



**UNIVERSIDAD DE LA FRONTERA**  
**FACULTAD DE INGENIERÍA Y CIENCIAS**  
**DEPARTAMENTO DE INGENIERÍA DE OBRAS CIVILES**

**“Estimación de un Modelo de Predicción de Frecuencia de Accidentes de Tránsito en  
Intersecciones Urbanas”**

**HUGO IGNACIO MARTÍNEZ MUÑOZ**

**2020**





**UNIVERSIDAD DE LA FRONTERA**  
**FACULTAD DE INGENIERÍA Y CIENCIAS**  
**DEPARTAMENTO DE INGENIERÍA DE OBRAS CIVILES**

**“Estimación de un Modelo de Predicción de Frecuencia de Accidentes de Tránsito en  
Intersecciones Urbanas”**

**TRABAJO PARA OPTAR AL TÍTULO DE INGENIERO CIVIL**

**Profesor Guía:** Maximiliano Exequiel Lizana Maldonado

**HUGO IGNACIO MARTÍNEZ MUÑOZ**  
**2020**

**ESTIMACIÓN DE UN MODELO DE PREDICCIÓN DE FRECUENCIA DE  
ACCIDENTES DE TRÁNSITO EN INTERSECCIONES URBANAS**

**HUGO IGNACIO MARTÍNEZ MUÑOZ**

**COMISIÓN EXAMINADORA**

**MAXIMILIANO EXEQUIEL LIZANA MALDONADO**

**Profesor Guía**

**RENE DUFEU FULLER**

Académico Evaluador

**RODRIGO VICTORIANO HABIT**

Académico Evaluador

**Calificación trabajo escrito :**

**Calificación examen :**

**Calificación final :**

## AGRADECIMIENTOS

En primer lugar, dar las gracias a mi profesor guía Maximiliano Lizana, a los miembros de la comisión evaluadora René Dufeu y Rodrigo Victoriano, y al profesor Tomas Echaveguren de la Universidad de Concepción. Ellos han hecho posible este trabajo de título y han sido parte del proceso, brindando todo el conocimiento y ayuda requerida.

Agradecer a mis profesores de la carrera de Ingeniería Civil de la UFRO por los conocimientos y apoyo brindados a lo largo de estos seis años, especialmente los profesores/as e ingeniero/as Alejandra Calabi, Angel Monsalve, Elisa Gallardo, Pablo Vergara y Gerald Gonzalez, por darme la oportunidad de trabajar con ellos, ya sea como ayudante de cátedra o en distintos proyectos, por la sabiduría entregada y por confiar en mis capacidades. Han sido una parte importante de mi crecimiento como profesional y persona.

Dar las gracias a mis compañeros y amigos de carrera, en especial a mis amigos y equipo titular de estudio, Alfonso, Alfredo y Conrado, con quienes dimos todo en las asignaturas, incluyendo largas noches de estudio, cansancio y muchas risas. Siempre nos hemos apoyado, espero siga siendo así, los quiero amigos. También, dar todo mi apoyo y ánimo a mis compañeros y amigos que están o estarán en este mismo proceso, son muy capaces y estoy seguro que les ira excelente en su vida, tanto profesional como personal.

Finalmente agradecer a mis padres Carolina y Cristian, a mi polola y compañera de muchos años, Angela, por todo el amor y apoyo incondicional, por los valores entregados, por estar siempre que los he necesitado, por siempre tener fe en mí y confiar en mis capacidades. Los amo.

A todos quienes han sido parte de este proceso universitario... ¡Muchas gracias!

## RESUMEN

Los accidentes de tránsito son un problema que crece en gran parte del mundo. Estos son la principal causa de muerte de niños y jóvenes entre 5 y 29 años y genera en la mayoría de los países un costo económico cercano al 3% del PIB. A su vez, en Chile los siniestros de tránsito han ido en aumento cada año con alrededor de 1600 fallecidos al año. Debido a estos antecedentes es que los accidentes de tránsito se califican como un problema prioritario de salud pública en todo el mundo y requieren ser disminuidos y ser la preocupación de usuarios y autoridades. De aquí surgió la motivación del presente estudio, con el objetivo de estimar modelos de predicción de frecuencia de accidentes de tránsito en intersecciones urbanas con el fin de predecir la accidentabilidad en una intersección e identificar los factores de riesgo que generan siniestros de tránsito. Los accidentes de tránsito se definen como eventos estocásticos, es decir, tienen una parte aleatoria y una parte determinística, por lo tanto, es posible prevenir su ocurrencia mediante modelos predictivos basados en técnicas econométricas.

La metodología incorporo un proceso que comenzó con el estudio del estado del arte hasta la realización de la modelación de la cantidad de accidentes en las entidades consideradas. Dichos modelos se generaron a partir de la base de datos del “Gran Concepción”, la cual contempló un total de 253 intersecciones semaforizadas de las comunas de Chiguayante, Concepción, Hualqui, San Pedro de la Paz, Talcahuano y Tomé. Las variables dependientes fueron estimadas separadamente considerando accidentes totales, severos y leves de las entidades durante 5 años. Las variables independientes explicativas utilizadas fueron variables de tipo infraestructura, operación y entorno que se seleccionaron de la base de datos. La estimación de accidentes se realizó para la base completa de intersecciones y para la subdivisión de la base de datos en dos *clusters* de intersecciones obtenidos de un proceso de agrupación con K-medios. Los modelos se desarrollaron a partir de una regresión Binomial Negativa (NB) en complemento con el método Empírico Bayesiano (EB).

En este trabajo se observó que el método Empírico Bayesiano utilizado generó un aumento significativo en los ajustes de todos los modelos de predicción de accidentes elaborados, siendo eficaz en mejorar la precisión de los modelos. Por otro lado, la agrupación de las intersecciones en dos grupos generó aumento en los ajustes solo para el *cluster* con mayor cantidad de datos, mientras

que el otro *cluster*, que consideró menor cantidad de intersecciones, no obtuvo buenos ajustes. En este sentido, el número de grupos a utilizar es recomendable que sea bajo para evitar obtener *clusters* con poca cantidad de entidades y posteriormente deriven en bajos ajustes de los modelos. En cuanto al modelo Binomial Negativo utilizado, se confirma su efectividad frente a datos con sobredispersión como es el caso de los accidentes totales y leves, mientras que, en los accidentes severos, cuyo set de datos no presenta sobredispersión, los ajustes fueron bajos.

Las variables explicativas de infraestructura que resultaron más significativas en los modelos fueron el número de pistas de la rama secundaria, la presencia de mediana en la rama principal y la presencia de mediana en la rama secundaria. El número de pistas de la rama secundaria resultó en una correlación positiva en la generación de accidentes de tránsito, por lo tanto, a mayor cantidad de pistas existe menor seguridad en la intersección. Por otro lado, la presencia de mediana en la rama principal y en la rama secundaria resultaron en una correlación negativa en la generación de accidentes de tránsito, es decir, la presencia de mediana es una variable que aumenta la seguridad en la intersección. Las variables explicativas de operación que resultaron más significativas en los modelos fueron el flujo total de la rama principal, flujo total de la rama secundaria y flujo total de giro derecha. Todas ellas resultaron en una correlación positiva en la generación de accidentes de tránsito, por lo tanto, a mayor TMD existe menor seguridad en la intersección. La variable explicativa de entorno que resultó más significativa en los modelos fue la presencia de tipo de zona (uso de suelo) comercial o mixta en los alrededores de la intersección. Esta resultó principalmente en una correlación positiva en la generación de accidentes de tránsito, por lo tanto, las zonas comerciales son intersecciones menos seguras y más vulnerables a la ocurrencia de siniestros de tránsito.

En definitiva, el estudio sugiere que es recomendable la presencia de infraestructura como medianas y pistas de viraje exclusivo que aumentan la seguridad de la intersección y disminuyen el riesgo de siniestros. Las intersecciones con flujos altos tienen mayor riesgo de accidentes, por lo tanto, es recomendable la implementación de elementos de seguridad, que protejan a los usuarios más vulnerables de la vía. El flujo de giro derecha resultó principalmente generador de accidentes severos, por lo tanto, es recomendable la instalación de señaléticas que adviertan al usuario conductor de la presencia de ciclistas y/o peatones.

