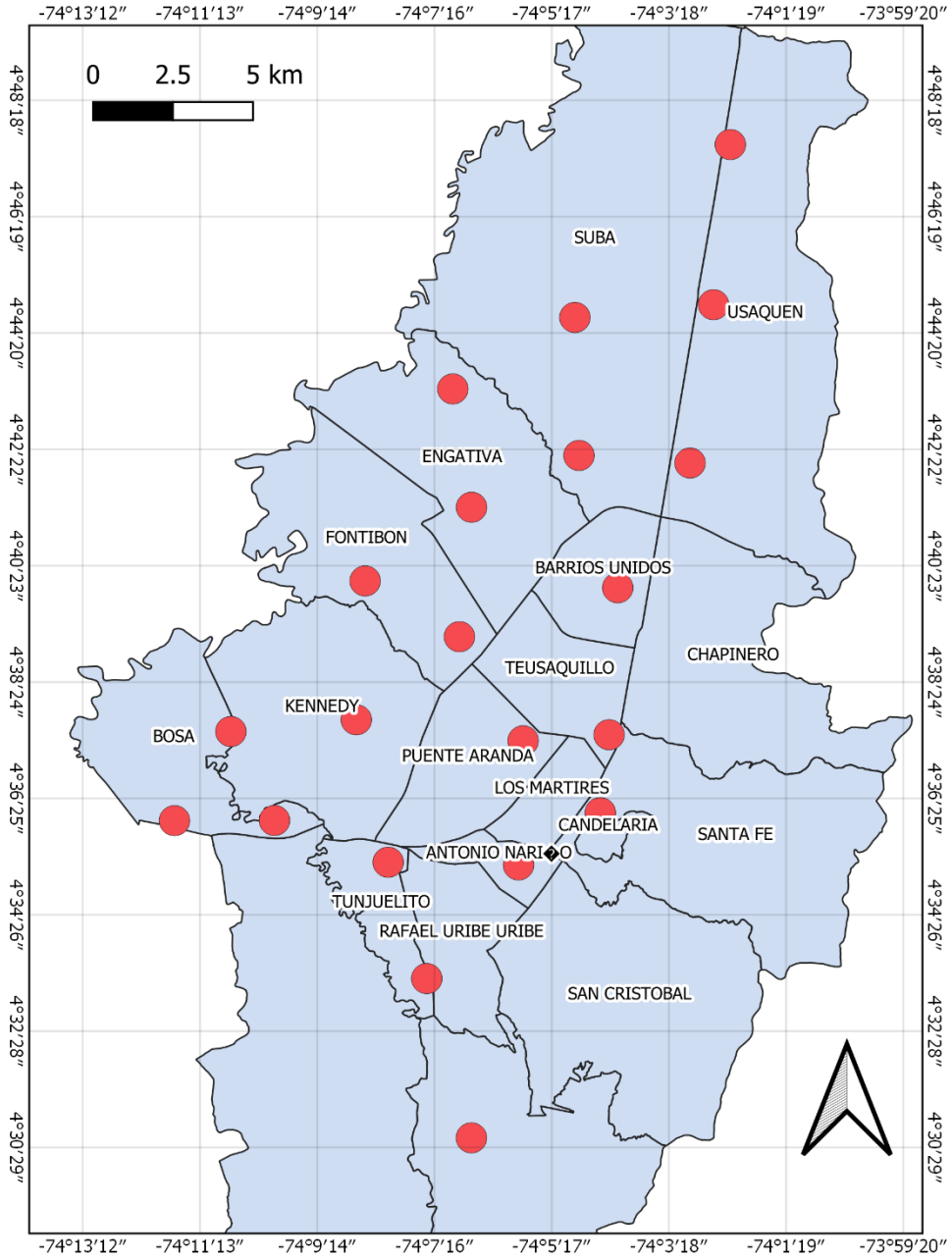


Fuente: elaboración propia.

De acuerdo con la **Figura 5-1**, es evidente un gran número de accidentes de tránsito en la zona centro y occidente de Bogotá. Existen algunas diferencias en cuanto a la densidad de su distribución en la ciudad. Por ejemplo, se observa una densidad significativa en la zona occidental en las localidades de Engativá, Fontibón y Kennedy. Es especial la localidad de Kennedy donde existe la mayor concentración con 27567 accidentes de

septiembre de 2018 y julio de 2019. Asimismo, en la zona centro, el cual abarca a la localidad de Chapinero, Teusaquillo y Barrios unidos donde se tienen grupos entre los 16100 y 22100 accidentes de tránsito. Estos se le agrega una baja densidad en las zonas sur y norte de la ciudad. En la **Figura 5-2,** se muestra la ubicación de los centroides de cada una de las agrupaciones en las diferentes localidades de Bogotá.

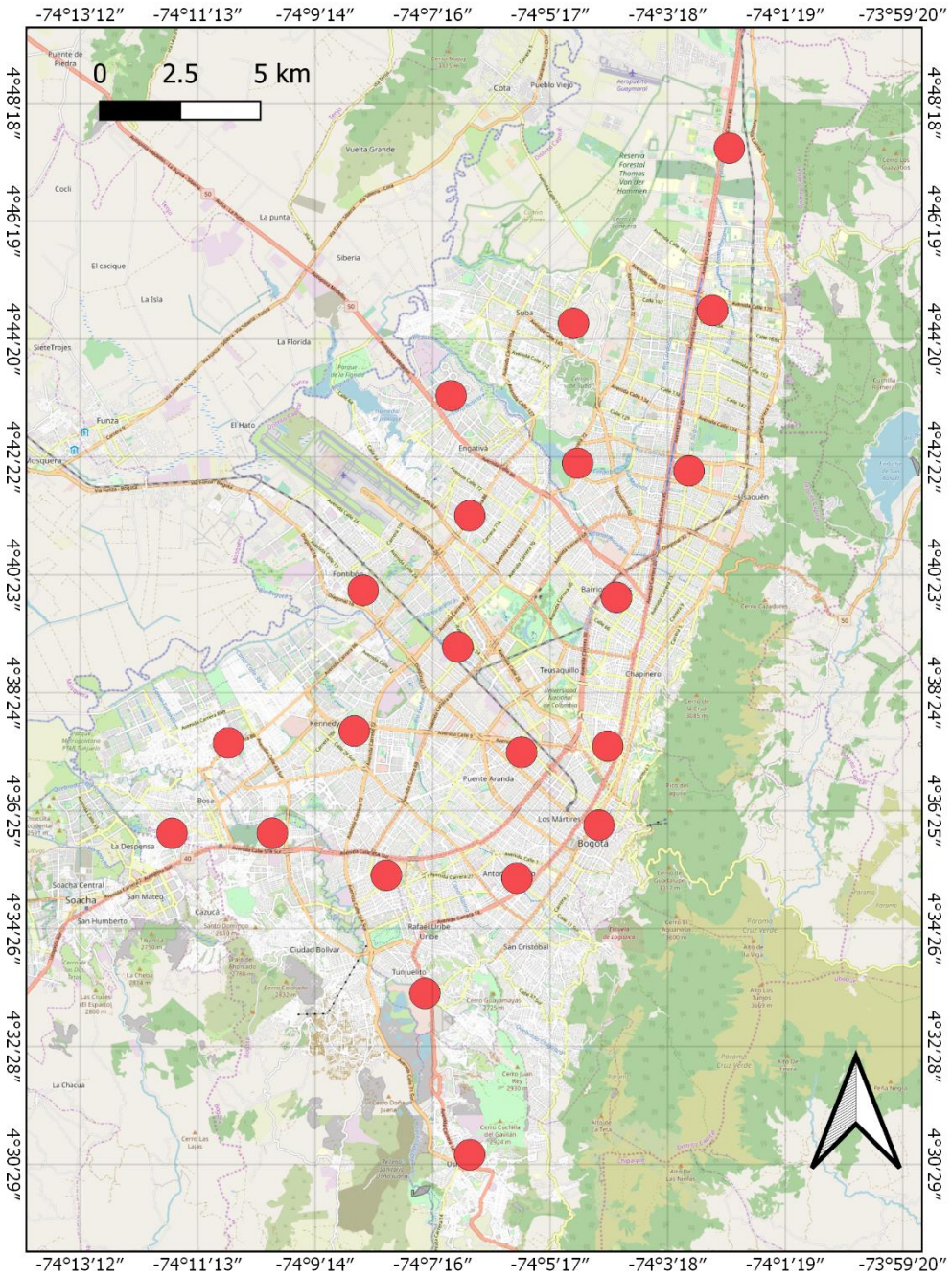
Figura 5-2. Centroides de los grupos conglomerados de accidentes de tránsito por localidades.



