Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería Mecánica Eléctrica

DISEÑO DE INVESTIGACIÓN PARA EL ANÁLISIS EXPLORATORIO *BIG DATA* APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO

Fernando Emmanuel Monzón Martinez
Asesorado por el MSc. Ing. Marco Alberto Villavicencio Sandoval

Guatemala, octubre de 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN PARA EL ANÁLISIS EXPLORATORIO *BIG DATA* APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

FERNANDO EMMANUEL MONZÓN MARTINEZ
ASESORADO POR EL MSC. ING. MARCO ALBERTO VILLAVICENCIO SANDOVAL

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO ELECTRÓNICO

GUATEMALA, OCTUBRE DE 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Pedro Antonio Aguilar Polanco
EXAMINADOR	Ing. Walter Giovanni Álvarez Marroquín
EXAMINADOR	Ing. Carlos Eduardo Guzmán Salazar
EXAMINADOR	Ing. Armando Alonso Rivera Carrillo
SECRETARIA	Inga. Lesbia Magalí Herrera López

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN PARA EL ANÁLISIS EXPLORATORIO *BIG DATA* APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 26 de abril de 2021.

Fernando Emmanuel Monzón Martinez

Escuela de Estudios de Postgrado, Edificio T-1
Teléfono: 2418-9142 / 24188000 ext. 1382
WhatsApp: 5746-9353
Email: informacion_eep@ingenieria.usac.edu.gt

https://postgrado.ingenieria.usac.edu.gt

Ref. EEPFI-0521-2021
Guatemala, 26 de abril de 2021

Director
Armando Alonso Rivera Carrillo
Escuela de Ingeniería Mecánica Eléctrica
Presente.

Estimado Ing. Rivera:

Reciba un cordial saludo de la Escuela de Estudios de Postgrado. El propósito de la presente es para informarle que se ha revisado y aprobado el DISEÑO DE INVESTIGACIÓN: **ANÁLISIS EXPLORATORIO BIG DATA APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO**, presentado por el estudiante **Fernando Emmanuel Monzón Martinez** carné número **201213383**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en Artes en Estadística Aplica.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Sin otro particular,

Atentamente.

"Id y Enseñad a Todos"

Mario Alberto Villavicencio Sandoval
Ingeniero en Ciencias y Sistemas
Colegiado No. 14,745

Mtro. Marco Alberto Villavicencio Sandoval
Asesor

"ID Y ENSEÑAD A TODOS" UNIVERSIDAD DE SAN CARLOS DE GUATEMALA ESCUELA DE POSTGRADO FACULTAD DE INGENIERIA

Mtro. Edwin Adalberto Bracamonte Orozco
Coordinador de Maestría
Estadística Aplicada

DIRECCIÓN

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingenieria

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

REF.EIME-012-2021

El Director de la Escuela de Ingeniería Mecánica Eléctrica de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **ANÁLISIS EXPLORATORIO BIG DATA APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO**, presentado por el estudiante universitario **Fernando Emmanuel Monzón Martínez**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS

Ing. Armando Alonso Rivera Carrillo
Director
Escuela de Ingeniería Mecánica Eléctrica

Guatemala, abril de 2021

Decanato
Facultad de Ingeniería
24189101 – 24189102
secretariadecanato@ingenieria.usac.edu.gt

DTG. 494-2021

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería Mecánica Eléctrica, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN PARA EL ANÁLISIS EXPLORATORIO BIG DATA APLICADO A LA MOVILIDAD EN EL MUNICIPIO DE ANTIGUA GUATEMALA, BASADO EN REGISTROS DE LLAMADAS, PARA ANALÍTICA PRESCRIPTIVA EN TURISMO Y COMERCIO,** presentado por el estudiante universitario: **Fernando Emmanuel Monzón Martinez,** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Anabela Cordova Estrada
Decana

Guatemala, octubre de 2021

AACE/asga

ACTO QUE DEDICO A:

Dios	Por haberme permitido realizar una más de mis metas.
Mis padres	Por haberme traído al mundo y guiado a través de él, mi eterno agradecimiento por su apoyo para hacer realidad este logro.
Mis hermanos	Luis, Sofía y Karen Monzón, por su apoyo y compañía durante mi vida.
Mis abuelos	Carlos Monzón (q. d. e. p.), Matilde Portillo, Antonio Martínez y Rosa Ruiz, por sus sabias enseñanzas y consejos durante toda mi vida.
Mi esposa	María de los Ángeles, por su apoyo y motivación para seguir hacia adelante.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser la *alma mater* que me brindó tanto en los últimos años.
Facultad de Ingeniería	Por proporcionarme los conocimientos y brindarme la oportunidad de conocer excelentes catedráticos y amigos.
Mis amigos	Por haberme acompañado durante la carrera.
Mi asesor	Msc. Ing. Marco Villavicencio, por haberme guiado durante el trabajo de graduación.

.

ÍNDICE GENERAL

.

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
D	Constante
s	Desviación estándar muestral
α	Error de estimación
Z_0	Estadístico de prueba
max	Función máximo valor
min	Función mínimo valor
H_A	Hipótesis alterna
H_0	Hipótesis nula
^	Indicador de estimador de un parámetro estadístico
B	Límite de error
$\bar{x}$	Media muestral
ID	Número único de identificación
θ	Parámetro estadístico
P	Proporción
IQR	Rango intercuartil
R	Rango muestral
n	Tamaño de la muestra
N	Tamaño de la población
Z_α	Valor crítico
s^2	Varianza muestral
σ^2	Varianza poblacional

GLOSARIO

Big Data	Término relacionado al crecimiento de datos generados por distintas tecnologías.
Business Analytics	Proceso por el cual las empresas utilizan técnicas estadísticas para analizar datos.
CDR	Registros de llamadas por sus siglas en inglés, *Call Detail Records*. Registro de datos producido por una central telefónica u otro equipo de telecomunicaciones que documenta los detalles de una llamada telefónica.
Dashboard	Es un tablero que presenta, de manera visual, métricas y gráficas.
Datawarehouse	Almacén de datos utilizado por empresas para guardar grandes cantidades de información.
INE	Instituto Nacional de Estadística, es un organismo descentralizado del estado, semiautónomo, cuyo principal fin es ejecutar la política estadística nacional
INGUAT	Instituto Guatemalteco de Turismo. Es la autoridad superior en materia de turismo en Guatemala, que rige y controla la promoción, fomento y desarrollo sostenible de la industria turística

R	Entorno y lenguaje de programación con enfoque al análisis estadístico.
Spark	*Framework* de computación en clúster para procesamientos *Big Data*.

RESUMEN

Actualmente para realizar analítica prescriptiva en el municipio de Antigua Guatemala existen fuentes de información que son recopiladas o generadas con métodos convencionales como encuestas y censos, sin embargo, estos tienen diversas limitaciones como largos tiempos de implementación, altos costos y distintas fuentes de error.

El presente diseño de investigación propone un análisis exploratorio de la movilidad urbana en el municipio de Antigua Guatemala, con base en CDR agregados, anonimizados y extrapolados de una empresa de telecomunicaciones.

La metodología propuesta busca disminuir limitaciones de los métodos convencionales brindando una mayor granularidad espacio temporal, proporcionando entre los resultados información agregada acerca del comportamiento de la gente como el uso que le dan al municipio, el lugar donde viven y trabajan, las regiones habitacionales, turísticas y comerciales.

Los resultados serán presentados de forma visual por medio de distintos gráficos como mapas de calor.