## **Índice de contenidos**

Capítulo 1. Introducción	2
1.1. Motivación	2
1.2. Descripción del problema	3
1.3. Objetivos	5
1.3.1. Objetivo general	5
1.3.2. Objetivos específicos	5
Capítulo 2. Antecedentes Generales	7
2.1. Modelos de predicción y clasificación	7
2.2. Método Empírico Bayesiano (EB)	13
2.3. Variables dependientes (a explicar)	14
2.4. Variables independientes (explicativas)	14
2.5. Problemas metodológicos	16
Capítulo 3. Metodología	20
3.1. Identificación modelos de predicción y variables explicativas	20
3.2. Formulación del modelo	21
3.2.1. Modelo de regresión Binomial Negativo	21
3.2.2. Método Empírico Bayesiano (EB)	22
3.3. Selección de variables explicativas	23
3.4. Evaluación del modelo	24
3.5. Zona de estudio	26
3.6. Base de datos de accidentes	27
3.6.1. Variables dependientes	28
3.6.2. Variables independientes	29
3.7. Agrupación de intersecciones	32
Capítulo 4. Resultados y discusión	39
4.1. Resultados respecto a accidentes totales	39
4.2. Resultados respecto a accidentes severos	45
4.3. Resultados respecto a accidentes leves	51
4.4. Comparación de modelos	57
4.5. Interpretación, análisis y discusión respecto a las variables explicativas	58
4.5.1. Variables explicativas de infraestructura	61
4.5.2. Variables explicativas de operación	64
4.5.3. Variables explicativas de entorno	65
Capítulo 5. Conclusiones	68
Bibliografía	71
Anexos	74

## Índice de tablas

<b>Tabla 2.1.</b> Resumen de estudios previos.	11
<b>Tabla 2.2.</b> Resumen de estudios previos (Continuación).	12
<b>Tabla 2.3.</b> Variables significativas en literatura.	15
<b>Tabla 2.4.</b> Resumen de problemas metodológicos en datos de accidentes de tránsito. (Lord y Mannering, 2010).	18
<b>Tabla 3.1.</b> Detalle de cantidad de intersecciones por comuna presentes en la base de datos.	27
<b>Tabla 3.2.</b> Variables dependientes a estimar en los modelos.	28
<b>Tabla 3.3.</b> Análisis descriptivo de variables dependientes.	29
<b>Tabla 3.4.</b> Variables explicativas de infraestructura seleccionadas.	30
<b>Tabla 3.5.</b> Análisis descriptivo de variables explicativas de infraestructura seleccionadas.	30
<b>Tabla 3.6.</b> Variables explicativas de operación seleccionadas.	31
<b>Tabla 3.7.</b> Análisis descriptivo de variables explicativas de operación seleccionadas.	31
<b>Tabla 3.8.</b> Variables explicativas de entorno seleccionadas.	31
<b>Tabla 3.9.</b> Análisis descriptivo de variables explicativas de entorno seleccionadas.	32
<b>Tabla 3.10.</b> Variables explicativas seleccionadas para agrupación.	33
<b>Tabla 3.11.</b> Ajuste de modelo de agrupación K-means en software JASP.	33
<b>Tabla 3.12.</b> Información de agrupación K-means en software JASP.	34
<b>Tabla 3.13.</b> Análisis descriptivo de variables predictoras por cluster.	36
<b>Tabla 4.1.</b> Modelo de predicción de accidentes totales en base completa.	40
<b>Tabla 4.2.</b> Modelo de predicción de accidentes totales en intersecciones del Cluster 1.	41
<b>Tabla 4.3.</b> Modelo de predicción de accidentes totales en intersecciones del Cluster 2.	43
<b>Tabla 4.4.</b> Modelo de predicción de accidentes severos en base completa.	46
<b>Tabla 4.5.</b> Modelo de predicción de accidentes severos en intersecciones del Cluster 1.	47
<b>Tabla 4.6.</b> Modelo de predicción de accidentes severos en intersecciones del Cluster 2.	49
<b>Tabla 4.7.</b> Modelo de predicción de accidentes leves en base completa.	52
<b>Tabla 4.8.</b> Modelo de predicción de accidentes leves en intersecciones del Cluster 1.	53
<b>Tabla 4.9.</b> Modelo de predicción de accidentes leves en intersecciones del Cluster 2.	55
<b>Tabla 4.10.</b> Resumen de coeficientes de determinación corregidos ( $R^2_{corr}$ ) normal y EB en los casos modelados.	58
<b>Tabla 4.11.</b> Resumen de variables explicativas resultantes de los modelos.	59
<b>Tabla 4.12.</b> Porcentaje de aumento o disminución por variable en la frecuencia de accidentes.	60
<b>Tabla B. 1.</b> Variables explicativas de infraestructura presentes en BD completa.	77
<b>Tabla B. 2.</b> Variables explicativas de operación presentes en BD completa.	78
<b>Tabla B. 3.</b> Variables explicativas de entorno presentes en BD completa.	79

## Índice de figuras

<b>Figura 2.1.</b> Línea del tiempo de los últimos diez años de estudios en el tema.	12
<b>Figura 3.1.</b> Mapa emplazamiento de intersecciones por comuna presentes en la base de datos.	26
<b>Figura 3.2.</b> Medias (centroides) de predictores por cluster en software JASP.	34
<b>Figura 3.3.</b> Clusters en el espacio de características, t-SNE en software JASP.	35
<b>Figura 3.4.</b> Mapa emplazamiento de intersecciones por cluster en zona de estudio.	37
<b>Figura 4.1.</b> Representación gráfica de modelo de predicción de accidentes totales en base completa. Accidentes estimados vs. Accidentes observados.	40
<b>Figura 4.2.</b> Representación gráfica de modelo de predicción de accidentes totales en base completa. Accidentes esperados (EB) vs. Accidentes observados.	41
<b>Figura 4.3.</b> Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.	42
<b>Figura 4.4.</b> Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.	42
<b>Figura 4.5.</b> Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados	43
<b>Figura 4.6.</b> Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.	44
<b>Figura 4.7.</b> Representación gráfica de modelo de predicción de accidentes severos en todas intersecciones. Accidentes estimados vs. Accidentes observados.	46
<b>Figura 4.8.</b> Representación gráfica de modelo de predicción de accidentes severos en todas intersecciones. Accidentes esperados (EB) vs. Accidentes observados.	47
<b>Figura 4.9.</b> Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.	48
<b>Figura 4.10.</b> Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.	48
<b>Figura 4.11.</b> Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados.	49
<b>Figura 4.12.</b> Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.	50
<b>Figura 4.13.</b> Representación gráfica de modelo de predicción de accidentes leves en base completa. Accidentes estimados vs. Accidentes observados.	52
<b>Figura 4.14.</b> Representación gráfica de modelo de predicción de accidentes leves en base completa. Accidentes esperados (EB) vs. Accidentes observados.	53
<b>Figura 4.15.</b> Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.	54
<b>Figura 4.16.</b> Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.	54
<b>Figura 4.17.</b> Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados.	55
<b>Figura 4.18.</b> Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.	56

<b>Figura A. 1.</b> Mapa emplazamiento de intersecciones en comuna de Chiguayante.	74
<b>Figura A. 2.</b> Mapa emplazamiento de intersecciones en comuna de Concepción.	74
<b>Figura A. 3.</b> Mapa emplazamiento de intersecciones en comuna de Hualqui.	75
<b>Figura A. 4.</b> Mapa emplazamiento de intersecciones en comuna de San Pedro de la Paz.	75
<b>Figura A. 5.</b> Mapa emplazamiento de intersecciones en comuna de Talcahuano.	76
<b>Figura A. 6.</b> Mapa emplazamiento de intersecciones en comuna de Tomé.	76
<b>Figura C. 1.</b> Mapa emplazamiento de intersecciones por cluster en Chiguayante.	79
<b>Figura C. 2.</b> Mapa emplazamiento de intersecciones por cluster en Concepción.	80
<b>Figura C. 3.</b> Mapa emplazamiento de intersecciones por cluster en Hualqui.	80
<b>Figura C. 4.</b> Mapa emplazamiento de intersecciones por cluster en San Pedro de la Paz.	81
<b>Figura C. 5.</b> Mapa emplazamiento de intersecciones por cluster en Talcahuano.	81
<b>Figura C. 6.</b> Mapa emplazamiento de intersecciones por cluster en Tomé.	82

# **CAPÍTULO 1**

## **INTRODUCCIÓN**

## **Capítulo 1. Introducción**

### **1.1. Motivación**

Según la OMS (2018) los accidentes de tránsito son la principal causa de muerte de niños y jóvenes entre 5 y 29 años. Actualmente en el mundo, según cifras del mismo organismo, entre 20 y 50 millones de personas sufren traumatismos en accidentes de tránsito, de los cuales una proporción importante provocan alguna discapacidad. Además, alrededor de 1.35 millones de personas mueren cada año en el mundo por causa de accidentes de tránsito, donde los principales afectados son los usuarios vulnerables de la vía pública (peatones, ciclistas y motociclistas) y generando, en la mayoría de los países, un costo económico cercano al 3% del PIB, el cual podría ser aprovechado en otras necesidades de la ciudadanía (OMS, 2018). El gran crecimiento del tráfico junto con la expansión no planificada de las ciudades ha generado un crecimiento excesivo en la tasa de accidentes de tránsito, derivando en un costo socioeconómico alto y siendo una carga para la salud, economía y reducción de la pobreza (Islam *et al.*, 2019). Debido a estos antecedentes es que los accidentes de tránsito se califican como un problema prioritario de salud pública en todo el mundo (OMS, 2018).

La situación en Chile no es distinta a la del resto del mundo. En Chile, los siniestros de tránsito han ido en aumento cada año, bordeando los 27 mil en 1972 hasta 90 mil para el año 2018 con diversas gravedades (CONASET, 2019). Esto se debe a que, tal como informa CONASET (2019) en su base de datos, tanto la población como el parque vehicular han ido aumentando a través de los años, pero este último a una tasa mucho mayor, lo cual se ve reflejado en las cifras de tasa de motorización, donde en el año 1972 era de 24.7 personas por vehículo y actualmente en el año 2018 esta fue de 3.4 personas por vehículo. Cabe destacar que este aumento de accidentes de tránsito se debe principalmente al aumento de accidentes con lesiones de tipo “leves” y los accidentes con lesiones “menos graves” y “graves” se han mantenido a lo largo de los años. Actualmente en el país, el 18% de los siniestros son “graves”, el 14% “menos graves”, el 68% “leves” y anualmente fallecen alrededor de 1600 personas (CONASET, 2019). En definitiva, los accidentes de tránsito en Chile y todo el mundo requieren ser disminuidos y ser la preocupación de usuarios y autoridades.