Fuente: elaboración propia.

El centroide de cada grupo o clúster es a su vez el promedio de las posiciones de todos los puntos dentro del clúster. En la **Figura 5-3**, se muestra el centroide de cada uno de los conglomerados jerárquicos, identificando la ubicación con la mayor concentración de accidentes.

Figura 5-3. Centroides de los grupos conglomerados de accidentes de tránsito en la ciudad de Bogotá.



Fuente: elaboración propia.

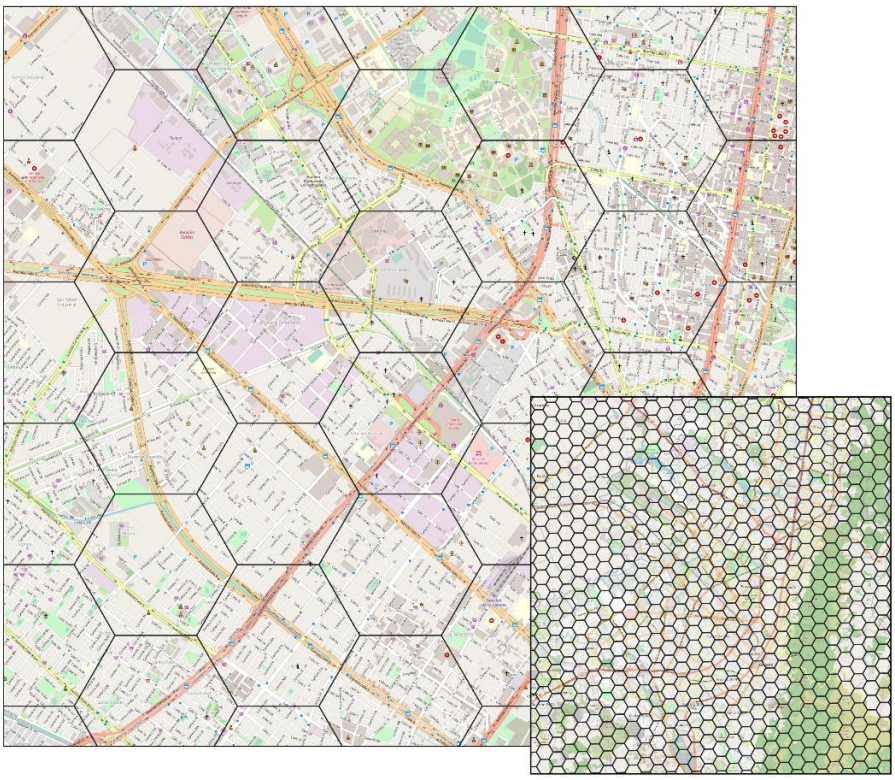
Teniendo en cuenta a las densidades agrupaciones presentadas en la **Figura 5-1**, se observa que estas se concentran en las avenidas principales. Partiendo de los centroides de la zona occidental el cual existe mayor accidentalidad, estos en gran parte se ubican cerca o en las intersecciones de las grandes avenidas. Por ejemplo, en el que se tiene

mayor concentración se ubica en la intersección de la avenida Boyacá con la avenida calle 6, otro punto se ubica entre las intersecciones de la avenida calle 26 y las avenidas Boyacá y avenida carrera 68. Otro punto de gran impacto es la avenida Ciudad de Cali a la altura de la calle 72.

5.2 Polígonos hexagonales

Con el objetivo de mostrar densidades de puntos, se crea una malla de hexágonos para el conteo del número de eventos públicos y accidentes de tránsito dentro de cada hexágono. Información utilizada como valor de agregación para posteriores análisis. La malla hexagonal, brinda una estructura de teselas más natural que la red de cuadrícula, con una mayor cobertura y continuidad espacial. En la **Figura 5-4**, se muestra el polígono de una malla hexagonal regular.

Figura 5-4. Malla hexagonal regular.



Fuente: elaboración propia.

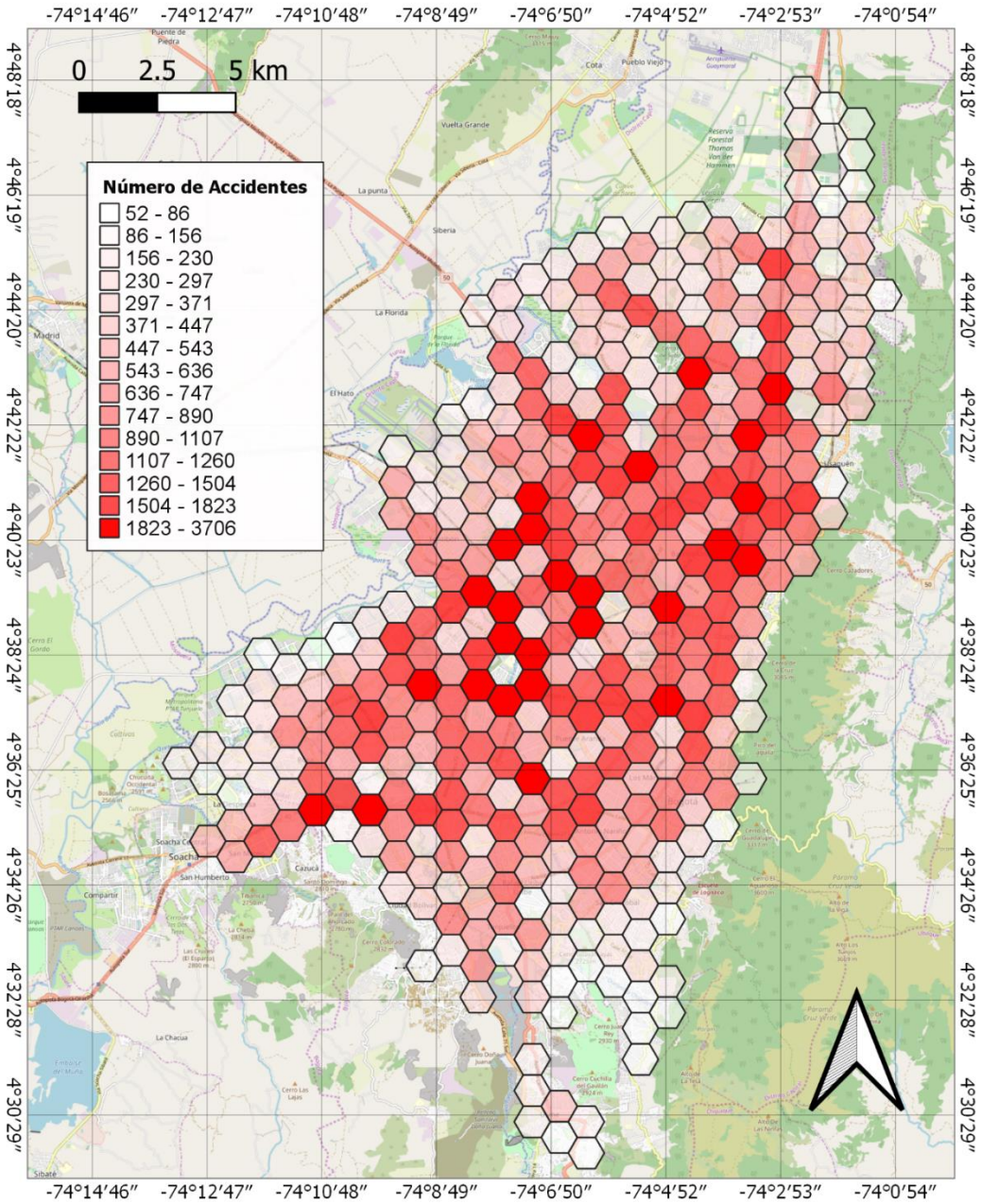
La creación del polígono cubre geográficamente la distribución de los eventos estudiados (Accidentes vehiculares y Eventos públicos), con un área de 855000 metros cuadrados, teniendo como referencia las zonas de influencias creadas en la sección 4.1.