1. INTRODUCCIÓN

Este trabajo de investigación, enmarcado en la estadística de población, propone un análisis exploratorio de la movilidad urbana en el municipio de Antigua Guatemala, basándose en CDR agregados, anonimizados y extrapolados de una empresa de telecomunicaciones. Brindando una herramienta innovadora al sector turismo, comercio y gubernamental para aplicar analítica prescriptiva.

Actualmente existen fuentes de información basadas en censos y encuestas, las cuales por su metodología y naturaleza tienen limitaciones que este trabajo busca eliminar.

Esta metodología es innovadora en el sentido que se basa en observaciones, en vez de encuestas tradicionales, teniendo alcance a una mayor muestra utilizando herramientas *big data*. También brinda costos más accesibles debido a que los datos son generados por infraestructuras de telecomunicaciones ya existentes, sin necesidad de modificaciones a los teléfonos móviles y equipos de red.

Entre los resultados se obtendrá información agregada acerca del comportamiento de la gente como el uso que le dan al municipio, el lugar donde viven y trabajan, las regiones habitacionales, turísticas y comerciales, entre otros. Estos serán presentados de forma visual utilizando distintos gráficos. Brindando una mayor granularidad espacio temporal con respecto a metodologías ya existentes.

El informe final del estudio se describe el marco referencial, y se presentarán las herramientas de la estadística descriptiva que serán utilizadas para poder explorar y describir de forma óptima la información extraída de los datos.

También se expondrán los resultados de la investigación, donde gráficamente se presentará la respuesta a las preguntas planteadas acerca del comportamiento de la gente que se moviliza en el municipio de Antigua Guatemala. Se encontrará la discusión de los resultados obtenidos, donde se desarrollarán los argumentos para llegar a las conclusiones y recomendaciones de la investigación realizada.

2. ANTECEDENTES

El estudio de la movilidad urbana es un tema de bastante interés alrededor del mundo, debido al potencial que tiene en distintas áreas.

En Sarraute y Minnoni (2018) se listan y describen distintas aplicaciones de análisis de movilidad, basadas en datos de teléfonos móviles, para planeación urbana, predicción de uso de tráfico de datos, generación de mapas de riesgo epidemiológicos, entre otras. Estos referentes ayudarán a delimitar los objetivos del presente estudio y como guía para encontrar más referencias importantes en el campo.

En Behadili, Bertelle y George (2016) se exploraron las características de los patrones de movilidad humana utilizando CDR obtenidos de un festival en Francia. Los resultados muestran que el tiempo entre eventos tiene un patrón heterogéneo y que las trayectorias humanas siguen una distribución de ley de potencias. Este artículo provee evidencia de las distribuciones probabilísticas que se esperan encontrar en los CDR por utilizar.

En Alhasoun *et. al.* (2014) se presenta la herramienta *city browser* desarrollada para analizar la movilidad humana en la ciudad de Riyadh, Arabia Saudita. La metodología utilizada consiste en cuatro pasos: “Capturar zonas densas, analizar CDR para capturar las ubicaciones hogar/trabajo de los individuos, investigar la formación de comunidades dentro de las ciudades y estimar flujos de personas en una escala diaria” (p. 2). Esta arquitectura está compuesta por *data warehouse*, módulos y algoritmos, y una interfaz de

visualización. De aquí se podría utilizar la arquitectura y la metodología para capturar las ubicaciones hogar/trabajo.

En Furletti *et. al.* (2014) se presenta un caso de uso en la provincia de Pisa, Italia utilizando CDR, obteniendo resultados de población. Se presenta la idea de matriz origen destino, a nivel municipal, y se categoriza a los residentes, usuarios y conmutadores de la ciudad. De aquí se puede utilizar la metodología para categorizar a las personas de acuerdo con el uso que le den al municipio.

En Feriancic *et. al.* (2015) se presentan algunos casos de uso de la herramienta de *big data*, llamada *smart steps*, de telefónica en Inglaterra, España y Brasil. Se presentan los beneficios en estudios estratégicos, tácticos y operacionales. Utiliza la infraestructura de telefonía móvil y resalta la seguridad, ya que utiliza datos en masa, anónimos, agrupados y extrapolados, preservando la privacidad de los clientes y población en general. De aquí se piensa utilizar la metodología para preservar la seguridad y privacidad de los datos utilizados.

En Liu y Pöllmann (2020) se propone un modelo bayesiano para estimar la población dinámica utilizando datos de movilidad y datos estáticos de censos como conocimiento previo. Presenta una resolución espacial de 1 km y temporal de 1 hora, señalando que podría ser más fina. También presenta un cuadro comparativo de trabajo relacionado. De aquí se podrían rescatar ideas para validar los resultados obtenidos.

En Zhou *et. al.* (2018) se presentan y discuten diferentes tipos de datos para análisis de movilidad humana y métodos de análisis y preprocesamiento de datos. Se comparan las ventajas y desventajas de los tipos de datos y se recomienda en qué aplicaciones utilizarlos. De este artículo se resaltan las

desventajas y limitaciones que podrían presentar los CDR al realizar estudios de movilidad humana.

En Wang *et. al.* (2010) se propone un método para inferir el modo de transporte, con base en datos recolectados de CDR en la ciudad de Boston. Los resultados muestran, para un punto de origen y destino, el porcentaje de viajeros que caminan, usan transporte público y usan automóviles. De este artículo se podría utilizar la metodología para extraer los puntos de origen y destino en un viaje.

En Dong *et. al.* (2015) se introduce el concepto de evento inusual, a partir del aumento de movilidad por eventos deportivos, conciertos, marchas, protestas, entre otros. También se propone la metodología utilizando CDR como fuente de información para la detección de dichos eventos. De aquí se podría utilizar la detección de eventos inusuales, ya que podrían ser indicadores útiles para el propósito de este trabajo de investigación.

El turismo y comercio tienen una estrecha relación con indicadores socioeconómicos. En Pappalardo *et. al.* (2015) se estudia la relación que existe entre los patrones de movilidad humana y desarrollo socioeconómico. Se introduce una métrica de movilidad humana calculada a partir de la información extraída de CDR. Alguno de estos indicadores pudiera ser de utilidad en los resultados de esta investigación.

Se observa que, a pesar de no haber antecedentes de la solución propuesta en el país, hay bastante literatura y antecedentes en otras partes del mundo señalando la viabilidad y ventajas de la solución propuesta. Además, se encuentran distintas técnicas y metodologías para resolver problemas relacionados con la movilidad urbana.

3. PLANTEAMIENTO DEL PROBLEMA

Para Bentley (2017) asume que la analítica prescriptiva es la tercera fase del *business analytics*, que también incluye la descriptiva y predictiva. Esta se basa en la aplicación de ciencias matemáticas y computacionales, para sugerir opciones de decisión con el objetivo de aprovechar al máximo los resultados de la analítica descriptiva y predictiva.

Las empresas de telecomunicaciones utilizan CDR, “los cuales consisten en información acerca de la actividad de usuarios en redes celulares. Usualmente son utilizados para el propósito de facturación, mantenimiento de infraestructura, y manejo de recursos por proveedores de servicio” (Saleh y Bahrak, 2019, p. 3).

Según el INE (2019) el municipio de Antigua Guatemala es la cabecera del departamento de Sacatepéquez, el cual se constituye en uno de los lugares turísticos más populares en Guatemala y de acuerdo con el último censo realizado en el 2018 “el municipio de Antigua Guatemala, departamento de Sacatepéquez cuenta con una población de 46 054 personas” (p. 99).

En INGUAT (2020) menciona que “el flujo de llegadas de visitantes residentes a Sacatepéquez durante el año 2018 es 4 766 251 visitantes siendo Semana Santa la festividad con mayor afluencia” (pp. 22-24). Determinando que la mayoría del turismo interno proviene del departamento de Guatemala y el país con mayor presencia es Estados Unidos. Estos datos son importantes y son los que actualmente tienen a su alcance el sector turismo, laboral y gubernamental para aplicar analítica prescriptiva.