Los accidentes de tránsito se deben a factores que abarcan un conjunto de dimensiones, tanto errores humanos como problemas en cuanto a la infraestructura vial. Entre los principales factores que da cuenta la OMS (2018) se encuentra la velocidad, que al aumentar en 1% la velocidad media del vehículo puede producir un aumento del 4% en accidentes mortales; la conducción bajo los efectos del alcohol o sustancias psicoactivas; la no utilización de implementos de seguridad como cascos, cinturones y sillas para niños; la infraestructura vial insegura, que a medida que se desarrollan los países esta va mejorando en temas de seguridad vial; el incumplimiento de las normas de tránsito por parte del usuario; y finalmente la conducción distraída, factor que ha tenido un aumento exponencial en las últimas décadas y se debe principalmente al uso del teléfono móvil en la conducción (OMS, 2018). En definitiva, se requiere tanto de usuarios responsables y atentos a las condiciones de tránsito, como también de una infraestructura vial adecuada que permita la disminución de los siniestros. Además, debido a los muchos movimientos de giro y puntos de conflicto resultantes, no es sorprendente que las intersecciones se encuentren entre los lugares más peligrosos de las redes viales (Chen y Xie, 2016). Agencias como la AASHTO, consideran como prioridad máxima la mejora de la seguridad en intersecciones, existiendo una necesidad urgente de investigar y mejorar la seguridad del tráfico en las intersecciones señalizadas (Guo *et al.*, 2010).

En base a todos los antecedentes entregados anteriormente, surge la motivación e importancia del presente estudio, en el cual se realiza la “estimación de modelos de predicción de frecuencia de accidentes de tránsito en intersecciones urbanas a partir de variables de infraestructura, operación y entorno” a modo de estimar la accidentabilidad presente en una intersección urbana y en especial lograr identificar los factores y variables de riesgo que generan accidentes de tránsito. El fin de este estudio fue lograr un aporte para que las autoridades y los consultores (diseñadores viales) puedan brindar una mayor seguridad vial para el tránsito mediante diseños más seguros en este ámbito.

## **1.2. Descripción del problema**

Un accidente de tránsito, según CONASET (2017), es un suceso eventual en el cual involuntariamente resultan daños y se le asocia un fenómeno casual, impredecible e incontrolable que resulta en un siniestro que ocurre en una vía e involucra al menos un vehículo en movimiento y resulta en una o más personas ilesas, lesionadas o fallecidas (CONASET, 2017). Sin embargo, según CONASET (2017), un accidente de tránsito no debiese ser llamado “accidente”, ya que estos

tienen poco de impredecible y se pueden evitar. De hecho, los accidentes de tránsito se definen, según Chías (2004), como eventos estocásticos, es decir, tienen una parte aleatoria y una parte determinística, por lo tanto, es posible prevenir su ocurrencia o realizar un análisis probabilístico (Dufeu, 2019). Es por esto que modelos predictivos basados en técnicas econométricas han sido implementados para la predicción de accidentes.

El principal problema al momento de predecir accidentes es definir el tipo de modelo que podría lograr el mejor ajuste. Entre las técnicas para análisis predictivo se encuentran, en general, las de regresión y las de aprendizaje computacional. Las técnicas de regresión se basan en establecer una ecuación matemática como modelo para representar interacciones entre variables. Entre los modelos de regresión más comunes se encuentran la regresión lineal, árboles de clasificación, etc. En cuanto al aprendizaje computacional, estos modelos son capaces de “ignorar” las relaciones entre variables, debido a que generalmente estas pueden ser muy complejas y la forma matemática es desconocida, por lo tanto, esta técnica de predicción emula la cognición humana y aprende a partir de ejemplos de entrenamientos para luego predecir las respuestas futuras. Entre estos modelos de aprendizaje computacional se encuentran las redes neuronales, método de Naive Bayes, aprendizaje de máquinas, etc. (Espino, 2017). En resumen, en la literatura existe una gran cantidad de modelos de estimación con diferentes consideraciones.

Para la modelación de accidentes de tránsito, en la literatura previa se han probado y propuesto distintos modelos por variados autores. La experiencia internacional indica que es recomendable realizar desagregaciones de los datos de accidentes, por ejemplo, por gravedad, ya que esto permite lidiar con la correlación de las observaciones. La literatura indica que la distribución Binomial Negativa (NB, también llamada Poisson-Gamma PG), que es una modificación a la distribución Poisson, sería consistente con la sobre dispersión que existe usualmente en los datos de accidentes de tránsito (SECTRA, 2014). En definitiva, se ha observado un consenso entre los especialistas en que la distribución Binomial Negativa suele ser la más adecuada y con mejor ajuste.

Como se ha mencionado, hay gran variedad de modelos estudiados y propuestos por autores, por lo tanto, la problemática a resolver es la estimación y la calibración de los modelos para cada situación local, ya que los distintos países en todo el mundo suelen presentar distintas características en cuanto a infraestructura, operación y entorno, además de existir diferencias en el

comportamiento de los usuarios, principalmente por los niveles de educación vial. Es por esto que uno de los objetivos principales de los estudios de seguridad en intersecciones es identificar los factores de alto riesgo en la generación de accidentes de tránsito entre las distintas variables de infraestructura, operación y entorno (Guo *et al.*, 2010), lo cual será aplicado para la zona del “Gran Concepción”.

### **1.3. Objetivos**

A partir de la problemática expuesta anteriormente, en el presente trabajo se pretenden lograr los objetivos que se mencionan a continuación.

#### **1.3.1. Objetivo general**

Estimar frecuencia de accidentes de tránsito en intersecciones urbanas mediante modelos de predicción a partir de variables de infraestructura, operación y entorno.

#### **1.3.2. Objetivos específicos**

- a) Identificar variables explicativas de importancia para el estudio.
- b) Elaborar una agrupación de intersecciones a partir de variables de infraestructura, operación y entorno.
- c) Desarrollar modelos de predicción de frecuencia de accidentes para cada grupo de intersecciones a partir de variables de infraestructura, operación y entorno.
- d) Analizar los factores con mayor influencia y/o efecto en la frecuencia de accidentes de tránsito.

**CAPÍTULO 2**  
**ANTECEDENTES GENERALES**

## **Capítulo 2. Antecedentes Generales**

En las siguientes secciones se encuentra una profunda revisión bibliográfica de los antecedentes que son necesarios manejar o tener presente para una mejor comprensión del trabajo realizado. Entre estos antecedentes se encuentran los diversos estudios previos en el tema de modelación de predicción de accidentes y clasificación, método Empírico Bayesiano, las variables dependientes modeladas en la literatura y las variables independientes utilizadas para explicar dichos fenómenos, además de los problemas metodológicos más usuales al momento de modelar accidentes de tránsito.

### **2.1. Modelos de predicción y clasificación**

A lo largo de los años, en la literatura se ha encontrado diversas técnicas y metodologías que implica un análisis de predicción de accidentes.

En cuanto a los métodos de estimación de parámetros para predicción de choques, los dos más utilizados son el de Máxima Verosimilitud (Maximum likelihood) y el método Bayesiano (Lord y Mannering, 2010).

Como se mencionó antes, también existen modelos de inteligencia artificial (AI) como redes neuronales y aprendizaje computacional. En el artículo de Kononov *et al.* (2012) se menciona que las redes neuronales no están limitadas por una forma funcional preseleccionada y supuestos de distribución específicos y que una forma de la función de respuesta utilizada en muchas aplicaciones de ingeniería es un modelo de red neuronal de avance con una sola capa de unidades ocultas (Kononov *et al.*, 2012). Es por esto que no es recomendable el uso de los métodos como redes neuronales o aprendizaje computacional, ya que estos funcionan bajo el mecanismo de “caja negra” en los cuales se desconoce qué es lo ocurre y como se infieren los resultados. Por lo tanto, los modelos de aprendizaje computacional son adecuados en problemas de predicción en los cuales no es de interés conocer la relación entre los valores de entrada y de salida, como es el caso del análisis del mercado de valores, detección de fraudes de tarjetas de crédito, reconocimiento facial, etc. (Espino, 2017).

Varios de los modelos utilizados en la literatura son explicados por Lord y Mannering (2010), entre los cuales se encuentran los modelos Poisson, Binomial Negativo (Poisson-Gamma), Poisson-LogNormal, Zero-Inflated Poisson y Zero-Inflated Negative Binomial, Conway-Maxwell-Poisson, Gamma, Redes Neuronales, Red Neuronal Bayesiana, etc.

A continuación, se describen las metodologías utilizadas por distintas entidades y autores en relación a modelos de predicción de accidentes de tránsito para el periodo de los últimos diez años.

La American Association of State Highway and Transportation Officials (AASHTO) ha establecido un manual titulado “Highway Safety Manual” (HSM) en 2010, en la cual se presentan recomendaciones para los profesionales y pretende ser un estándar de práctica. En el HSM se utilizan los métodos Safety Performance Functions (SPFs) para la estimación de accidentes de tránsito y se desarrollan asumiendo la distribución NB para el número observado de choques (Cunto *et al.*, 2015). En el artículo de Cunto *et al.*, (2015) se presentan los resultados con respecto a la transferibilidad de los componentes del método predictivo del HSM para estimar la frecuencia promedio esperada de choques de intersecciones urbanas en la ciudad de Fortaleza, Brasil. Tasic *et al.*, (2016) realizaron el análisis de frecuencia de choque esperado choques entre vehículos, peatones y bicicletas. En este caso se utilizó la distribución NB con un modelo del tipo Generalized Additive Models (GAM) que son una extensión de los Modelos Lineales Generalizados (GLM) y se desarrollan con predictores lineales, pero incluyendo una suma de funciones de suavizado de las variables explicativas. En dicho estudio, a modo de comparación, se utilizó el Bayesian Hierarchical Model en el cual se asume la distribución Poisson. La desventaja de ambos modelos mencionados es que puede que no sean fácilmente transferibles a otros conjuntos de datos (Lord y Mannering, 2010).