5.3 Conteo de puntos en polígonos

Una forma de visualizar la concentración y distribución de los puntos se lleva a cabo por medio de conteo de puntos en polígonos. Como paso posterior a la creación de los polígonos hexagonales, se procede a identificar la cantidad de accidentes por área. Esta técnica que suele denominarse enlace espacial, el cual consiste contar el número de punto dentro de un área o polígono. Para este caso se crean polígonos de 570 metros de lado, donde se cubran los diferentes lugares o zonas de gran extensión y sus vías aledañas. Lo anterior, teniendo en cuenta el análisis realizado por zona de influencia a los lugares de presencia de eventos públicos.

Como resultado se obtiene una capa de polígonos, junto con sus atributos una nueva columna con el número de puntos que hay dentro de cada polígono de la capa resultante. Adicional del conteo se procede a categorizar en gama de colores la cantidad de puntos dentro de polígono, para la identificación visual de aquellas áreas en las que se tiene mayor concentración de accidentes. Los polígonos con mayor intensidad de color rojo son las áreas con un mayor número de accidentes. En la **Figura 5-5**, se muestra el conteo de accidentes dentro de polígonos hexagonales.

Figura 5-5. Conteo de accidentes vehiculares en polígonos hexagonales.

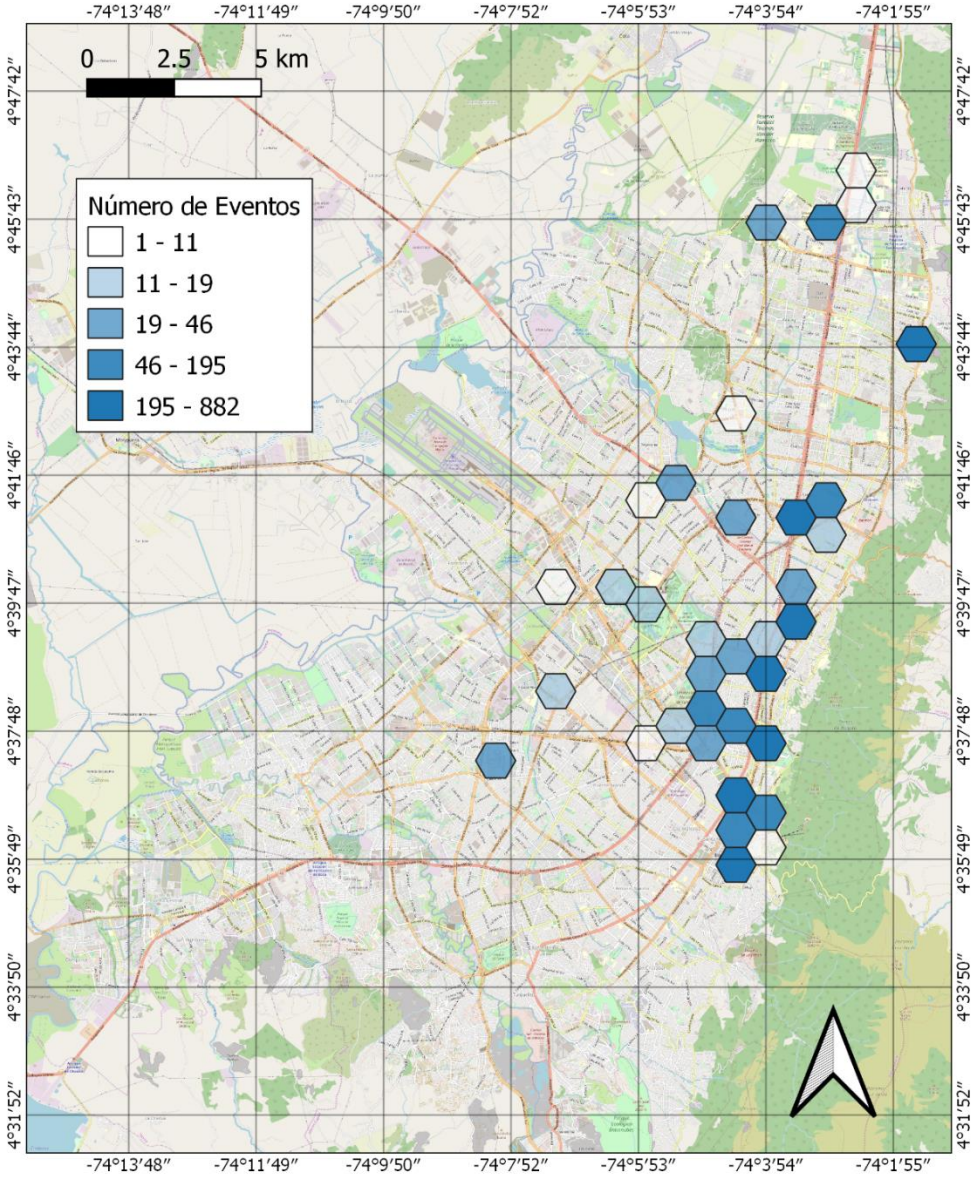


Fuente: elaboración propia.

Como resultado al análisis de densidad de kernel en polígonos hexagonales, se tiene una distribución espacial heterogénea de la accidentalidad. Observando que, al no tener una densidad continua de polígonos, se presenta concentraciones de forma dispersa. El conteo de puntos también es aplicado a los eventos públicos con respecto al lugar de ocurrencia.

Con el objetivo de identificar una asociación espacial en la ocurrencia dentro de estos dos eventos en las diferentes zonas de la ciudad. En la **Figura 5-6**, se visualiza el número de eventos que ocurrieron en los distintos lugares identificados en el capítulo anterior.

Figura 5-6. Conteo de eventos públicos en polígonos hexagonales.