Actualmente existen distintas fuentes de información que el municipio de Antigua Guatemala puede utilizar para realizar analítica prescriptiva. Estas se generan o recopilan utilizando métodos convencionales como encuestas y censos, sin embargo, tienen limitaciones.

Según Sarraute y Minnoni (2018) “estas metodologías consisten principalmente en realizar encuestas de movilidad en casas, diarios de viaje o encuestas de transporte público y privado. Estos estudios son metodológicamente complejos, requieren largos tiempos de implementación y cantidades significativas de recursos económicos” (p. 1).

De acuerdo con Scheaffer *et. al.* (2012) las encuestas pueden tener distintas fuentes de error, inducidas por el encuestado, entrevistador o por el mismo diseñador de los instrumentos de medición. El encuestado puede inducir errores a propósito por desconfianza al encuestador, mala memoria para recordar hechos pasados o mala intención al dar información falsa. Los instrumentos pueden influenciar al encuestado con el orden en que se presentan las preguntas, las palabras utilizadas y con las opciones de respuestas en preguntas de opción múltiple.

Para Sarraute y Minnoni (2018) los métodos tradicionales tienen muestras pequeñas, en el orden de decenas de miles o incluso menos en ciudades más pequeñas, para mantener costos razonables. Además, estos estudios se realizan máximo una vez por década, en el mejor de los casos, en muchos países.

Debido a la importancia que tiene esta información para la analítica prescriptiva se requiere una solución, basada en observaciones, que permita analizar el comportamiento de la movilidad urbana en la Antigua Guatemala con suficiente granularidad espacio temporal. Pudiendo distinguir a las personas que

habitan, laboran y visitan el municipio de Antigua Guatemala; encontrar sus lugares de origen y destino; distinguir las zonas domiciliarias, laborales y turísticas; y cualquier otro patrón de utilidad.

No se conoce cuál es el comportamiento de las personas, respecto a la forma de movilizarse por el municipio y las áreas o destinos más frecuentados. Se desconoce la distribución espacial de los visitantes, trabajadores y habitantes.

Esto lleva a plantear la pregunta principal de este estudio: ¿Cuál es el comportamiento de la movilidad de la gente que se mueve dentro del municipio de Antigua Guatemala?

Para responder a esta interrogante se deberán contestar las siguientes preguntas auxiliares:

- ¿Cuál es el comportamiento de las ubicaciones que visitan las personas que viven en Antigua Guatemala?

- ¿En qué sectores hay mayor frecuencia de actividades laborales en Antigua Guatemala y de qué puntos geográficos provienen los colectivos de trabajadores?

- ¿Qué sectores tienen la mayor actividad turística y de dónde provienen los visitantes?

- ¿Cómo es la distribución espacial de los sectores turístico, laboral y domiciliar en la Antigua Guatemala?

- ¿Cuál es el grado de validez de los resultados para la estimación de la población en un municipio de Guatemala?

4. JUSTIFICACIÓN

La investigación propuesta se justifica en el campo de la estadística de población haciendo un análisis exploratorio de la movilidad urbana en el municipio de Antigua Guatemala.

Se espera que la metodología a partir de CDR tenga varias ventajas como: El aumento del tamaño de muestra, en comparación con encuestas tradicionales, aumentando la confianza de la información; resultados basados en observación del comportamiento realizado, en vez del declarado; reducción de tiempo de ejecución de toma de datos, pudiendo aumentar la frecuencia de los reportes y la automatización de estos y mayor precisión espacio temporal en la extrapolación.

Así mismo, se espera que los sectores turismo, comercio y gubernamental se vean beneficiados teniendo información más completa, confiable y actualizada para la toma de decisiones.

5. OBJETIVOS

5.1. General

Desarrollar un análisis exploratorio del comportamiento de la movilidad urbana en el municipio de Antigua Guatemala, aplicando técnicas descriptivas en registros de llamadas agrupadas y anonimizadas de una empresa de telecomunicaciones, para mejorar la analítica prescriptiva en turismo y comercio.

5.2. Específicos

- Segmentar las ubicaciones, que visita la gente que habita en Antigua Guatemala, en mapas geográficos utilizando técnicas descriptivas en los registros de llamadas para comprender su movilidad.

- Identificar y tabular las ubicaciones de trabajo en la Antigua Guatemala y los municipios de origen del sector laboral, utilizando medidas de tendencia central y dispersión en registros de llamadas para entender su movilidad del municipio de Antigua Guatemala.

- Establecer las zonas turísticas y el origen de los visitantes, nacionales y extranjeros, utilizando las ubicaciones más frecuentes para identificar su lugar de origen utilizando técnicas descriptivas por medio de un mapa geográfico.

- Distinguir las regiones residenciales, comerciales y turísticas dentro de la ciudad con la movilidad de la gente para comprender su distribución utilizando técnicas descriptivas por medio de mapas geográficos.

- Comprobar la validez del estudio con la cantidad de usuarios extrapolados en los registros de llamadas, comparando los resultados con las fuentes oficiales del INE y del INGUAT por medio de inferencia estadística.

6. NECESIDADES POR CUBRIR Y ESQUEMA DE LA SOLUCIÓN

Con base en los CDR de una empresa de telecomunicaciones y un catálogo de celdas, con sus coordenadas geográficas, se construirá una base de datos con los registros de llamadas de personas que se hayan movilizado en Antigua Guatemala durante las fechas delimitadas.

Tomando en cuenta los horarios en que la gente se conecta a la red, se obtendrán las celdas hogar/trabajo. Después se podrá clasificar a los usuarios de la red con respecto al uso que le dan al municipio. Con base en las frecuencias de registros y horarios se clasificaron áreas del municipio en zonas habitacionales, comerciales y turísticas, procediendo a visualizarlas.

La validez del estudio se hará comparando con las fuentes de datos ya existentes, tomando en cuenta sus limitaciones.

7. MARCO TEÓRICO

Para Ross (2004) se presentan varias definiciones de la estadística, entre las cuales se resalta que “es una herramienta matemática útil para analizar datos observacionales para aprender y estudiar algo acerca de la naturaleza de poblaciones” (p. 1). Además, Singpurwalla (2013) agrega que generalmente incluye recolectar, clasificar, resumir, organizar, analizar e interpretar datos.

7.1. Muestreo

Según Ross (2004) el interés de la estadística es estudiar información de una colección de elementos, es decir, de una población “cuando el tamaño de la población es demasiado grande basta con tomar un subgrupo de elementos, conocido como muestra” (p. 22). Para que la muestra sea útil es necesario que esta sea representativa de la población.

Para Scheaffer *et. al.* (2012) definen algunos términos técnicos que son importantes entender antes de diseñar algún muestreo:

- Elemento: objeto sobre el cual se realiza una medición.
- Población: colección de objetos sobre la que se desea realizar alguna inferencia.
- Unidad de muestreo: las unidades de muestreo son colecciones, sin traslape, de elementos que cubren la población completa.
- Marco: es una lista de unidades de muestreo.
- Muestra: es una colección de unidades de muestreo tomadas de uno o varios marcos.