Varios estudios sugieren que las formas funcionales no lineales son las más adecuadas. Un ejemplo de esto es el artículo de Chen y Xie (2016), en el cual se utilizaron los modelos GAM y modelos de regresión Piecewise Linear Negative Binomial (PLNB) para el ajuste de los datos de accidentes en intersecciones. En este estudio se concluyó que los modelos PLNB proporcionan un mejor rendimiento de bondad de ajuste que los GAM. Además, se menciona que la proporción del Annual Average Daily Traffic (AADT) menor y mayor de la intersección tiene un impacto variable en la seguridad de la intersección y merece más investigaciones. En el caso de Rizaldi *et al.* (2017) y

Wang *et al.* (2017) analizaron autopistas con los modelos de predicción Zero Inflated Negative Binomial (ZINB) y Poisson Log-Normal (PLN) respectivamente. La desventaja de estos modelos es que pueden verse afectados negativamente por la baja media muestral y el sesgo de tamaño en muestras pequeñas (Lord y Mannering, 2010). Alarifi *et al.* (2018) realizaron la predicción del conteo de choques por tipo en intersecciones y en corredores mediante modelos de articulación espacial Hierarchical Poisson-LogNormal (HPLN), de forma multivariante y univariante. Las desventajas de los modelos Hierarchical es que puede que no sea fácilmente transferible a otros conjuntos de datos y los resultados de correlación pueden ser difíciles de interpretar (Lord y Mannering, 2010). Los resultados indicaron que el modelo multivariante supera al modelo univariante. Además, las variables explicativas significativas y la magnitud de las estimaciones de los parámetros para la misma variable independiente son diferentes entre los tipos de accidentes. Los resultados enfatizan la necesidad de estimar los recuentos de accidentes por tipo en una forma multivariante para detectar mejor los problemas y proporcionar las contramedidas adecuadas (Alarifi *et al.*, 2018).

En el estudio de Mafi *et al.* (2018) se investigó cómo la gravedad de las lesiones difiere entre los conductores de varias edades y grupos de género utilizando modelos sensibles al costo (consecuencias) de predicciones incorrectas al desarrollar modelos de aprendizaje automático (Machine Learning) como son C4.5, Instance-Bases (IB) y Random Forest (RF). Los resultados de dicho trabajo indicaron que los clasificadores de aprendizaje sensibles a los costos eran superiores a los clasificadores regulares para predecir con precisión las lesiones y muertes por accidentes, siendo el mejor el modelo RF. En el artículo de Islam *et al.* (2019) se investigan los efectos de los atributos no geométricos en las intersecciones urbanas señalizadas de los países en desarrollo a través de Modelos de Predicción de Accidentes (CPM) en el marco Empírico Bayesiano (EB). Los autores de ese artículo esperan que los modelos propuestos se parezcan a otras intersecciones señalizadas urbanas de países en desarrollo como Bangladesh. Los mayores desafíos se encuentran en las predicciones de choques en tiempo real. Yuan *et al.* (2019) intentan verificar la viabilidad de la predicción del riesgo de accidentes en tiempo real mediante el uso de un algoritmo Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) con un conjunto de datos sobre muestreados con la Técnica de Sobre Muestreo Minoritario Sintético (SMOTE).

Otro punto de vista en cuanto a modelación, es la predicción de la severidad de un accidente de tránsito. El estudio realizado por Mohammadian *et al.* (2020) busca predecir la gravedad de los choques de objetos fijos entre conductores de edad avanzada mediante la comparación de dos tipos de técnicas de Inteligencia Artificial (AI), la Red Neuronal Artificial (ANN) y el Algoritmo Genético Inteligente Híbrido (GA en conjunto con ANN). Aunque los resultados de dicho estudio mostraron que el ANN desarrollado superó al Algoritmo Genético Inteligente Híbrido, el enfoque híbrido fue más capaz de predecir choques de alta gravedad. Esto se basa en la forma en que se entrenó el modelo híbrido aprovechando el Algoritmo Genético (GA).

Una metodología interesante es la propuesta por Pooyan *et al.* (2020) donde se realiza un modelo conjunto entre recuento y gravedad de choques. El conteo de choques es una medida de la probabilidad de que ocurra un evento potencialmente dañino, mientras que la gravedad del choque es una medida del impacto social y el daño a la sociedad. Los resultados del mencionado trabajo muestran que el modelo conjunto ha mejorado sustancialmente la precisión de la predicción e identificación de un mayor número de accidentes fatales y graves en comparación con los modelos de recuento tradicionales.

En la Tabla 2.1 y Tabla 2.2 se presenta un resumen de los estudios previos relacionados a los temas expuestos y a los modelos revisados, algunos de estos ya fueron mencionados y otros no.

En resumen, en diversos trabajos estudiados se han utilizado modelos de regresión y de aprendizaje computacional. Se observa que los modelos de regresión se suelen utilizar en la predicción de frecuencia de accidentes, en cambio los modelos de aprendizaje computacional se utilizan más en temas de clasificación y desagregación de los datos, predicción de clasificación por gravedad de los accidentes, además de la predicción en tiempo real del riesgo de accidentes de tránsito. Esto se debe a que un modelo de aprendizaje computacional se entrena y predice los resultados sin dar una explicación en base a la influencia de las variables asociadas.

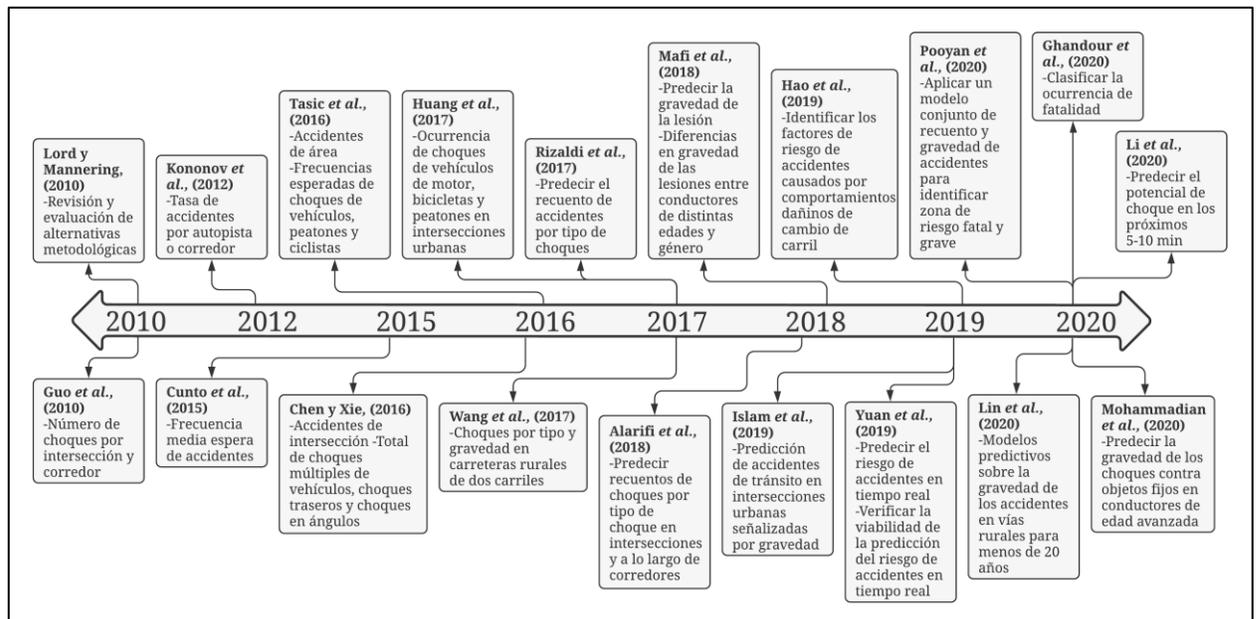
**Tabla 2.1.** Resumen de estudios previos.