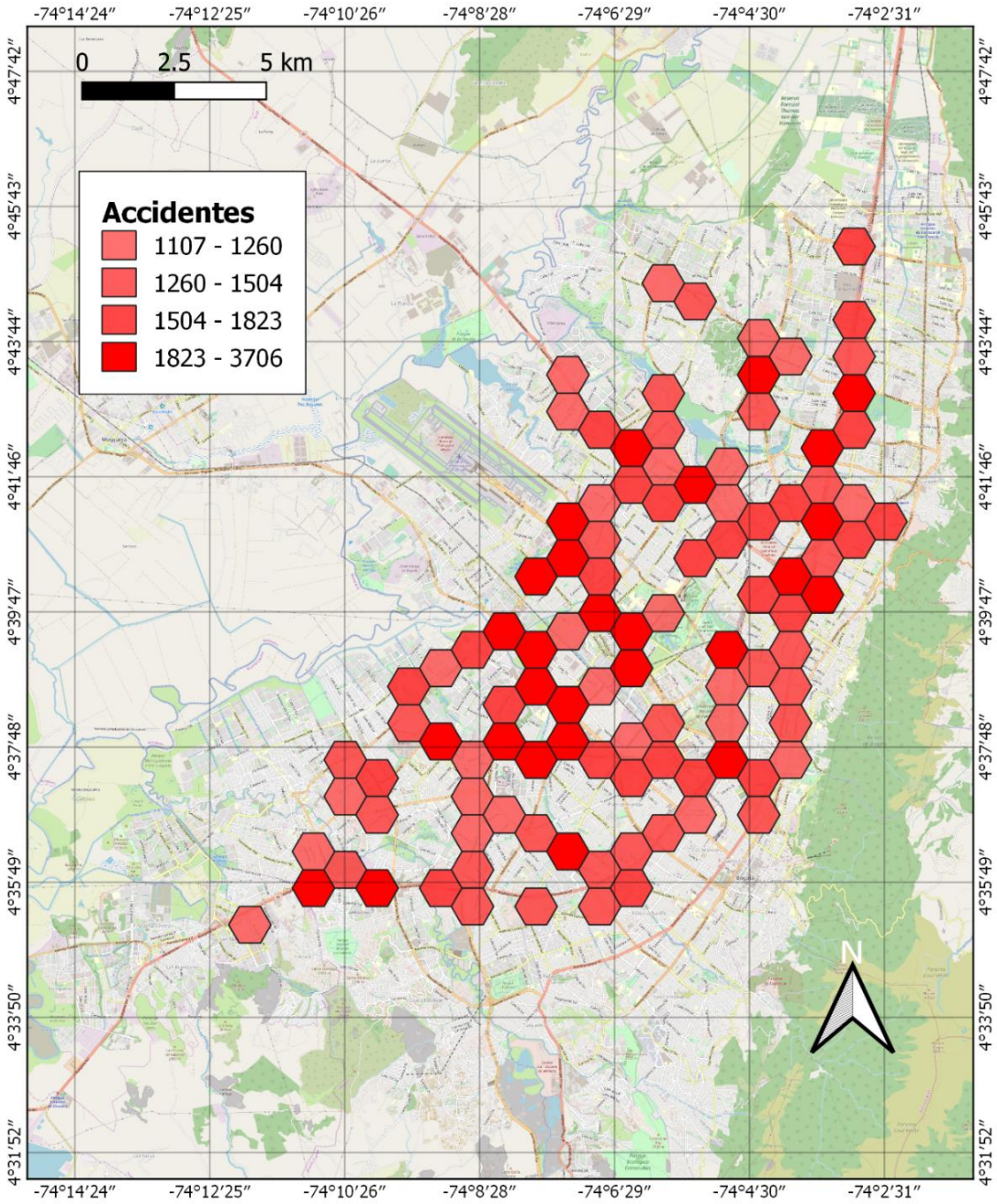


Fuente: elaboración propia.

Como punto importante se realiza la unión de las dos capas de densidad en polígonos, para el análisis conjunto de la distribución de los eventos públicos y los accidentes de tránsito. Aquellas unidades fisiográficas con resultados de menor concentración o

alrededor de cero se descartaron con la finalidad de tener un panorama de los polígonos de mayor accidentalidad con respecto a la ocurrencia de los eventos públicos (**Figura 5-7**).

Figura 5-7. Polígonos de eventos públicos y accidentes de tránsito con densidades mayores a 1100.



Fuente: elaboración propia.

De acuerdo con la dependencia espacial de las observaciones en términos de valores de agrupación **Figura 5-7**, las unidades con valores de 1107 a 3706 (tono rojo) indicaron clúster de puntos de calor considerablemente agrupados para el caso de los accidentes de tránsito, mientras que los eventos públicos se tiene datos desde 11 hasta 882 presentaron puntos de calor dispersos (tonos azules).

5.4 Índice de autocorrelación espacial global

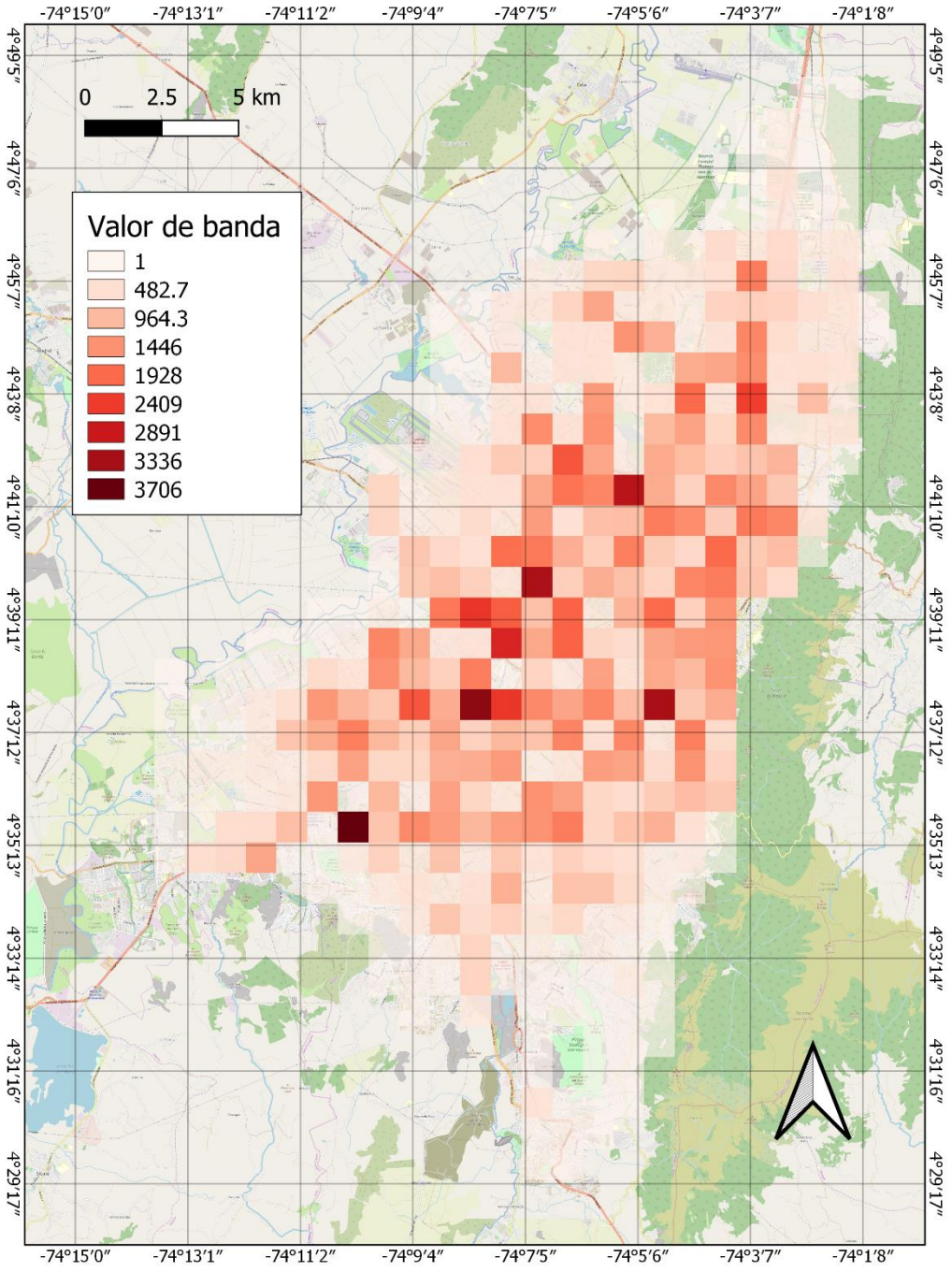
En el análisis de autocorrelación espacial se requiere alguna medida de contigüidad. La contigüidad tiene una definición bastante amplia según la pregunta de investigación, sin embargo, la mayoría de los análisis en autocorrelación espacial se adhieren a una definición común de las relaciones de vecindad. Para ello se aplica el Índice de Moran [55]. Este índice trata de contrastar la ausencia de autocorrelación espacial (aleatoriedad espacial) frente a la existencia de autocorrelación espacial (positiva o negativa). El índice global se limita a establecer la autocorrelación espacial en la totalidad del espacio geográfico en estudiado. El índice I de Moran viene dado por la siguiente expresión:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad para\ i \neq j \qquad (5\text{-}1)$$

donde $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} \ \ para\ i \neq j$, siendo n el tamaño muestral, esto es, el número total de unidades espaciales analizadas.