7.1.1. Error de estimación

Para Scheaffer *et. al.* (2012) "si θ es el parámetro de interés y es un estimador de $\hat{\theta}$, deberíamos especificar un límite en nuestro error de estimación" (p. 10). En el caso de un parámetro θ y un estimador $\hat{\theta}$, se definen el error y su límite como:

$$Error\ de\ estimación = \left|\theta - \hat{\theta}\right| < B \tag{1}$$

También se define la probabilidad de que el error de estimación sea menor al límite B como:

$$P\left[\left|\theta - \hat{\theta}\right| < B\right] = 1 - \alpha \tag{2}$$

7.1.2. Muestreo aleatorio simple

Existen diferentes tipos de muestreos, el más sencillo y fácil de implementar es el muestreo simple aleatorio. Scheaffer *et. al.* (2012) lo definen como: "si una muestra de tamaño n es seleccionada de una población de tamaño N, tal que todas las posibles muestras de tamaño n tenga la misma probabilidad, el procedimiento de muestreo es llamado muestreo aleatorio simple" (p. 76).

7.1.2.1. Tamaño de la muestra

Para el tamaño de la muestra para estimar algún parámetro cumpliendo con el límite de error B, lo presentan como:

$$n = \frac{N\ \sigma^2}{(N-1)\ D + \ \sigma^2} \tag{3}$$

"Donde σ^2 es la varianza poblacional del parámetro, N el tamaño de la población y $D = B^2/4$, para estimar la media o proporción y $D = \frac{B^2}{4\,N^2}$ para estimar el total de la población" (Scheaffer *et. al.,* 2012, p. 90).

7.2. Estadística descriptiva

La estadística descriptiva según Singpurwalla (2013) "tiene como objetivo describir un conjunto de datos, para esto utiliza métodos numéricos y gráficos para encontrar patrones, resumir la información y presentarla de forma visual" (p. 9).

7.2.1. Medidas de tendencia central

Una forma de resumir grandes conjuntos de datos es utilizando medidas de tendencia central, entre ellas, las más utilizadas son la media aritmética, la mediana, y la moda.

7.2.1.1. Media muestral

Es un estadístico que se calcula a partir de la media aritmética de un conjunto de valores de una variable aleatoria, suponiendo que un conjunto de datos tiene n valores numéricos $x_1, x_2, \dots, x_n$. La media de la muestra, $\bar{x}$, está definida por:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad (4)$$

7.2.1.2. Mediana muestral

Se debe partir de n datos ordenados “si n es impar la mediana muestral es el valor en la posición $\frac{n+1}{2}$; si n es par, es el promedio de los valores en las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$” (Ross, 2004, p. 20). La mediana muestral está definida por:

$$mediana = \begin{Bmatrix} x_{\left\{\frac{n+1}{2}\right\}}, & si\ n\ es\ impar \\ \frac{x_{\left\{\frac{n}{2}\right\}} + x_{\left\{\frac{n}{2}+1\right\}}}{2}, & si\ n\ es\ par \end{Bmatrix} \quad (5)$$

7.2.1.3. Moda muestral

La moda es el valor que en una muestra para Ross (2004) “ocurre con la mayor frecuencia. Si no es solo uno, entonces todos los valores con la mayor frecuencia son llamados valores modales” (p. 22). En un grupo de datos puede haber dos modas y se conoce como bimodal, se llama amodal cuando en un conglomerado no se repiten los valores.

Para Rosenkrantz (2009) el uso de la moda es más razonable con variables categóricas. La mediana es más robusta ante los valores atípicos y la media es más apropiada cuando la distribución es simétrica.

7.2.2. Medidas de dispersión

Otras medidas de utilidad para resumir son las de dispersión, las cuales son indicadores de la variabilidad y la dispersión de los datos con respecto a la media. Las más utilizadas son:

7.2.2.1. Varianza muestral

Mide la variación de los datos con respecto a la media. Está definida como:

$$s^2 = \sum_{i=1}^{n} \frac{\left(x_i - \underline{x}\right)^2}{(n-1)} \tag{6}$$

7.2.2.2. Desviación estándar muestral

Es la raíz cuadrada de la varianza y tiene la ventaja que está en las mismas unidades que los datos. Está definida por:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}} \tag{7}$$

7.2.2.3. Percentiles muestrales

Para Ross (2004):

El percentil 100p es el valor que es mayor o igual que el 100p % de los datos, y es menor o igual que el 100 (1-p) % de los datos. Si dos valores cumplen esta condición, entonces el percentil es el promedio aritmético de estos dos valores.

El percentil 50 es la mediana y junto con el 25 y 75 también se conocen como cuartiles. El percentil 25 es el primer cuartil, la mediana el segundo cuartil y el percentil 75 el tercer cuartil. (p. 25)

7.2.2.4. Rango muestral e IQR

Según Rosenkrantz (2009):

El rango de muestra es la longitud de el intervalo más pequeño que contiene todos los valores observados; el rango intercuartílico es la longitud del intervalo que contiene la mitad central de los datos el cual está definido por:

$$R = max(x_i) - min\,(x_i) \tag{8}$$

Y el rango intercuartílico se define por,

$$IQR = Q_3 - Q_1 \tag{9}$$

Teniendo en cuenta que se considera un dato como atípico cuando este se encuentra fuera del intervalo:

$$(Q_1 - 1.5 * IQR,\ Q_3 + 1.5 * IQR) \tag{10}$$

> El intervalo [Q1, Q3] se denomina rango medio del 50 %. (pp. 29–30)

7.2.3. Tablas de frecuencia

Para Singpurwalla (2013) “las tablas de frecuencias proporcionan recuentos y porcentajes de datos cualitativos por clase. Una clase es una de las categorías en las que se pueden clasificar los datos cualitativos” (p.18).

La frecuencia de clase es el número de observaciones a un grupo de datos y la frecuencia relativa de clase es la frecuencia de clase dividida por el número total de observaciones en el conjunto de datos de la tabla.

Tabla I. **Ejemplo de tabla de frecuencia**

Salario ($k)	Frecuencia	Frecuencia relativa
47	4	0.095
48	1	0.024
49	3	0.071
50	5	0.119
51	8	0.190
52	10	0.238
53	0	0.000
54	5	0.119
56	2	0.048
57	3	0.071
60	1	0.024
Total	**42**	**1**

Fuente: Ross (2004) *Introduction to Probability and Statistics for Engineers and Scientists*.

En la tabla I, se presenta un ejemplo con datos de salarios anuales, en miles de dólares, de ingenieros eléctricos recién graduados en Estados Unidos.

7.2.4. Gráficos

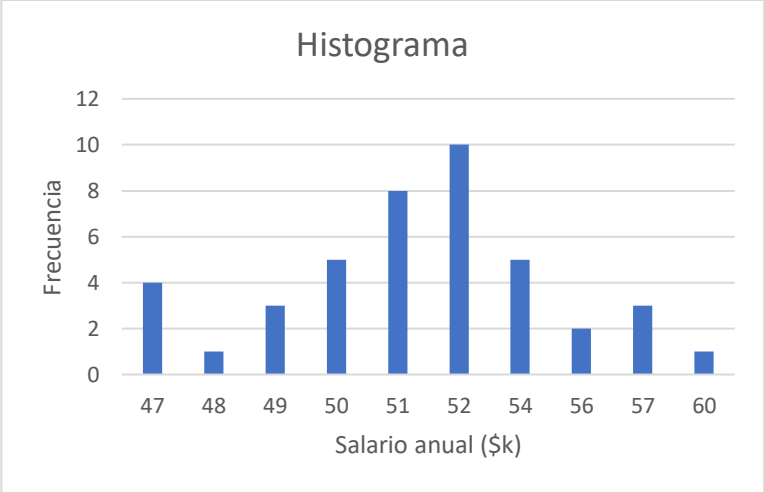
Para representar los datos de forma visual se utilizan gráficos. Existen distintos tipos y su elección dependerá de los atributos que el autor desee presentar y resaltar.