<b>Fuente/Autor(es)</b>	<b>Método/Modelo/Algoritmo</b>	<b>Tema/Variable estimada/Modelado</b>
Lord y Mannering (2010)	Varios	-Revisión y evaluación de alternativas metodológicas.
Guo <i>et al.</i> (2010)	-Poisson Bayesiano. - Binomial Negativo (NB) Bayesiano.	-Número de choques por intersección y corredor.
Kononov <i>et al.</i> (2012)	-Funciones de Rendimiento de Seguridad (SPFs). -Redes Neuronales.	-Tasa de accidentes por autopista o corredor (medida en número de accidentes por millón de millas recorridas por vehículos).
Cunto <i>et al.</i> (2015)	-SPFs de Manual de Seguridad Vial (HSM). -Empírico Bayesiano (EB).	-Frecuencia media esperada de accidentes.
Tasic <i>et al.</i> (2016)	-Modelos Aditivos Generalizados (GAMs). -NB. -Bayesiano Jerárquico.	-Accidentes de área. -Frecuencias esperadas de choques de vehículos, peatones y ciclistas.
Chen y Xie (2016)	-GAMs. -Binomial Negativo Lineal por partes (PLNB).	-Accidentes de intersección. -Total de choques múltiples de vehículos, choques traseros y choques en ángulo.
Huang <i>et al.</i> (2017)	-Poisson Log-Normal Univariante. -Poisson Log-Normal Multivariante.	-Ocurrencia de choques de vehículos de motor, bicicletas y peatones en intersecciones urbanas.
Wang <i>et al.</i> (2017)	-Poisson Log-Normal Multivariante.	-Choques por tipo y gravedad en carreteras rurales de dos carriles.
Rizaldi <i>et al.</i> (2017)	-Binomial Negativo Cero Inflado (ZINB).	-Predecir el recuento de accidentes por tipo de choques.
Alarifi <i>et al.</i> (2018)	-Conjunto Espacial Jerárquico Multivariante Bayesiano.	-Predecir recuentos de choques por tipo de choque en intersecciones y segmentos a lo largo de corredores.
Mafi <i>et al.</i> (2018)	-C4.5. -Basado en Instancia (IB). -Bosque Aleatorio (RF).	-Predecir la gravedad de la lesión. -Diferencias en gravedad de las lesiones entre conductores de distintas edades y grupos de género.
Islam <i>et al.</i> (2019)	-NB. -EB.	-Predicción de accidentes tránsito en intersecciones urbanas señalizadas por gravedad.
Hao <i>et al.</i> (2019)	-Tiempo de Aproximación de Vehículos (VAT). -Regresión de Ángulo Mínimo (Lasso-LARS).	-Identificar los factores de riesgo de accidentes causados por comportamientos dañinos de cambio de carril.
Yuan <i>et al.</i> (2019)	-Red Neuronal Recurrente de Memoria a Corto Plazo (LSTM-RNN). -Técnica de Sobre Muestreo de Minorías Sintéticas (SMOTE). -Modelo Logístico Condicional, Bernoulli.	-Predecir el riesgo de accidentes en tiempo real. -Verificar la viabilidad de la predicción del riesgo de accidentes en tiempo real.
Pooyan <i>et al.</i> (2020)	-Conteo de choques: Parámetros Aleatorios Binomial Negativo (RPNB). -Severidad: Logit. -EB	-Aplicar un modelo conjunto de recuento y gravedad de accidentes para identificar los segmentos de la carretera con alto riesgo de accidentes fatales y con lesiones graves.

**Tabla 2.2.** Resumen de estudios previos (Continuación).

Fuente/Autor(es)	Método/Modelo/Algoritmo	Tema/Variable estimada/Modelado
Lin <i>et al.</i> (2020)	-RF. -Árbol de Decisión de Aumento de Gradiente (GBDT).	-Modelos predictivos sobre la gravedad de los accidentes en vías rurales para menores de 20 años.
Ghandour <i>et al.</i> (2020)	-Optimización Mínima Secuencial (SMO). -Arboles de decisión J48. -RF, Red Neuronal Artificial (ANN), Regresión Logística, Naive Bayes.	-Clasificar la ocurrencia de fatalidad.
Mohammadian <i>et al.</i> (2020)	-ANN. -Algoritmo Genético (GA). -Algoritmo Genético Inteligente Híbrido.	-Predecir la gravedad de los choques contra objetos fijos en conductores de edad avanzada.
Li <i>et al.</i> (2020)	-LSTM.	-Predecir el potencial de choque en los próximos 5-10 min.

Finalmente, en la Figura 2.1 se presenta una línea temporal de los estudios previos de los últimos 10 años mostrados anteriormente. En esta se observa que en los últimos años en los estudios se han centrado y dado un mayor énfasis a la predicción de gravedad del siniestro, más que a la cantidad o frecuencia de este. Esto se debe a la mayor accesibilidad de herramientas de inteligencia artificial que facilitan dicha modelación.



**Figura 2.1.** Línea del tiempo de los últimos diez años de estudios en el tema.

## 2.2. Método Empírico Bayesiano (EB)

El método Empírico Bayesiano (EB) se utiliza para refinar la estimación del número esperado de accidentes de tránsito en una entidad (que corresponde a una intersección en este trabajo) combinando el número observado de accidentes con el número estimado por un modelo (Islam *et al.*, 2019).

Como se explica en el estudio de Hauer *et al.* (2002), el método consiste en hacer uso conjunto de dos aspectos de seguridad de una entidad: el historial de accidentes y la frecuencia de accidentes estimada en entidades similares. Dicha frecuencia de accidentes estimada en entidades similares proviene de un modelo de predicción. Según Hauer *et al.* (2002) este uso conjunto de ambos aspectos se implementa en la estimación EB mediante un promedio ponderado donde se utiliza un término matemático llamado “peso” (weight) y su cálculo requiere el uso del parámetro de sobredispersión del modelo utilizado.

Diversos autores han hecho uso de esta metodología. Miao y Lord (2003) incluyen el método EB en la modelación de la relación entre accidentes de tránsito y flujos de tráfico para intersecciones de Toronto en Canadá, mencionando que este método generalmente resulta en una buena precisión. Cheng y Washington (2005) utilizan el método EB para la evaluación experimental de métodos de identificación de puntos críticos usando datos simulados derivados experimentalmente y concluyendo que los resultados ilustran que la técnica EB supera significativamente a técnicas de clasificación e intervalo de confianza, solo con ciertas salvedades. Elvik (2008) realiza un análisis comparativo de cinco técnicas para identificar ubicaciones de carreteras peligrosas en datos de carreteras noruegas y concluyendo que la técnica EB es la que funciona mejor según criterios de sensibilidad y especificidad. Wang *et al.* (2017) realizaron un examen de la fiabilidad de los factores de modificación de la colisión (CMFs) mediante el método empírico de Bayes con técnica de remuestreo para modelos de predicción de accidentes en una base de datos de 153 intersecciones, concluyendo así que los CMFs generados con la técnica EB tuvieron la mayor precisión. Lee *et al.* (2018) utilizan el método EB para la identificación de puntos críticos en las carreteras y predecir el número de choques, además proponen una mejora a este método mediante la incorporación de una medida de similitud basada en el índice de discordancia proporcional (PDR) para identificar segmentos viales similares. Islam *et al.* (2019) utilizan el método EB para mejorar la precisión en

modelos de predicción de tránsito con regresión Binomial Negativa para intersecciones de Dhaka en Bangladesh.

En resumen, se observa que el método EB es recomendado por variados autores y sirve como complemento a los modelos de predicción de accidentes de tránsito, con la ventaja de mejorar la precisión de la estimación de los siniestros de tránsito.

### **2.3. Variables dependientes (a explicar)**

En la literatura se observan principalmente dos alternativas de modelación de accidentes de tránsito. Estos son la estimación de la cantidad de accidentes de tránsito por entidad en cierto periodo, ya sea para la totalidad de los accidentes o por severidad, y la estimación de la severidad o gravedad que podría tener un accidente ocurrido en una entidad.

En la Tabla 2.1 y Tabla 2.2 se encuentran trabajos y estudios realizados previamente por distintos autores. En estas se observan los trabajos realizados por Guo *et al.*, (2010), Cunto *et al.*, (2015), Rizaldi *et al.*, (2017), Islam *et al.*, (2019), entre otros, quienes realizaron modelos de estimación del conteo o frecuencias de accidentes de tránsito. Por otro lado, los trabajos de Mafi *et al.*, (2018), Pooyan *et al.*, (2020), Ghandour *et al.*, (2020), entre otros, abordan modelos de predicción de la fatalidad o severidad del accidente de tránsito.

En los trabajos mencionados anteriormente se observa que para la predicción de conteo o frecuencia de accidentes de tránsito se utiliza principalmente modelos de regresión, los cuales se basan en un modelo matemático. En cambio, en la predicción de fatalidad o severidad se hace uso generalmente de herramientas de aprendizaje de máquinas (*machine learning*) que, sin una base matemática, el modelo se entrena y predice como resultado la gravedad del siniestro.

### **2.4. Variables independientes (explicativas)**

Entre las variables explicativas que son significativas en la frecuencia de accidentes de tránsito suelen encontrarse tres tipos, las cuales son infraestructura, operación y entorno. Como explica

Dufeu (2019), las variables de infraestructura hacen referencia al diseño geométrico y equipamiento vial y de seguridad presente en la intersección; las de operación caracterizan el funcionamiento de la intersección como el flujo vehicular, peatonal, giros izquierda y derecha, semaforización, entre otros; y finalmente las de entorno hacen referencia al uso de suelo en el que se emplaza la entidad y servicios presentes en los alrededores como paradas de autobús, centros de salud, centros educacionales, etc.

A continuación, en la Tabla 2.3 se resumen las variables explicativas de los tres tipos que han sido encontradas significativas en literatura previa. A partir de la Tabla 2.3 se destaca la variable flujo vehicular, de tipo operación, la cual es la variable que suele encontrarse más a menudo en la literatura, existiendo incluso modelos de predicción de frecuencia de accidentes de tránsito generados solo con dicha variable. Esto se debe al simple hecho de que para que ocurra un siniestro de tránsito debe haber presencia de vehículos en la vía, por lo tanto, mientras más flujo vehicular exista habrá más posibilidades de que dicho suceso ocurra.