Para verificar la existencia o ausencia de autocorrelación espacial, se crea la capa tipo ráster con base a las densidades de los polígonos hexagonales (**Figura 5-5**) creados en la sección 5.3. En la **Figura 5-8** se visualiza en forma de ráster la matriz de celdas con valores que representan el grado de accidentalidad, para determinar la autocorrelación espacial de los accidentes de tránsito.

Figura 5-8. Mapa ráster de los accidentes de tránsito.



Fuente: elaboración propia.

En el estudio se hace uso de los casos de Rook's case, y Queen's case, para la validación de la existencia de aleatoriedad espacial de los accidentes de tránsito.

Tabla 5-1. Índice de autocorrelación global en accidentes de tránsito

Grilla	*Contigüidad*	*Índice Moran*	*Vecinos*	*Promedio celdas*
Rasterizado	Queen's case	0.38424	1376	628.561151
Rasterizado	Rook's case	0.37631	1376	628.561151

Fuente: elaboración propia.

El resultado de la aplicación de la autocorrelación espacial con el uso de los casos Queen´s y Rook´s se puede observar en la **Tabla 5-1**. Donde se presenta un índice de 0.384 y 0.376 respectivamente, el cual indica que existe una correlación positiva baja, dado que se acerca más a cero que al valor uno, llegando a tener una aleatoriedad espacial en la ocurrencia de accidentes de tránsito.

5.5 Nearest Neighbor Index

Por medio del método de vecino más cercano promedio se puede evaluar la distribución de puntos espaciales, validando el patrón que este presenta dispersión o agrupación. Para el análisis se calcula la distancia media observada, el índice de vecino más próximo, la puntuación z y la distancia media esperada. El índice de Vecino más próximo se expresa como la relación entre la distancia media observada y la distancia media esperada ver (ecuación 1) [56].

$$ANN = \frac{DO}{DE} \tag{5-2}$$

Donde DO es la distancia promedio observada entre cada punto y su vecino más cercano:

$$DO = \frac{\sum_{i=1}^{n} d_i}{n} \tag{5-3}$$

Y DE es la distancia promedio esperada para los puntos dado un patrón aleatorio.

$$DE = \frac{0.5}{\sqrt{n/a}} \tag{5-4}$$

En las anteriores ecuaciones d_i es igual a la distancia entre los puntos espaciales i y sus vecinos más cercanos. n Corresponde al total de puntos y a es el área que encierra todos los centroides. El z-score del promedio de vecino más cercano es calculado como:

$$z = \frac{DO - DE}{SE} \tag{5-5}$$

Donde:

$$SE = \frac{0.26136}{\sqrt{n^2/A}} \tag{5-6}$$

Como resultado se debe tener en cuenta que, si el índice es menor que 1, el patrón exhibe una agrupación de puntos espaciales; si el índice es mayor que 1, la tendencia es la dispersión.

Tabla 5-2. Índice de vecino más cercano en accidentes de tránsito y eventos públicos

Distribución	***Distancia media observada***	***Distancia media esperada***	***Índice vecino más cercano***	***Número de puntos***	***Z-Score***
Accidentes	5,0944	18,4031	0,2768	304030	0,0174
Eventos	981,7754	836.3883	1,1738	44	65.9099

Fuente: elaboración propia.

Comparando los resultados obtenidos de aplicar el índice de vecino más cercano en las distribuciones espaciales (Eventos públicos y Accidentes de tránsito) para verificar el patrón que estos forman en la ciudad de Bogotá ver (**Tabla 5-2**). Se encontró que la distribución de los accidentes con un valor de 0.2768 muestra una agrupación, teniendo como precedente que los accidentes se presentan mayormente en las principales avenidas e intersecciones de las mismas. A diferencia de los eventos públicos el cual presenta una

tendencia de dispersión con un valor 1.17. Dado que estos se encuentran en su gran mayoría en la zona central de la ciudad y los demás, pero con menor presencia, dispersos en las diferentes zonas de la ciudad de Bogotá. La dispersión que se presenta los eventos públicos es leve al tener un z-score de 65.9 el cual no es un valor relativamente grande. Para el caso de los accidentes de tránsito con un z-score de 0,017 al no ser negativo, la agrupación de los puntos no es significativa.

6. Conclusiones y recomendaciones

6.1 Conclusiones

En este trabajo se propuso el análisis de datos para determinar la relación entre los accidentes de tránsito y los eventos públicos a partir de la extracción de datos de páginas web y métodos geoestadísticos. Para ello, se partió de la realización de un estado del arte que permitiera establecer cada una de las etapas y métodos para la extracción y procesamiento de datos espaciales y los diferentes métodos y algoritmos utilizados en la actualidad para dicho análisis. Permitiendo determinar la relación que presenta los accidentes de tránsito con los eventos públicos en la ciudad de Bogotá.

A partir del método propuesto en la extracción de información de los eventos públicos, se logró la extracción de los datos necesarios para realizar el análisis espacio temporal con respecto a la ocurrencia de los accidentes de tránsito. Como primera limitación se obtuvo al no tener las direcciones de forma directa de los lugares de ocurrencia del evento extraído en las páginas web. Proceso posterior adicional que se realizó para obtener la dirección y la posición geográfica por medio de APIs de georreferenciación. Finalmente, como

resultado del proceso de extracción se crea una base de datos de los eventos públicos y los accidentes de tránsito con la fecha y ubicación de ocurrencia.

En el proceso de análisis espacio temporal se hace uso de sistemas de información geográficos, métodos de visualización y geoestadísticos. El cual se realiza un enfoque a nivel de zona de influencia y por días a diferentes franjas horarias. Inicialmente se obtuvo 46 lugares de ocurrencia de eventos, que para el caso de estudio se seleccionaron 5 lugares que contemplan diferentes características.
En términos generales respecto a los datos utilizados, hay un desequilibrio de información ya que la cantidad de reportes de momentos en los que no hay evento es considerablemente superior a la cantidad de reportes en momentos previos, esto tiene una justificación temática que en trabajos futuros debería revisarse para la generación de conjuntos de datos más pertinentes para responder a las preguntas de investigación.

En el caso de análisis espacio temporal para cada uno de los lugares se encontró:

- En el estadio el Campín se puede evidenciar una leve diferencia entre la accidentalidad que representa el conjunto de registros cuando no hay evento que la que representa el conjunto de registros que se tienen cuando si hay eventos. En los resultados desde un enfoque descriptivo, presentó mayor accidentalidad cuando no hay presencia de eventos públicos en las distintas zonas de influencia, sin embargo, existieron días donde hay mayor accidentalidad como los días martes en las zonas de menor distancias y los sábados a mayor distancia.