7.2.4.1. Histogramas

Para Ross (2004) un histograma “es un diagrama de gráfico de barras de datos de clase, con las barras colocadas adyacentes entre sí” (p. 16). El histograma se utiliza para representar datos en tablas de frecuencia el cual consiste en barras verticales para cada clase, con el eje vertical que representa la frecuencia o frecuencia relativa. En la figura 1, se presentan los datos de la tabla I.

Según Cohen y Cohen (2008) con los histogramas se busca observar la distribución de los datos.

Figura 1. **Ejemplo de histograma**



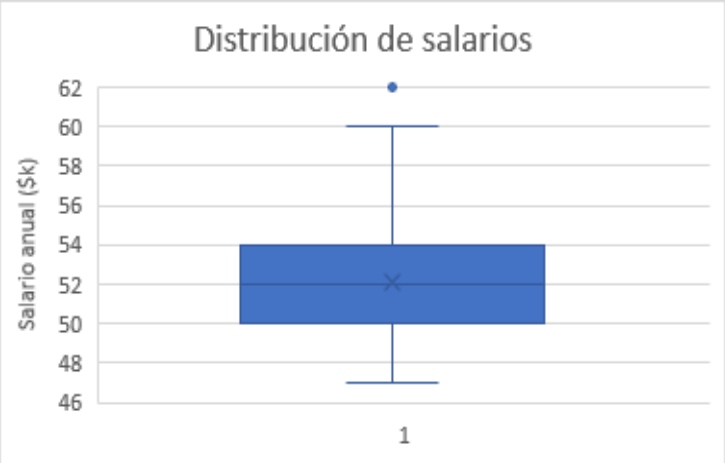
Fuente: elaboración propia.

7.2.4.2. Diagramas de cajas

Según Rosenkrantz (2009) “el diagrama de caja es una representación gráfica de la mediana, los cuartiles rangos intercuartiles y datos atípicos de un conjunto de datos” (p. 30).

> Un diagrama de caja se usa a menudo para trazar algunas de las estadísticas resumidas de un conjunto de datos. Un segmento de línea recta que se extiende desde el valor de datos más pequeño al más grande se dibuja en un eje horizontal impuesto en la línea es una "caja", que comienza en el primero y continúa hasta el tercer cuartil, con el valor del segundo cuartil indicado por una línea vertical. (Ross, 2004, p. 27)

Figura 2. **Ejemplo de diagrama de cajas**

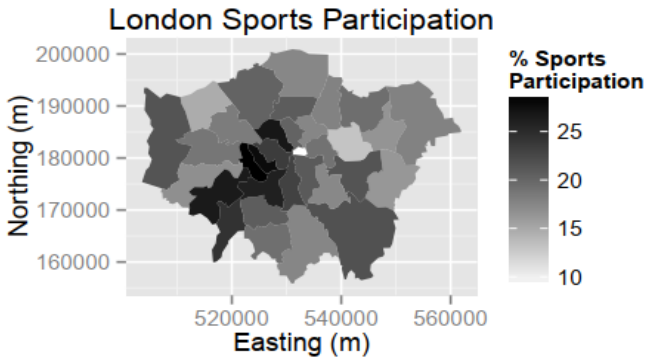


Fuente: elaboración propia.

7.2.4.3. Mapas de calor

Para Lovelace y Cheshire (2014) un mapa de color “es un gráfico que muestra las escalas de varios valores de profundidad de color en un mapa geográfico” (p. 18). En la figura 3, se observa un mapa de Londres donde se puede identificar las regiones con mayor participación deportiva.

Figura 3. **Ejemplo de mapa de calor**



Fuente: Lovelace y Cheshire (2014). *Introduction to visualising spatial data in R*.

7.2.5. Inferencia estadística

Según Gutiérrez y Vara (2008) “el objetivo de la inferencia estadística es hacer afirmaciones válidas acerca de la población o proceso con base en la información contenida en una muestra. La inferencia estadística por lo general se divide en estimación y prueba de hipótesis” (p. 20)

7.2.5.1. Prueba de hipótesis

Para Gutiérrez y Vara (2008) la prueba de hipótesis es una herramienta estadística utilizada para poner a prueba supuestos o creencias a priori, esto con base en datos.

Una hipótesis estadística para Gutiérrez y Vara (2008) es una "afirmación sobre los valores de los parámetros de una población o proceso, que es susceptible de probarse a partir de la información contenida en una muestra representativa que es obtenida de la población" (p. 30).

En una prueba de hipótesis se debe plantear dos hipótesis opuestas, la hipótesis nula (H_0) y la hipótesis alternativa (H_A). La hipótesis nula es la creencia a priori. Para comprobarla se supone que es verdadera y con base en un estadístico de prueba se rechaza, aceptando la hipótesis alternativa o no se rechaza.

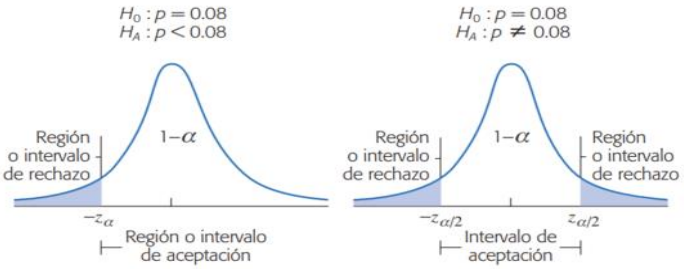
- Estadístico de prueba

Para Gutiérrez y Vara (2008) un estadístico de prueba es un "número calculado a partir de los datos y de H_0, cuya magnitud permite discernir si se rechaza o no la hipótesis nula" (p. 61).

- Criterio de rechazo

Para Gutiérrez y Vara (2008) "el valor del estadístico de prueba debería caer dentro del rango de valores más probables de su distribución asociada, el cual se conoce como región de aceptación" (p. 31). Por otro lado, si el estadístico de prueba cae fuera, es decir en la región de rechazo, se considera como evidencia de que la hipótesis nula es falsa, rechazándola y aceptando la alternativa.

Figura 4. **Hipótesis unilateral y bilateral**



Fuente: Gutiérrez y Vara (2008). *Análisis y diseño de experimentos*.

7.2.5.2. Diferencia de proporciones

Para Gutiérrez y Vara (2008) "una situación de frecuente interés es investigar la igualdad de las proporciones de dos poblaciones o tratamientos" (p. 44). Para poder verificar la igualdad se requiere probar la siguiente hipótesis:

$$H_0: p_1 = P_2$$
$$H_A: p_1 \neq p_2$$

Donde p_1 y p_2 son las proporciones de una población, de acuerdo con el estadístico de prueba, bajo el supuesto de distribución binomial está dado por:

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (11)$$

Donde $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$. Se rechaza H_0 si $|z_0| > z_{\alpha/2}$.

7.2.6. Analítica prescriptiva

Para Bentley (2017) “la analítica es la comprensión y comunicación de patrones importantes de datos. La analítica se aplica en las empresas para mejorar su rendimiento” (p. 24). Esta se basa en la aplicación de ciencias matemáticas y computacionales para extraer valor de los datos.

El *business analytics* se divide en tres partes: analítica descriptiva, la cual observa datos históricos para comprender el pasado; analítica predictiva, la cual predice que es lo más probable que ocurra en el futuro; y finalmente la analítica prescriptiva, la cual anticipa qué va a pasar, cuándo y por qué, además sugiere opciones de decisión para maximizar el aprovechamiento de las ventajas y disminución de riesgos.