**Tabla 2.3.** Variables significativas en literatura.

Tipo de variable	Variable	Referencias
Infraestructura	Nº de accesos a la intersección	SECTRA (2014); Huang <i>et al.</i> (2017)
	Nº de pistas por acceso	SECTRA (2014)
	Existencia de pistas de viraje	SECTRA (2014); Chen y Xie (2016)
	Existencia de mediana	SECTRA (2014); Alarifi <i>et al.</i> (2018)
	Tamaño de la intersección	Guo <i>et al.</i> (2010)
	Presencia de carril de bicicleta	Huang <i>et al.</i> (2017)
	Presencia estacionamiento	Islam <i>et al.</i> (2019)
Operación	Presencia de semáforo	SECTRA (2014); Huang <i>et al.</i> (2017)
	Flujo vehicular rama principal y rama secundaria	SECTRA (2014); Chen y Xie (2016); Guo <i>et al.</i> (2010); Huang <i>et al.</i> (2017); Islam <i>et al.</i> (2019); Alarifi <i>et al.</i> (2018)
	Diferencia de límite de velocidad entre rama principal y rama secundaria	Alarifi <i>et al.</i> (2018)
	Flujo de giro a la izquierda en rama principal y rama secundaria	Guo <i>et al.</i> (2010)
	Densidad peatonal	Huang <i>et al.</i> (2017)
	Rango etario	Huang <i>et al.</i> (2017)
	Límite de velocidad en rama principal	Huang <i>et al.</i> (2017)
	Descenso y abordaje de pasajeros	Islam <i>et al.</i> (2019)
Entorno	Zona céntrica o periférica	SECTRA (2014)
	Parada de autobús	Huang <i>et al.</i> (2017)
	Presencia de centros de salud	Islam <i>et al.</i> (2019)
	Presencia de establecimiento religioso	Islam <i>et al.</i> (2019)
	Zona urbana o rural	Islam <i>et al.</i> (2019)

## 2.5. Problemas metodológicos

Los modelos de predicción de accidentes de tránsito pueden presentar variados problemas metodológicos debido, generalmente, a las características de la base de datos, es decir, de la frecuencia de accidentes de tránsito y los atributos de las intersecciones involucradas. Lord y Mannering (2010) explican resumidamente los principales problemas metodológicos asociados a los datos y variables de accidentes de tránsito que se han encontrado en diversos estudios. Entre ellos se encuentra la sobre dispersión, baja dispersión, variables explicativas que varían en el tiempo, correlación temporal y espacial, media muestral baja y tamaño muestral pequeño, correlación entre la gravedad de la lesión y el tipo de accidente, infra registro, sesgo de variables omitidas, variables endógenas, forma funcional y parámetros fijos.

Estos problemas metodológicos serán explicados a continuación a partir de la información expuesta por Lord y Mannering (2010).

- Sobredispersión: se da cuando la varianza excede la media de los conteos de accidentes de tránsito. Cuando hay datos dispersos en exceso, la estimación de un modelo puede resultar en estimaciones de parámetros sesgadas e inconsistentes que, a su vez, podrían conducir a inferencias erróneas con respecto a los factores que determinan las frecuencias de los choques.
- Baja dispersión: ocurre cuando la media excede la varianza. Esto es poco común en datos de accidentes de tránsito, pero en caso de ocurrir puede generar una estimación de parámetros incorrecta.
- Variables explicativas variables en el tiempo: El hecho de que una variable explicativa cambie significativamente durante un periodo de tiempo generalmente no se considera en los modelos debido a la falta de datos detallados de esta variable en ese periodo de tiempo.
- Correlación temporal y espacial: Se da cuando un dato, por ejemplo, de un año, se divide en 12 observaciones, 1 por mes. Estas observaciones se correlacionarán con el tiempo porque muchos de los efectos no observados asociados con una entidad de carretera específica permanecerán iguales a lo largo del tiempo. De manera similar, puede existir

correlación sobre el espacio, porque las entidades de la calzada que estén muy próximas pueden compartir efectos no observados.

- Media muestral baja y tamaño muestral pequeño: Ocurre generalmente debido al costo que conlleva una recopilación de datos de choques y de las características de una intersección, sobre todo en cuanto al tiempo que se requiere para lograr una base de datos extensa.
- La gravedad de la lesión y correlación del tipo de accidente: Se da al momento de registrar la información y al clasificar una intersección por gravedad. Esta gravedad queda muy a criterio del ente que esté realizando esta clasificación de accidentes (generalmente Carabineros).
- Bajo registro: Este aspecto es muy usual. Ocurre generalmente ya que los choques leves no se reportan ante la autoridad policial y se llega a un acuerdo entre los mismos involucrados al momento de ocurrencia del choque para no tener problemas ante la ley.
- Sesgo de variables omitidas: Se da al omitir variables que podrían ser significativas pero que por falta de información no se consideró.
- Variables endógenas: Ocurre cuando la variable explicativa depende del número de choques. Por ejemplo, si se considerara una señalética de precaución por zona de hielo como variable independiente, la cual generalmente se emplaza en el lugar luego de registrarse choques debido al hielo.
- Forma funcional: La mayoría de los modelos asumen que las variables explicativas influyen en la variable dependiente de forma lineal, pero variados trabajos sugieren que una forma funcional no lineal funciona mejor.
- Parámetros fijos: Los modelos estadísticos tradicionales no permiten que las estimaciones de parámetros varíen en las observaciones. Puede existir el caso que en algunas zonas este varié en la realidad, pero los modelos sin parámetros fijos se vuelven bastante complejos.

En la Tabla 2.4 se resumen los problemas metodológicos asociados a accidentes de tránsito según Lord y Mannering (2010).

**Tabla 2.4.** Resumen de problemas metodológicos en datos de accidentes de tránsito. (Lord y Mannering, 2010).

<b>Dato o problema metodológico</b>	<b>Problemas asociados</b>
Sobre dispersión	Puede violar supuestos básicos de la modelación.
Baja dispersión	Puede violar supuestos básicos de la modelación.
Variables explicativas variables en el tiempo	Estimaciones de parámetros erróneas.
Correlación temporal y espacial	Pérdidas en la eficiencia de la estimación.
Media muestral baja y tamaño muestral pequeño	Causar errores en las estimaciones de los parámetros.
La gravedad de la lesión y la correlación del tipo de accidente	Pérdidas en la eficiencia de la estimación.
Bajo registro	Distorsiona las predicciones del modelo y dar lugar a inferencias erróneas con respecto a la influencia de las variables explicativas.
Sesgo de variables omitidas	Inferencias erróneas con respecto a la influencia de las variables explicativas.
Variables endógenas	Estimación de parámetros sesgadas e inferencias erróneas respecto a variables explicativas.
Forma funcional	Estimación de parámetros sesgadas e inferencias erróneas respecto a variables explicativas.
Parámetros fijos	Estimación de parámetros sesgadas e inferencias erróneas respecto a variables explicativas.

Finalmente, se han encontrado variadas metodologías en relación a los datos de frecuencia de accidentes de tránsito además de los problemas asociados a sus variables, los cuales pueden llevar a una predicción errónea. La existencia de una variedad de modelos y técnicas para la predicción se asocian a distintas ventajas y desventajas respecto a uno y otro, y los métodos de estimación de parámetros también juegan un rol importante. Todos estos aspectos, considerados en conjunto, definirán la aproximación que tenga la predicción estudiada.

# **CAPÍTULO 3**

## **METODOLOGÍA**

### **Capítulo 3. Metodología**

En el presente estudio se elaboró un modelo para la predicción de la frecuencia de accidentes de tránsito en intersecciones urbanas semaforizadas del “Gran Concepción”. Para esto, se realizó una agrupación de la base de datos de accidentes que separa las intersecciones y logra la desagregación de los datos, mediante el uso de la técnica de agrupación (*clusters*) *K-Means* (K-medios), para luego pasar a la estimación del modelo de predicción de accidentes mediante un modelo de regresión Binomial Negativa (Poisson-Gamma) en conjunto al método Empírico Bayesiano (EB). La base de datos utilizada contiene información del área metropolitana del “Gran Concepción” de la Región del Biobío en Chile, la cual incluye intersecciones de las comunas de Chiguayante, Concepción, Hualqui, San Pedro, Talcahuano y Tomé, con datos de un periodo de 5 años de registro de accidentes de tránsito. A continuación, se describen los procesos realizados en la elaboración de este trabajo, necesarios para la obtención de los modelos de predicción.

#### **3.1. Identificación modelos de predicción y variables explicativas**

En primer lugar, se realizó una revisión bibliográfica de los métodos, modelos, técnicas y algoritmos más utilizados en la predicción de accidentes de tránsito por distintos autores a lo largo de los últimos diez años. En el capítulo anterior de Antecedentes Generales del presente informe se hace referencia a destacados artículos y estudios en los cuales se encuentran distintos tipos de metodologías para la predicción de accidentes de tránsito con un enfoque en la frecuencia o en la severidad de estos. Estos métodos van desde modelos simples de regresión lineal, arboles de decisión y hasta aplicaciones de inteligencia artificial.

En segundo lugar, se realizó una revisión bibliográfica de las variables explicativas (cuantitativas y/o cualitativas) relacionadas a infraestructura, operación y entorno que han sido utilizadas en dichas metodologías, identificando cuales son las más significativas y que han tenido un mayor efecto en la frecuencia de accidentes y/o de la gravedad de estos.

### 3.2. Formulación del modelo

El modelo de predicción de accidentes de tránsito a implementar fue realizado a partir del trabajo en conjunto de dos metodologías, estas son el modelo de regresión Binomial Negativo (NB) y el método Empírico Bayesiano (EB). En primer lugar, se hace uso del modelo de regresión NB para la obtención de los accidentes de tránsito estimados en intersecciones y posteriormente se aplica el método EB para mejorar la precisión y obtener los accidentes de tránsito esperados. A continuación, se explicará la formulación de cada uno de ellos.