- En Movistar Arena, la cantidad de accidentes observados en las dos subpoblaciones interés oscila entre cero (0) y máximo dos (2) accidentes, lo que evidencia que el comportamiento de los accidentes no es diferente en los momentos cuando no se desarrollan eventos públicos que cuando sí se desarrollan. Aunque se tiene mayor accidentalidad cuando no hay eventos públicos en las diferentes zonas de influencia, para los días martes y viernes.

- Teatro nacional la castellana se tiene que en las observaciones realizadas con o sin presencia de eventos públicos la cantidad de accidentes ocurridos está entre

cero (0) y máximo de dos (2) accidentes viales aproximadamente, los resultados obtenidos de la prueba aplicados en el conjunto de datos agregados por día, concluyen que no hay significancia estadística para rechazar la diferencia del comportamiento de ocurrencia de accidentes para ambas muestras. El cual se observa que la ocurrencia de los accidentes es similar cuando hay y no hay eventos públicos, en los diferentes días y zonas de influencia.

- Para el parque Simón Bolívar dado los resultados estadísticos se tiene que no hay significancia estadística para afirmar que la accidentalidad vial tiene comportamiento diferencial en el conjunto de datos con desarrollo de eventos públicos y el que no contempla ningún evento de esta índole. Que para este caso solo se tiene eventos los días sábados y domingos.

- En el teatro Jorge Eliecer Gaitán no hay suficiencia estadística para soportar que hay una concordancia entre los valores obtenidos de estas dos variables de análisis. Lo implica que no hay diferencia entre el comportamiento de la cantidad de accidentes viales habiendo o no eventos públicos. A pesar que se presente mayor accidentalidad cuando no hay eventos públicos en las diferentes zonal de influencia y una aleatoriedad en los diferentes días de la semana.

Lo que significa que, para los diferentes casos de estudio si bien no se puede concluir que a medida que hay más eventos públicos más (o menos) accidentalidad vial se genera, dada la limitación de la diferencia significativa de datos cuando hay eventos frente a cuando no hay eventos y a la normalidad de los datos. Se puede dar a conocer que se presenta mayor accidentalidad cuando no hay presencia a cuando si los hay, relación que también se observa en el análisis realizado por las franjas horarias. Lo anterior tomando la variable de los eventos en un modo categórico.

En el análisis espacial planteado se encontró que no hay evidencia en una relación espacial en la ocurrencia de los eventos públicos con los accidentes de tránsito. Dada la dispersión de los accidentes en las diferentes zonas de la ciudad, teniendo en cuenta la ubicación de los puntos críticos de los mayores conglomerados de los accidentes y de los eventos públicos. En términos de la interdependencia espacial, Los valores observados en

los coeficientes de autocorrelación espacial (I de Moran) sugieren que la ocurrencia de los accidentes de tránsito es al azar, mostrando un fenómeno que espacialmente no son autocorrelacionados en la zona demarcada geográficamente de la ciudad. Sin embargo, los accidentes de tránsito son definidos por ciertos puntos críticos de las principales avenidas de la ciudad de Bogotá.

6.2 Discusión

Un factor común encontrado en el estudio, en los cincos lugares analizados, es que se encuentran en áreas de gran accidentalidad. Pero se podría hacer énfasis en algunas otras variables el cual pueden influir en el acontecimiento de un accidente de tránsito cuando hay o no eventos públicos.

Dado el gran aforo que presentan algunos sitios, como el Movistar Arena, el estadio el Campín y el parque Simón Bolívar. Estos presentan para algunos casos, controles de tránsito por parte de las autoridades durante la presencia de un evento, incluyendo cierres de vías y zonas de ingreso a parqueaderos. Otro valor a tener en cuenta son las zonas donde estos se encuentran y las vías de acceso.

El estadio el Campín como resultado del análisis de la ocurrencia de los accidentes, existieron días donde hay mayor accidentalidad en presencia de eventos como los días martes en las zonas de menor distancias y los sábados a mayor distancia. Teniendo como vía de acceso la carrera 30, el cual es avenida principal, se podría validar los controles de tránsito que se realizan entre semana y fines de semana. Validando que se hagan cierres de vías paralelas los días sábados, días de mayor ocurrencia de eventos deportivos, aumentando la accidentalidad en las vías aledañas. Siendo de igual forma para el Movistar Arena. Para el caso del parque Simón Bolívar, este presenta más de una vía de acceso, como la avenida 68, avenida calle 63 y la avenida carrera 60, avenidas de gran concurrencia. Por lo que se tiene congestión vehicular en las horas picos y en la presencia de los eventos, lo que generaría una baja diferencia de la ocurrencia de los accidentes de tránsito, cuando hay y no eventos públicos.

Por otra parte, el Teatro nacional la castellana y el teatro Jorge Eliecer Gaitán no están directamente sobre avenidas principales, pero si se encuentra en zonas comerciales. Estos sitios se podrían ver influenciados por la existencia constante de aglomeración de personas y transporte público, generando que la ocurrencia de los eventos es similar cuando hay y no hay eventos públicos, en los diferentes días y zonas de influencia.

6.3 Recomendaciones

Del trabajo realizado surgen las siguientes ideas y recomendaciones sobre las que se debe desarrollar en futuros trabajos:

1. Este método de extracción abre una oportunidad ya que, al aumentar la generación de grandes conjuntos de datos, el análisis desde un enfoque descriptivo, exploratorio en el que se involucren variables geo-temporales puede aportar información mucho más nutrida en una primera instancia.
2. Dada la técnica de extracción desarrollada en este trabajo de grado, para la estadística se generan grandes conjuntos de datos con mucho contenido para analizar ya que no hay restricciones desde el método para la obtención de múltiples atributos de una ubicación y un momento en el tiempo en particular.
3. Este trabajo abre la oportunidad para que se analice el comportamiento de variables de movilidad en contraposición de otras que se puedan extraer y que sean atravesadas por referenciación geo-temporal.
4. Si bien en los conjuntos de datos analizados predominaban los registros sin eventos públicos y eso pudo afectar los resultados estadísticos, la herramienta de extracción aportaría a lograr conjuntos más equilibrados y analizar comportamientos diferenciables sin que se vean afectados por cantidad de registros.

A. Anexo:

Anexo 1 Seudocódigo de *Web Scraping* de eventos públicos

```
browser.RealizarPeticion('https://vive.tuboleta.com')
codigoFuente ← browser.obtenerCodigoFuente()
CodigoHTML ← parsing(codigoFuente)

lista categoriaURL
Tabla DatosEventosPublicos(columnas=('Nombre', 'Categoria', 'Lugar', 'Fecha', 'Hora','Latitud','Longitud',
'direccion'))

para index en CodigoHTML BuscarURLCategoria('URLCategoria'){
   categoriaURL(index) ← index.seleccionar["URL"]
}
para index en categoriaURL {

   browser.RealizarPeticion(index.categoriaURL)
   codigoFuente ← browser.obtenerCodigoFuente ()
   CodigoHTMLCategoria ← parsing(codigoFuente)

   lista EventoPublicoURL
        para index en CodigoHTMLCategoria BuscarURLEventoPublico('URLEventoPublico'){
           EventoPublicoURL ← index.seleccionar["URL"])

           Para index en EventoPublicoURL{

              browser.RealizarPeticion(index.EventoPublicoURL)
              codigoFuente ← browser.obtenerCodigoFuente()
              CodigoHTMLEventoPublico ← parsing(codigoFuente)

              Para index en CodigoHTMLEventoPublico BuscarDatosPorIdClassHTML(class_="datos"){

                 DatosEventosPublicos(index).nombre ← index.seleccionar("nombre").obtener_texto())
                 DatosEventosPublicos(index).categoria ← index.seleccionar("categoria").obtener_texto())
                 DatosEventosPublicos(index).lugare ← index.seleccionar("lugare").obtener_texto())
                 DatosEventosPublicos(index).fecha ← index.seleccionar("fecha").obtener_texto())
                 DatosEventosPublicos.(index).hora ← index.seleccionar('hora').obtener_texto())
              }
           }
     }
para index en DatosEventosPublicos{
     resultado = gmaps.geocode(index.columnaLugar)

     DatosEventosPublicos(index).Latitud ← resultado.obtenerLatitud()
     DatosEventosPublicos(index).Longitud ← resultado.obtenerLongitud()
     DatosEventosPublicos(index).direccion ← resultado.obtenerDireccion()
}
DatosEventosPublicos ← DatosEventosPublicos.eliminarDuplicados por ['Nombre','Fecha']
DatosEventosPublicos.descargar.to_csv(Ruta_descarga)
```