7.3. Conceptos básicos de telecomunicaciones

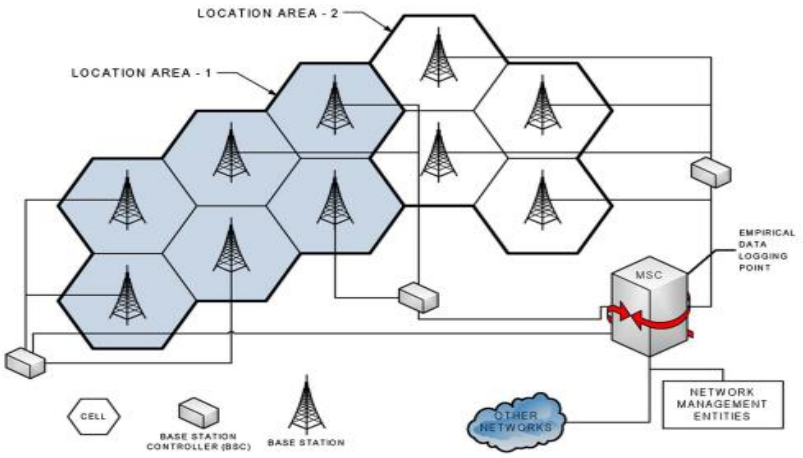
Las comunicaciones celulares tienen un papel importante en nuestras vidas conectando personas a través de llamadas y transmisiones de datos, contribuyendo al desarrollo económico, social y mejora de la calidad de vida de la población en todo el mundo las telecomunicaciones se componen de diferentes infraestructuras las cuales son vitales para su funcionamiento.

Las telecomunicaciones son fundamentales para las personas y las empresas, y su importancia puede duplicarse con el número de usuarios existentes. Cuantos más usuarios se conecten al sistema de comunicación, mayores serán las capacidades y necesidades de comunicación.

7.3.1. Antenas de radiofrecuencia y celdas

Las comunicaciones celulares están distribuidas con radio-bases con posiciones fijas sobre regiones de tierra llamadas celdas, y cada una de estas está cubierta por al menos una antena transmisora de microondas. “La distancia cubierta por las antenas debe ser considerada dependiendo de las características de la antena, banda de frecuencia y nivel de potencia, desde algunos kilómetros hasta decenas de kilómetros” (Penttienen, 2015, p. 64).

Figura 5. **Modelo simplificado de una red GSM**



Fuente: Büyükçorak, Kurt, y Cengaver (2014). *A Probabilistic Framework for Estimating Call Holding Time Distributions*.

7.3.2. Registros de llamadas (CDR)

Para Junbo, Wu, Hsu, y Cheng (2017) los sistemas de las comunicaciones celulares generan dos tipos de datos, los orientados a usuarios y orientados a sistemas. Los registros de llamadas (CDR por sus siglas en inglés, *call detail*

records) son la principal fuente de datos orientados a usuarios en una empresa de telecomunicaciones.

Además, menciona que cada actividad realizada por los usuarios, como mensajes de texto, llamadas y navegación por internet, es procesada por los equipos en las radio-bases escribiendo registros que incluyen datos como números de teléfono, parámetros de los aparatos, fecha y hora, duración de la llamada y la celda que dio el servicio, entre otros.

7.3.3. Big data

El *big data* es un término que ha surgido debido al crecimiento de datos generados por personas y distintas tecnologías como 5G, internet de las cosas, entre otras.

De acuerdo con Prajapati (2013):

El *big data* normalmente menciona el modelo 3V, que son velocidad, volumen y variedad.

- La velocidad se refiere a la baja latencia y la velocidad en tiempo real a la que se deben aplicar los análisis.
- El volumen se refiere al tamaño del conjunto de datos. Puede estar en KB, MB, GB, TB o PB según el tipo de aplicación que genera o recibe los datos.
- La variedad se refiere a los diversos tipos de datos que pueden existir, por ejemplo, texto, audio, video y fotos.

Los volúmenes son un objetivo en constante movimiento, a partir de 2012 que van desde unas pocas docenas de terabytes a muchos petabytes

de datos en un solo conjunto de datos. Frente a este desafío aparentemente insuperable, las plataformas completamente nuevas se denominan plataformas de *big data*. (pp. 4-6)

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

9. METODOLOGÍA

9.1. Características del estudio

El enfoque del estudio propuesto es mixto, debido a que se obtendrá información agregada acerca del comportamiento de la gente como el uso que le dan al municipio, el lugar donde viven y trabajan, las regiones habitacionales, turísticas y comerciales, entre otros.

El alcance es exploratorio y descriptivo, dado que se busca identificar tendencias y comportamientos de movilidad urbana, analizando las variables de forma independiente.

El diseño adoptado será descriptivo, pues la información de movilidad urbana se analizará en su estado original sin ninguna manipulación; La información se obtendrá a partir de registros de llamadas sin inducir ningún tipo de influencia o control sobre los usuarios. Y para la validación de resultados se realizará un contraste hipótesis para la diferencia de proporciones entre la cantidad de usuarios extrapolados en los registros de llamadas, comparando los resultados con las fuentes oficiales del INE y del INGUAT.

9.2. Unidades de análisis

La población en estudio será la gente que habita o visita el municipio de Antigua Guatemala la cual se encuentra dividida en subpoblaciones dadas por esta misma gente dependiendo en el uso que le den al municipio. Se extraerá una muestra simple aleatoria donde los elementos serán usuarios de la red de

una empresa de telecomunicaciones. Luego se obtendrá todos los registros de llamadas de los usuarios seleccionados que estén dentro de los límites espaciotemporales del problema en cuestión.

9.3. Variables

En la tabla II, se presentan las variables que se utilizarán en el proyecto propuesto.

Tabla II. **Variables del estudio**

Variable	Definición teórica	Definición operativa
Teléfono	Número utilizado para identificar a un dispositivo y/o usuario dentro de la red telefónica. Variable cuantitativa discreta.	Número entero, útil para agregar los registros de llamadas por usuario. Escala de razón
ID celda	Identificador de celda. Variable cuantitativa discreta.	Número entero, útil para agregar los registros de llamadas por celda. Escala de razón
Latitud y Longitud de celda	Coordenadas geográficas que permiten que cada ubicación en la Tierra sea especificada por un conjunto de números. Variable cuantitativa continua.	Número real, en grados sexagesimales. Escala de razón.
Marca de fecha-hora	Registro digital de fecha y la hora del momento de ocurrencia de un evento. Variable cuantitativa	Número real, Escala de razón.
Tipo de celda	Número de celda correspondiente al lugar del hogar y trabajo por usuario. Variable cuantitativa discreta	Número entero, útil para identificar el lugar de trabajo y hogar por usuario.

Continuación tabla II.

Zona geográfica	Región más pequeña en la que se dividirá el municipio de Antigua Guatemala. Variable cuantitativa.	Variable en escala categórica. Posibles valores por definir.
Ubicación de origen	Lugar de origen de turistas y ubicación	Variable en escala categórica. Posibles valores por definir entre departamentos de Guatemala y otros países.
Frecuencia de usuarios	Cantidad de usuarios que utilizan la red en determinada fecha, hora y región. Variable cuantitativa discreta.	Escala de razón.

Fuente: elaboración propia.

9.4. Fases del estudio

La investigación se desarrolla en seis fases entre las cuales se encuentra la revisión de literatura gestión o recolección de la información, exploración y análisis de información, interpretación de información, visualización de información y la redacción de informe final.

9.4.1. Fase 1: Revisión de literatura

Búsqueda de antecedentes y fuentes teóricas para observar el estado del arte en las posibles metodologías por utilizar.

9.4.2. Fase 2: Gestión o recolección de la información

La información se obtendrá a partir de CDR los cuales son recolectados por la empresa de telecomunicaciones y almacenados en su *datawarehouse*. Se hará un muestreo aleatorio simple representativo de la población, basado en las fechas de interés y ubicaciones de las celdas en la región en cuestión.