#### 3.2.1. Modelo de regresión Binomial Negativo

Como la cantidad de choques son discretos, no negativos y eventos aleatorios, se asume que estos siguen una distribución Poisson (Sawalha y Sayed, 2006), es decir:

$$Y_i | \theta_i \sim \text{Poisson}(\theta_i) \quad (3.1)$$

Donde,  $Y_i$  es el número de accidentes en la intersección  $i$  ( $i = 1, 2, 3, \dots, n$ ), se asume que la cantidad de accidentes en cada una de las  $n$  intersecciones es independiente.  $\theta_i$  es el parámetro de Poisson y  $\theta_i > 0$ . Luego, la probabilidad de que una intersección  $i$  tenga  $y_i$  accidentes esta dado por:

$$\text{Pr}\{Y_i = y_i | \theta_i\} = \frac{e^{-\theta_i} \cdot \theta_i^{y_i}}{y_i!} \quad (3.2)$$

Miaou y Lord (2003) mencionan que el parámetro de Poisson,  $\theta_i$ , se expresa comúnmente como una función exponencial de los atributos de la intersección, es decir:

$$\theta_i = \exp(X_i' \cdot \beta) \quad (3.3)$$

Donde,  $X_i'$  es un vector fila de los atributos que caracterizan a la intersección y  $\beta$  es un vector de los parámetros de regresión estimados a partir de los datos. En la los modelos Poisson se asume que la media y la varianza son iguales, es decir:

$$E(Y_i) = \text{Var}(Y_i) = \theta_i \quad (3.4)$$

Esta suposición suele verse infringida en las bases de datos de accidentes de tránsito debido a la sobredispersión que suele presentarse (Cameron y Trivedi, 1998). Esto puede resultar en malas estimaciones de los errores estándar de los coeficientes de la regresión, viéndose afectado su nivel de significancia.

Para sobrellevar dichos problemas relacionados a los modelos de regresión Poisson los autores proponen el uso de la Poisson-Gamma que conduce al modelo de regresión Binomial Negativo (Cameron y Trivedi, 1998). Para abordar esta sobredispersión, se asume que:

$$\theta_i = \mu_i \cdot \exp(u_i) \quad (3.5)$$

$$\ln(\theta_i) = \ln(\mu_i) + u_i \quad (3.6)$$

Donde,  $\mu_i$  es una función exponencial de atributos de la intersección expresada como:

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^J (\beta_j \cdot X_{ji}) \quad (3.7)$$

Donde,  $X_{ji}$  es una matriz de variables,  $j$  es el número de variables ( $j = 1,2,3, \dots, J$ ),  $\beta_0$  es el intercepto, y  $\beta_j$  es el vector de los coeficientes de la regresión. El termino  $\exp(u_i)$  sigue una distribución gamma con un parámetro de dispersión inversa  $k$  (Islam *et al.*, 2019), es decir:

$$\exp(u_i) | k \sim \text{Gamma}(k, k) \quad (3.8)$$

El parámetro de dispersión,  $\alpha$ , se define como:

$$\alpha = \frac{1}{k} \quad (3.9)$$

La función de densidad de probabilidad del modelo NB esta dado por:

$$\Pr\{Y_i = y_i | \mu_i, k\} = \frac{\Gamma(y_i+k)}{y_i! \Gamma(k)} \cdot \left(\frac{k}{k+\mu_i}\right)^k \cdot \left(\frac{\mu_i}{k+\mu_i}\right)^{y_i} \quad (3.10)$$

Por lo tanto, bajo un modelo NB, la media y la varianza se expresarán como:

$$E(Y_i) = \mu_i \quad (3.11)$$

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{k} \quad (3.12)$$

La aplicación del modelo de regresión Binomial Negativa sobre la base de datos y la estimación de parámetros se realizaron en un script o código elaborado por el alumno en el software R, haciendo uso de los paquetes MASS, stats y Metrics.

### 3.2.2. Método Empírico Bayesiano (EB)

El método EB, como se mencionó en el capítulo de Antecedentes Generales, consiste en hacer uso conjunto de dos aspectos de seguridad de una entidad: el historial de accidentes de la entidad y la frecuencia de accidentes estimada por un modelo de predicción para entidades similares (Hauer *et al.*, 2002)

En la estimación EB, el uso conjunto de ambos aspectos se implementa mediante un promedio ponderado, por lo tanto, los accidentes esperados se calculan como:

$$EB_i = w_i \cdot \mu_i + (1 - w_i) \cdot y_i \quad (3.13)$$

El término  $w_i$  es el peso que se le da a los accidentes estimados y se determina por:

$$w_i = \frac{k}{k + \mu_i} \quad (3.14)$$

Donde,  $EB_i$  son los accidentes esperados en la intersección  $i$ ,  $w_i$  es el peso que se le da a los accidentes estimados,  $\mu_i$  son los accidentes estimados,  $y_i$  son los accidentes históricos y  $k$  es el parámetro de sobredispersión entregado por el modelo de predicción NB.

La aplicación del método EB se realiza en un script o código elaborado por el alumno en el software R, donde para cada intersección con registro histórico de accidentes y accidentes estimados por el modelo NB se procede al cálculo de sus accidentes esperados. Este cálculo es fácil de hacer mediante el uso de las ecuaciones 3.13 y 3.14.

### 3.3. Selección de variables explicativas

La selección de variables explicativas (atributos de la intersección), ya sean de tipo infraestructura, operación o entorno, es un proceso iterativo, en el cual a priori no se sabe que variables pueden ser las más significativas, congruentes y capaces de lograr la explicación del fenómeno de accidente de tránsito. Por lo tanto, las variables con las que se trabajara en primera instancia serán las que a partir de la literatura se ha llegado a la conclusión de que son influyentes en la generación de accidentes de tránsito en intersecciones. Dichas variables serán mencionadas más adelante.

Las variables explicativas presentes en los modelos finales se seleccionaron mediante procesos iterativos de adición, eliminación e intercambios paso a paso de estas variables. Las variables que se mantienen en el modelo deben cumplir con el criterio estadístico de que el valor Z (parámetro estimado dividido por su error estándar) tenga significancia a un 95% de nivel de confianza para los grados de libertad que correspondan (Islam *et al.*, 2019), es decir:

$$\text{valor } p \leq 0.05 \quad (3.15)$$

### 3.4. Evaluación del modelo

Para la evaluación de los modelos, existen dos tipos de medidas o parámetros. El primero es para validar la bondad de ajuste de un modelo y el segundo para realizar comparaciones y seleccionar entre diferentes modelos, el mejor. A continuación, se listan algunos de estos parámetros explicados en los trabajos de Camarillo *et al.* (2013), Alcaide (2015) e Islam *et al.* (2019).

- Estadístico Pearson  $\chi^2$ : Este estadístico indica que la hipótesis  $H_0$  no se rechaza en el caso que  $\chi^2 < \chi^2_{crítico}$ .  $\chi^2_{crítico}$  se determina para un intervalo de confianza de 95% con grados de libertad  $gl = n - p$ . Este estadístico indica la bondad de ajuste del modelo. Para el modelo NB, se determina como:

$$Pearson \chi^2 = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\mu_i + \frac{\mu_i^2}{k}} \quad (3.16)$$

Donde,  $n$  es el tamaño de la muestra,  $p$  es la cantidad de parámetros estimados,  $y_i$  son los accidentes observados,  $\mu_i$  son los accidentes estimados y  $k$  el parámetro de sobredispersión.

- Coeficiente de determinación ( $R^2$ ): Este coeficiente es la varianza explicada por el modelo ajustado. Es el cuadrado del coeficiente de correlación entre el valor de accidentes estimado y el observado. Se determina como:

$$R^2 = \frac{Cov(y, \mu)^2}{Var(y) \cdot Var(\mu)} \quad (3.17)$$

Donde,  $Cov(y, \mu)$  es la covarianza entre los accidentes observados y estimados,  $Var(y)$  y  $Var(\mu)$  son las varianzas respectivas de cada uno.

- Coeficiente de determinación corregido ( $\bar{R}^2$ ): Cumple la misma función que el coeficiente de determinación, pero considerando el número de variables incluidas en el modelo. Se utiliza para comparar entre modelos alternativos. Se define como:

$$\bar{R}^2 = 1 - \frac{n-1}{n-(p-1)-1} \cdot (1 - R^2) \quad (3.18)$$

Donde,  $n$  es el tamaño de la muestra y  $(p - 1)$  son las variables incluidas en el modelo.

- Error medio absoluto (*MAE*): Mide la magnitud promedio del error entre los accidentes observados y los estimados por el modelo. Se utiliza para comparar entre modelos alternativos. Se calcula como:

$$MAE = (1/n) \cdot \sum_{i=1}^n |y_i - \mu_i| \quad (3.19)$$

Donde,  $y_i$  son los accidentes observados,  $\mu_i$  son los accidentes estimados y  $n$  el tamaño de la muestra.

- Raíz del error cuadrático medio (*RMSE*): Mide el valor medio cuadrático del error y pondera los pronósticos que están más alejados del valor medido. Se utiliza para comparar entre modelos alternativos. Se determina por:

$$RMSE = \sqrt{(1/n) \cdot \sum_{i=1}^n (y_i - \mu_i)^2} \quad (3.20)$$

Donde,  $y_i$  son los accidentes observados,  $\mu_i$  son los accidentes estimados y  $n$  el tamaño de la muestra.