Anexo 2. Script de estandarización de fecha

```
Update  eventos.eventos_civico set fecha = (RIGHT("fecha",4) || '-' ||
                    case when (SUBSTR("fecha",7,3)) = 'ene' then 01
                                when (SUBSTR("fecha",7,3)) = 'feb' then 02
                                when (SUBSTR("fecha",7,3)) = 'mar' then 03
                                when (SUBSTR("fecha",7,3)) = 'abr' then 04
                                when (SUBSTR("fecha",7,3)) = 'may' then 05
                                when (SUBSTR("fecha",7,3)) = 'jun' then 06
                                when (SUBSTR("fecha",7,3)) = 'jul' then 07
                                when (SUBSTR("fecha",7,3)) = 'ago' then 08
                                when (SUBSTR("fecha",7,3)) = 'sep' then 09
                                when (SUBSTR("fecha",7,3)) = 'oct' then 10
                                when (SUBSTR("fecha",7,3)) = 'nov' then 11
                                when (SUBSTR("fecha",7,3)) = 'dic' then 12
                    end
                    || '-' || left("fecha",2)
                    || ' ' || coalesce((nullif (SUBSTR("hora",1,7),null)),'00:00' ))::timestamp;
```

Anexo 3. Script de estandarización de hora.

```
Update  eventos.eventos_civico set hora = case when (left("hora",2)) = '1:'
          then regexp_replace(hora, '1:', '01:')
                    when (left("hora",2)) = '2:' then regexp_replace(hora, '2:', '02:')
                    when (left("hora",2)) = '3:' then regexp_replace(hora, '3:', '03:')
                    when (left("hora",2)) = '4:' then regexp_replace(hora, '4:', '04:')
                    when (left("hora",2)) = '5:' then regexp_replace(hora, '5:', '05:')
                    when (left("hora",2)) = '6:' then regexp_replace(hora, '6:', '06:')
                    when (left("hora",2)) = '7:' then regexp_replace(hora, '7:', '07:')
                    when (left("hora",2)) = '8:' then regexp_replace(hora, '8:', '08:')
                    when (left("hora",2)) = '9:' then regexp_replace(hora, '9:', '09:')
          end;
```

Bibliografía

[1] M. Novkovic, M. Arsenovic, S. Sladojevic, A. Anderla, and D. Stefanovic, “Data science applied to extract insights from data - weather data influence on traffic accidents,” *Infoteh-Jahorina*, vol. 16, no. March, pp. 387–392, 2017.

[2] Y. He, Z. Liu, X. Zhou, and B. Zhong, “Analysis of Urban Traffic Accidents Features and Correlation with Traffic Congestion in Large-Scale Construction District,” in *2017 International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 2017, pp. 641–644.

[3] L. Martín, L. Baena, L. Garach, G. López, and J. de Oña, “Using Data Mining Techniques to Road Safety Improvement in Spanish Roads,” *Procedia - Soc. Behav. Sci.*, vol. 160, no. Cit, pp. 607–614, 2014.

[4] D. T. Akomolafe and A. Olutayo, “Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways,” *Am. J. Database Theory Appl.*, vol. 1, no. 3, pp. 26–38, 2013.

[5] J. Xi, Z. Zhao, W. Li, and Q. Wang, “A Traffic Accident Causation Analysis Method Based on AHP-Apriori,” *Procedia Eng.*, vol. 137, pp. 680–687, 2016.

[6] C. Grimm, A. Fristo, M. Amin-Naseri, M. Hong, and A. Sharma, “Investigating the relationship between traffic incidents and public events: A case study,” *2017 Syst. Inf. Eng. Des. Symp. SIEDS 2017*, pp. 198–201, 2017.

[7] F. Á. Cerquera Escobar, “Modelo patrón de evaluación de la accidentalidad vial en áreas urbanas de Bogotá D.C. (Colombia),” *Carreteras*, vol. 4, no. 202, pp. 16–32, 2015.

[8] F. Á. Cerquera Escobar, “Análisis espacial de los accidentes de tráfico en

Bogotá D.C. Fundamentos de investigación," *Perspect. Geográfica*, vol. 18, no. 1, p. 9, 2013.

[9] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*. Longman Scientific & Technical, 1995.

[10] P. Legendre and M. J. Fortin, "Spatial pattern and ecological analysis," *Vegetatio*, vol. 80, no. 2, pp. 107–138, Jun. 1989.

[11] H. Shekhar, S. Setty, and U. Mudenagudi, "Vehicular traffic analysis from social media data," in *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, 2016, pp. 1628–1634.

[12] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. Part C Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.

[13] P. Tiwari, "Accident Analysis by using Data Mining Techniques," no. June, 2017.

[14] A. Serna, J. K. Gerrikagoitia, U. Bernabé, and T. Ruiz, "Sustainability analysis on Urban Mobility based on Social Media content," in *Transportation Research Procedia*, 2017, vol. 24, pp. 1–8.

[15] M. Sameen and B. Pradhan, "Severity Prediction of Traffic Accidents with Recurrent Neural Networks," *Appl. Sci.*, vol. 7, no. 6, p. 476, 2017.

[16] O. V.A and E. A.A, "Traffic Accident Analysis Using Decision Trees and Neural Networks," *Int. J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 22–28, 2014.

[17] R. S.Shanthi and Geetha Ramani, "Feature Relevance Analisys and Clasification of Road Traficc Accident Data through Data Mining Techniques," *World Congr. Eng. Comput. Sci.* , vol. 1, 2012.

[18] A. T. Kashani, A. Shariat-Mohaymany, and A. Ranjbari, "A Data Mining Approach To Identify Key Factors of Traffic Injury Severity," *Prelim. Commun. Saf. Secur. Traffic*, vol. 23, no. 1, pp. 11–17, 2011.

[19] W. Chaozhong, L. Hu, M. Ming, and Y. Xinping, “Severity analyses of single-vehicle crashes based on rough set theory,” *Proc. 2009 Int. Conf. Comput. Intell. Nat. Comput. CINC 2009*, no. 2, pp. 59–62, 2009.

[20] M. Chong, A. Abraham, and M. Paprzycki, “Traffic Accident Analysis Using Machine Learning Paradigms,” *Informatica*, vol. 29, pp. 89–98, 2005.

[21] C. M. Fuentes and V. Hernández, “La estructura espacial urbana y la incidencia de accidentes de tránsito en Tijuana, Baja California (2003-2004),” *Front. Norte*, vol. 21, no. 42, pp. 109–138, 2009.

[22] W.-S. Tseng, H. Nguyen, J. Liebowitz, and W. W. Agresti, “Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files,” *Ind. Manag. Data Syst.*, vol. 105, pp. 1188–1205, 2005.

[23] A. Jain, G. Ahuja, Anuranjana, and D. Mehrotra, “Data mining approach to analyse the road accidents in India,” *2016 5th Int. Conf. Reliab. Infocom Technol. Optim. ICRITO 2016 Trends Futur. Dir.*, pp. 175–179, 2016.