9.4.3. Fase 3: Exploración y análisis de información

Se realizará el análisis exploratorio obteniendo medidas de tendencia central y dispersión para observar el comportamiento y tipos de variables. Los CDR, generalmente, tienen gran cantidad de variables que no aportan valor para el presente estudio, por lo tanto, se filtrarán.

Se prepararon los datos, anonimizando los números telefónicos. Se extrapolan para obtener las ubicaciones, tomando como base las posiciones geográficas de las celdas utilizadas en cada evento.

9.4.4. Fase 4: Interpretación de información

Se agruparán datos de acuerdo con distintas variables, como mes, franja de horario, día de la semana y ubicación geográfica para obtener tablas de frecuencias.

Se hará las clasificaciones y segmentaciones de los usuarios para determinar el uso que le dan al municipio. Lo mismo se hará con las ubicaciones geográficas para determinar su uso más común.

9.4.5. Fase 5: Visualización de información

Se implementará un *dashboard* para analizar los resultados de forma visual e interactiva. Contendrá diagramas de caja e histogramas para analizar las variables de forma independiente. También se harán gráficos de dispersión y mapas de calor para visualizar de forma más intuitiva el municipio. Se comprobará la validez del estudio comparando con las fuentes oficiales del INE y del INGUAT.

9.4.6. Fase 6: Redacción de informe final

Se redactará el informe final presentando el marco teórico, metodológico, los resultados obtenidos y las conclusiones del estudio.

10. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

Debido al alcance exploratorio del estudio se utilizarán únicamente técnicas de la estadística descriptiva las cuales se detallan a continuación:

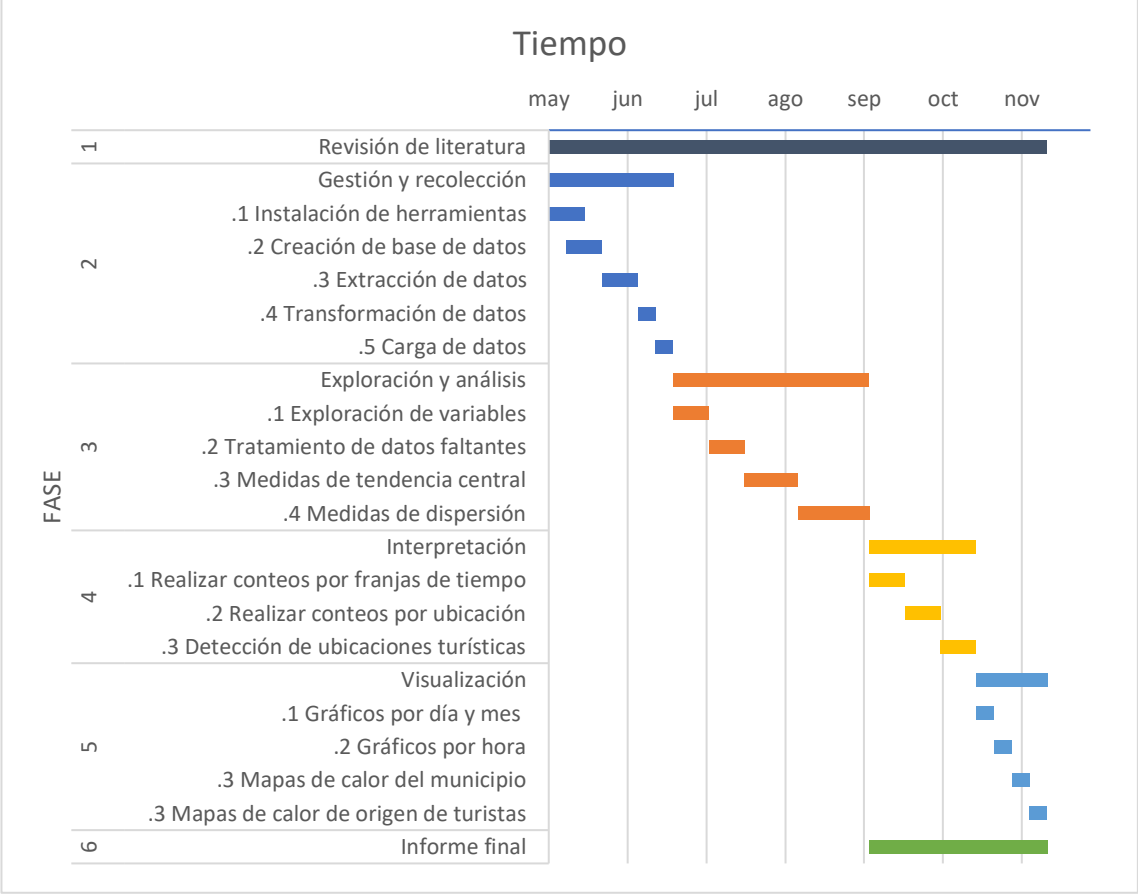
- Medidas de tendencia central y dispersión: se calcularán las medidas de tendencia central como media, moda y mediana, a los conteos de eventos por franja horaria, ubicación, día, entre otros. También dispersión para tener una visión general de los datos. Se busca poder responder preguntas como ver en promedio qué región tiene más eventos, o en qué mes hay más turistas.

- Tablas de frecuencia: se realizarán tablas de frecuencia para presentar la cantidad de eventos que hay a determinadas franjas de horario, mes, días de la semana y ubicaciones geográficas. También se presentará el uso que los usuarios le dan al municipio y el origen de los visitantes en tablas de frecuencias

- Gráficos: los conteos presentados en las tablas de frecuencias se observarán en distintos gráficos como diagramas de caja y mapas de calor para poder analizarlos de forma más intuitiva con respecto a la ubicación geográfica. Por ejemplo, mapas de calor con las ubicaciones con mayor frecuencia en determinado mes, día de la semana y franja de horario, las zonas con mayor presencia turística y los meses con mayor afluencia.

11. CRONOGRAMA

Tabla III. **Cronograma del proyecto**



Fuente: elaboración propia.

12. FACTIBILIDAD DEL ESTUDIO

12.1. Recursos humanos

En los recursos humanos se contemplan a dos personas, el autor del proyecto y el asesor.

12.2. Recursos financieros

En la Tabla IV se presentan los costos de los recursos necesarios para llevar a cabo el proyecto propuesto.

Tabla IV. **Costos de recursos**

Elemento	Unidad	Costo Unitario (Q)	Cantidad necesaria	Costo (Q)
Gestión y recolección de datos				
CDR	registro	1	Por definir	Por definir
Exploración y análisis de información				
Computador personal		10 000	1	10,000
Ambiente en la nube, Watson Studio Cloud Lite de IBM		800	1	800
Watson Studio Cloud Standard		800	1	800
Software, R y Spark		10	1	10
Visualización de información				
Software, Microsoft PowerBI		100	1	100
TOTAL				**Q. 11,710.00**

Fuente: elaboración propia.

12.3. Recursos tecnológicos

Se utilizarán herramientas de análisis de datos y *big data* (R y *Spark*). Y una versión gratuita de Microsoft Power BI para realizar las visualizaciones.

12.4. Acceso a información y permisos

La información será recolectada, por el asesor, en los sistemas de la empresa de telecomunicaciones en la cual labora. La información será anonimizada, resumida y extrapolada para no infringir las políticas de la empresa.

12.5. Equipo e infraestructura

Se utilizará un computador personal y un ambiente en la nube. Para el ambiente en la nube, se presentan dos opciones, una gratis y la otra con un costo de Q800 por mes, en caso se llegue a necesitar más capacidad de cómputo.

13. REFERENCIAS

1. Alhasoun, F., Almaatouq, A., Greco, K., Campari, R., Alfaris, A. y Ratti, C. (agosto, 2014). The city browser: Utilizing massive call data to infer city mobility dynamics. *SIGKDD International Workshop on Urban Computing.* Congreso llevado a cabo en New York, Estados Unidos.

2. Behadili, S., Bertelle, C. y George, L. (enero, 2016). Human Mobility Patterns Modelling using CDRs. *International Journal of UbiComp, 7*(1), 13-19. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3618870

3. Bentley, D. (2017). *Business Intelligence and Analytics.* New York, Estados Unidos: Library Press.

4. Büyükçorak, S., Kurt, G. y Cengaver, O. (febrero, 2014). A Probabilistic Framework for Estimating Call Holding Time Distributions. *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, 63*(2), 811-821. Recuperado de https://ieeexplore.ieee.org/abstract/document/6570525

5. Cohen, Y. y Cohen, J. (2008). *Statistics and Data with R.* West Sussex, Reino Unido: John Wiley & Sons Ltd.

6. Dong, Y., Pinelli, F., Gkoufas, Y., Nabi, Z., Calabrese, F. y Chawla, N. (abril, 2015). Inferring unusual crowd events from mobile phone call

detail records. *Joint European conference on machine learning and knowledge discovery in databases*. Congreso llevado a cabo en Porto, Portugal.

7. Feriancic, G., Celeiro, F. y Silva, L. (junio, 2015). Planejamento da Mobilidade com Big Data de Telefonia Móvel. *20º Congresso Brasileiro de Transporte e Trânsito*. Congreso llevado a cabo en Santos, Brasil.

8. Furletti, B., Gabrielli, L., Giannotti, F., Milli, L., Nanni, M., Pedreschi, D. y Garofalo, G. (junio, 2014). Use of mobile phone data to estimate mobility flows. measuring urban population and inter-city mobility using big data in an integrated approach. *47th SIS Scientific Meeting*. Congreso llevado a cabo en Cagliari, Italia.

9. Gutiérrez, H., y Vara, R. (2008). *Análisis y diseño de experimentos.* Ciudad de México, México: McGraw-Hill. Recuperado de https://gc.scalahed.com/recursos/files/r161r/w19537w/analisis_y_diseno_experimentos.pdf

10. Instituto Nacional de Estadística Guatemala (2019). *Resultados Censo 2018*. Guatemala: Autor. Recuperado de https://censopoblacion.gt/archivos/resultados_censo2018.pdf

11. Instituto Guatemalteco de Turismo (2020). *Perfil del visitante del departamento de Sacatepéquez.* Guatemala: Autor. Recuperado de http://www.inguat.gob.gt/index.php/informacion-estadistica/estadisticas/category/78-2018

12. Junbo, W., Wu, Y., Hsu, H.-H. y Cheng, Z. (2017). *Big Data Analytics for Sensor-Network Collected Intelligence.* Massachusetts, Estados Unidos: Academic Press.

13. Liu, X., y Pöllmann, P. (noviembre, 2020). Dynamic Population Estimation Using Anonymized Mobility Data. *28th International Conference on Advances in Geographic Information Systems.* Congreso llevado a cabo en Seattle, Estados Unidos.

14. Lovelace, R., y Cheshire, J. (marzo, 2014). Introduction to visualising spatial data in R. *National Centre for Research Methods Working Paper. 14*(3), 1-24. Recuperado de https://eprints.ncrm.ac.uk/id/eprint/3295/4/intro_to_R.pdf

15. Pappalardo, L., Pedreschi, D., Smoreda, Z. y Giannotti, F. (diciembre, 2015). Using big data to study the link between human mobility and socio-economic development. *2015 IEEE International Conference on Big Data (Big Data)*. Congreso llevado a cabo en California, Estados Unidos.

16. Penttinen, J. (2015). *The Telecommunications Handbook.* Nueva Jersey, Estados Unidos: John Wiley & Sons, Ltd. Recuperado de https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118678916

17. Prajapati, V. (2013). *Big Data Analytics with R and Hadoop.* Birmingham, Reino Unido: Packt Publishing Ltd. Recuperado de http://index-of.co.uk/Big-Data-Technologies/Big%20Data%20Analytics%20with%20R%20and%20Hadoop.pdf

18. Rosenkrantz, W. (2009). *Introduction to Probability and Statistics for Science, Engineering and Finance*. Florida, Estados Unidos: CRC Press.

19. Ross, S. (2004). *Introduction to Probability and Statistics for Engineers and Scientists*. California, Estados Unidos: Elsevier Inc. Recuperado de http://www.r-5.org/files/books/computers/algo-list/statistics/probability/Sheldon_M_Ross-Introduction_to_Probability_and_Statistics-EN.pdf

20. Saleh, M. y Bahrak, B. (2019). *A Regression Framework for Predicting User's Next Location using Call Detail Records*. Teheran, Irán: Computer Networks.

21. Sarraute, C., y Minnoni, M. (2018). *Brief survey of Mobility Analyses based on Mobile Phone Datasets*. California, Estados Unidos: Grandata Labs.

22. Scheaffer, R., Mendenhall, W., Ott, L. y Gerow, K. (2012). *Elementary Survey Sampling.* Boston, Estados Unidos: Cengage Learning.

23. Singpurwalla, D. (2013). *A Handbook of Statistics*: *An Overview of Statistical Methods.* Londres, Reino Unido: Bookboon. Recuperado de http://thuvienso.bvu.edu.vn/bitstream/TVDHBRVT/15777/1/A-Handbook-of-Statistics.pdf

24. Wang, H., Calabrese, F., Di Lorenzo, G. y Ratti, C. (noviembre, 2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. *13th International IEEE*

Conference on Intelligent Transportation Systems. Congreso llevado a cabo en Funchal, Portugal.

25. Zhou, Y., Lau, B. P., Yuen, C., Tunçer, B., y Wilhelm, E. (diciembre, 2018). Understanding urban human mobility through crowdsensed data. *IEEE Communications Magazine, 56*(11), 52-59. Recuperado de https://arxiv.org/pdf/1805.00628.pdf

14. APÉNDICES

Apéndice 1. **Fechas de inicio y final por actividad**

FASE	ACTIVIDAD	INICIO	DURACIÓN (DÍAS)	FINAL
1	Revisión de literatura	1-may	196	13-nov
2	Gestión y recolección	1-may	49	19-jun
	Instalación de herramientas	1-may	14	15-may
	Creación de base de datos	8-may	14	22-may
	Extracción de datos	22-may	14	5-jun
	Transformación de datos	5-jun	7	12-jun
	Carga de datos	12-jun	7	19-jun
3	Exploración y análisis	19-jun	77	4-sep
	Exploración de variables	19-jun	14	3-jul
	Tratamiento de datos faltantes	3-jul	14	17-jul
	Medidas de tendencia central	17-jul	21	7-ago
	Medidas de dispersión	7-ago	28	4-sep
4	Interpretación	4-sep	42	16-oct
	Realizar conteos por franjas de tiempo	4-sep	14	18-sep
	Realizar conteos por ubicación	18-sep	14	2-oct
	Detección de ubicaciones turísticas	2-oct	14	16-oct
5	Visualización	16-oct	28	13-nov
	Gráficos por día y mes	16-oct	7	23-oct
	Gráficos por hora	23-oct	7	30-oct
	Mapas de calor del municipio	30-oct	7	6-nov
	Mapas de calor de origen de turistas	6-nov	7	13-nov
6	**Informe final**	4-sep	70	13-nov

Fuente: elaboración propia.