[24] S. Kumar and D. Toshniwal, “A data mining framework to analyze road accident data,” *J. Big Data*, vol. 2, no. 1, p. 26, 2015.

[25] S. Kumar and D. Toshniwal, “A novel framework to analyze road accident time series data,” *J. Big Data*, vol. 3, no. 1, p. 8, Dec. 2016.

[26] A. Gupta, A. Mohammad, A. Syed, and M. N. Halgamuge, “A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA,” *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 7, pp. 374–381, 2016.

[27] S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, “Citywide traffic congestion estimation with social media,” *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS ’15*, pp. 1–10, 2015.

[28] S. Krishnaveni and M. Hemalatha, “A Perspective Analysis of Traffic Accident using Data Mining Techniques,” *Int. J. Comput. Appl.*, vol. 23, no. 7, pp. 975–8887, 2011.

[29] T. Beshah and S. Hill, “Mining road traffic accident data to improve safety:

Role of road-related factors on accident severity in Ethiopia," *AAAI Spring Symp. - Tech. Rep.*, vol. SS-10-01, no. 1997, pp. 14–19, 2010.

[30] L. Guan, W. Liu, X. Yin, and L. Zhang, "Traffic Incident Duration Prediction Based on Artificial Neural Network," *2010 Int. Conf. Intell. Comput. Technol. Autom.*, no. 1997, pp. 1076–1079, 2010.

[31] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising," *Artif. Intell. Res.*, vol. 2, no. 1, pp. 44–54, 2012.

[32] N. R. Haddaway, "The use of web-scraping software in searching for grey literature," *Grey J.*, vol. 11, no. February, pp. 186–190, 2016.

[33] "Web Scraping with Python: Collecting More Data from the Modern Web - Ryan Mitchell - Google Libros." [Online]. Available: https://books.google.es/books?hl=es&lr=&id=TYtSDwAAQBAJ&oi=fnd&pg=PT8&dq=web+scraping+beautifulsoup+&ots=y0A_uKnjgn&sig=8zX3poVfULv539E_EAPUy1btDFI#v=onepage&q=web scraping beautifulsoup&f=false. [Accessed: 28-Oct-2019].

[34] "Fundamentos de las Infraestructuras de Datos Espaciales (IDE) - Miguel A. Bernabé-Poveda y Carlos M. López-Vázquez - Google Libros." [Online]. Available: https://books.google.com.co/books?id=lBxg1llU0S8C&pg=PT165&dq=base+de+datos+espaciales&hl=es&sa=X&ved=0ahUKEwjDu7rtmcDlAhXJqlkKHU_WAfcQ6AEIUTAG#v=onepage&q=base de datos espaciales&f=false. [Accessed: 28-Oct-2019].

[35] "Spatial Databases: With Application to GIS - Ph Rigaux, Michel Scholl, Agnès Voisard - Google Libros." [Online]. Available: https://books.google.com.co/books?id=o8LfhpFOnPwC&printsec=frontcover&dq=spatial+database&hl=es&sa=X&ved=0ahUKEwiZ5uDgncDlAhWls1kKHZ8lCe0Q6AEIMzAB#v=onepage&q=spatial database&f=false. [Accessed: 28-Oct-2019].

[36] L. Zhang and J. Yi, “Management methods of spatial data based on PostGIS,” *2010 2nd Pacific-Asia Conf. Circuits, Commun. Syst. PACCS 2010*, vol. 1, pp. 410–413, 2010.

[37] J. C. Martínez Llario, *PostGIS 2 : análisis espacial avanzado*. CreateSpace Independent, 2018.

[38] “PostGIS — Documentation.” [Online]. Available: http://postgis.net/documentation/. [Accessed: 28-Oct-2019].

[39] P. Compieta, S. Di Martino, M. Bertolotto, F. Ferrucci, and T. Kechadi, “Exploratory spatio-temporal data mining and visualization,” *J. Vis. Lang. Comput.*, vol. 18, no. 3, pp. 255–279, Jun. 2007.

[40] D. Guo and J. Mennis, “Spatial data mining and geographic knowledge discovery-An introduction.”

[41] “Spatial Database Systems: Design, Implementation and Project Management / Cheap-Library.com.” [Online]. Available: https://cheap-library.com/book/645412fb3af06f98e91ed080f2585b8f. [Accessed: 28-Oct-2019].

[42] C. F. Chen, C. Y. Chang, and J. Bin Chen, “Spatiial knowledge discovery using spatial data mining method,” in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2005, vol. 8, pp. 5602–5605.

[43] A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba, “Discovery of spatial association rules in geo-referenced census data: A relational mining approach,” IOS Press, 2003.

[44] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012.

[45] D. A. Keim, C. Panse, M. Sips, and S. C. North, “Pixel based visual data mining of geo-spatial data,” *Comput. Graph.*, vol. 28, no. 3, pp. 327–344, Jun. 2004.

[46] “A database clustering methodology and tool | Request PDF.” [Online]. Available: https://www.researchgate.net/publication/296860286_A_database_clusterin

g_methodology_and_tool. [Accessed: 28-Oct-2019].

[47] L. Wang, K. Xie, T. Chen, and X. Ma, “Efficient discovery of multilevel spatial association rules using partitions,” *Inf. Softw. Technol.*, vol. 47, no. 13, pp. 829–840, Oct. 2005.

[48] W. Sen Chen and Y. K. Du, “Using neural networks and data mining techniques for the financial distress prediction model,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 4075–4086, 2009.

[49] “QGIS: Becoming a GIS Power User - Anita Graser, Ben Mearns, Alex Mandel, Victor Olaya Ferrero, Alexander Bruy - Google Libros.” [Online]. Available: https://books.google.com.co/books?id=M1QoDwAAQBAJ&pg=PA545&dq=buffers+qgis&hl=es&sa=X&ved=0ahUKEwje_sfsnrnmAhVNwVkKHTFRB24Q6AEIPTAC#v=onepage&q=buffers qgis&f=false. [Accessed: 15-Dec-2019].

[50] N. Baghdadi, C. Mallet, and M. Zribi, *QGIS and applications in territorial planning*, vol. 3. 2018.

[51] “QGIS 2 Cookbook - Alex Mandel, Victor Olaya Ferrero, Anita Graser, Alexander Bruy - Google Libros.” [Online]. Available: https://books.google.com.co/books?id=h9vJDAAAQBAJ&pg=PA127&dq=buffers+qgis&hl=es&sa=X&ved=0ahUKEwje_sfsnrnmAhVNwVkKHTFRB24Q6AEIMTAB#v=onepage&q=buffers qgis&f=false. [Accessed: 15-Dec-2019].

[52] M. J. Marquez Dos Santos, *Estadistica Basica Un enfoque no parametrico*. 2000.

[53] D. Levin, Richard; Rubin, “ESTADISTICA PARA ADMINISTRACION Y ECONOMIA - Richard I. Levin, David S. Rubin - Google Libros,” 2003. [Online]. Available: https://books.google.com.co/books?id=uPhtNCqC4isC&pg=PA359&dq=prueba+de+hipotesis&hl=es&sa=X&ved=0ahUKEwib27yV0bPmAhWHjFkKHSSxCgYQ6AEIMzAB#v=onepage&q=prueba de hipotesis&f=false. [Accessed: 15-Dec-2019].

[54] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded*. Taylor & Francis, 2014.

[55] A. Mitchell, “Finding what’s nearby,” in *The ESRI guide to GIS analysis. Volume 1: Geographic patterns & relationships*, First., Redlands, California: ESRI Press, 1999, pp. 116–146.

[56] D. Ebdon, *Statistics in geography*. B. Blackwell, 1985.