- Criterio de información Akaike (*AIC*): Este criterio recompensa la bondad de ajuste e incluye una penalidad en función del número de parámetro estimados. Se utiliza para comparar entre modelos alternativos, donde el mejor modelo es el de menor AIC. Se define como:

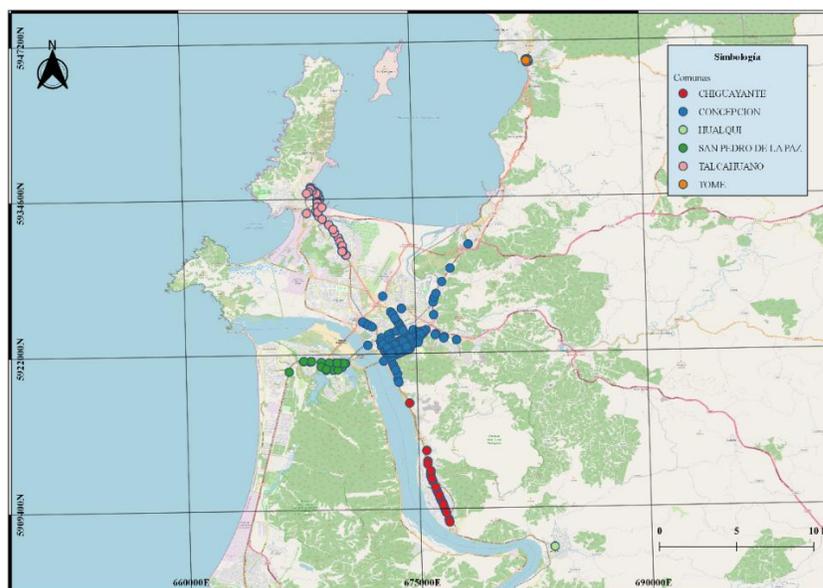
$$AIC = p - 2 \cdot \ln(\hat{L}) \quad (3.21)$$

Donde,  $p$  es el número de parámetro estimados y  $\hat{L}$  es el máximo valor de la función de verosimilitud.

### 3.5. Zona de estudio

La zona de estudio, considerada en la base de datos del presente trabajo de investigación, se encuentra en la Provincia de Concepción de la Región del Biobío de Chile. Las intersecciones contempladas forman parte de seis comunas de esa provincia, Chiguayante, Concepción, Hualqui, San Pedro de La Paz, Talcahuano y Tomé. En la Figura 3.1 se presenta el área de estudio junto con la ubicación de cada una de las intersecciones simbolizadas por comuna. Para ver detalle por comuna revisar **Anexo A**.

Según información del sitio web de la Gobernación de la Provincia de Concepción (2021), esta provincia es la que alberga la mayoría de la población regional, así como a todas las comunas que conforman el Gran Concepción. Además, es una de las zonas industriales más importantes de Chile, comprendiendo empresas de las áreas de siderúrgica, alimentaria, refinación de petróleo, petroquímica, pesquera, plantas de celulosa, aserraderos, entre otros (Gobernación Provincia de Concepción, 2021). Beyer (2021) menciona que el Gran Concepción posee poca infraestructura con capacidad de cruzar la metrópolis en dirección norte-sur y este-oeste, que permita viajar desde Talcahuano a Chiguayante y desde Penco a San Pedro de La Paz. Además, el crecimiento urbano, aglomeración de comercio y la demanda por el uso de espacios, provocan problemas de congestión en las grandes ciudades, como Concepción. (Beyer, 2021).



**Figura 3.1.** Mapa emplazamiento de intersecciones por comuna presentes en la base de datos.

Una intersección se define como un área formada por la conexión a nivel de todas las ramas de acceso hacia esa conexión y el área colindante a ella (Islam *et al.*, 2019). Las intersecciones observadas en la Figura 3.1 y que son consideradas en este estudio suman un total de 253 intersecciones (ver Tabla 3.1), todas reguladas por semaforización. En cuanto a la geometría de las intersecciones se encuentran de tipo T y tipo X, con ángulos entre las ramas principal y secundaria de 30° a 90°. El flujo total de estas intersecciones, tránsito medio diario (TMD), van desde los 6000 a los 76000 vehículos aproximadamente. Dichas intersecciones se emplazan en zonas con usos de suelo residencial, comercial o mixto (residencial y comercial).

En la Tabla 3.1 se observa el detalle de la cantidad de intersecciones por comuna consideradas en este trabajo de investigación y contempladas en la base de datos.

**Tabla 3.1.** Detalle de cantidad de intersecciones por comuna presentes en la base de datos.

Comuna	Cantidad de intersecciones
Chiguayante	22
Concepción	169
Hualqui	1
San Pedro	17
Talcahuano	38
Tomé	6
Total	253

### 3.6. Base de datos de accidentes

Una vez identificadas las variables explicativas más significativas en la literatura existente, estas son seleccionadas a partir de la base de datos de las intersecciones del “Gran Concepción” facilitado por el profesor Dr. Tomás Echaveguren de la Universidad de Concepción y recopilados en la tesis magister del Sr. René Dufeu en el programa de Magister de Ingeniería Civil de la Universidad de Concepción. Dicha base de datos del “Gran Concepción” incluye intersecciones de las comunas de Chiguayante, Concepción, Hualqui, San Pedro, Talcahuano y Tomé con un periodo de 5 años (2009 a 2013) de información de accidentes de tránsito y atributos (variables explicativas) de las entidades, es decir las intersecciones.

A continuación, se mostrarán las variables utilizadas en este trabajo, dependientes (a explicar) e independientes (explicativas) que fueron seleccionadas a partir de los antecedentes recopilados de literatura estudiada previamente. Luego se presentará el análisis descriptivo de la muestra considerando dichas variables.

### 3.6.1. Variables dependientes

Las variables dependientes, que fueron estimadas mediante el modelo de predicción planteado, son los accidentes de tránsito en intersecciones urbanas. Estas variables consideran datos de accidente de tránsito ocurridos para un periodo de 5 años, del 2009 al 2013. Estos datos provienen del registro e información recopilada por Carabineros de Chile, quienes son los que van al lugar al momento del siniestro y registran los antecedentes del hecho, entre ellos la gravedad del accidente, quedando a criterio de dicha entidad la gravedad en que clasifica cada siniestro. A partir de ello, se realiza la separación en accidentes totales, severos y leves, en un periodo de 5 años para cada entidad correspondiente (intersección). Cabe mencionar que a partir de la base de datos del “Gran Concepción” se observa que, respecto a los accidentes totales, los accidentes severos representan un 16.8% y los accidentes leves un 83.2%.

En la Tabla 3.2 se presenta el detalle de cada una de las tres variables a modelar.

**Tabla 3.2.** Variables dependientes a estimar en los modelos.

<b>Variable</b>	<b>Valor</b>	<b>Descripción</b>
Accidentes Totales	Continuo	Accidentes totales (todas las gravedades) en un periodo de 5 años.
Accidentes Severos	Continuo	Accidentes fatales, graves y medianamente graves en un periodo de 5 años.
Accidentes Leves	Continuo	Accidentes leves y sin daños en un periodo de 5 años.

En la Tabla 3.3 se encuentra el análisis descriptivo de las variables dependientes mencionadas. Para análisis descriptivo se utilizó los softwares JASP y Microsoft Excel.

**Tabla 3.3.** Análisis descriptivo de variables dependientes.

Variable	Media	Desviación Estándar	Varianza de la muestra	Valor mínimo	Valor máximo
Accidentes Totales	11.64	8.99	80.84	0	44
Accidentes Severos	1.96	1.99	3.95	0	9
Accidentes Leves	9.68	7.65	58.49	0	36

A partir de la Tabla 3.3 se observa, en las tres variables, la existencia de sobredispersión en los datos (varianza mayor a la media), acentuándose más en Accidentes Totales y Accidentes Leves. Por lo cual se justifica la elección del modelo Binomial Negativo.

### 3.6.2. Variables independientes

En el capítulo 2 Antecedentes Generales se revisaron las variables más significativas de literatura y que han sido utilizadas en diversos trabajos para explicar el fenómeno de accidentes de tránsito. A partir de esa revisión bibliográfica y considerando las variables disponibles en la base de datos del “Gran Concepción”, se han seleccionado las variables con las que se trabajara y se realizaran las iteraciones de cada uno de los casos a modelar.

Como explica Dufeu (2019), las variables de infraestructura hacen referencia al diseño geométrico y equipamiento vial y de seguridad presente en la intersección; las de operación caracterizan el funcionamiento de la intersección como el flujo vehicular, peatonal, giros izquierda y derecha, semaforización, entre otros; y finalmente las de entorno hacen referencia al uso de suelo en el que se emplaza la entidad y servicios presentes en los alrededores como paradas de autobús, centros de salud, centros educacionales, etc.

A continuación, en la Tabla 3.4, Tabla 3.6 y Tabla 3.8 se presentan las variables explicativas de infraestructura, operación y entorno, respectivamente, además de los valores que toman y la descripción de cada una. En la Tabla 3.5, Tabla 3.7 y Tabla 3.9 se presenta el análisis descriptivo de cada una de las variables explicativas de infraestructura, operación y entorno, respectivamente.

**Tabla 3.4.** Variables explicativas de infraestructura seleccionadas.

Variable	Valor	Descripción
PISTA EXCLUSIVA GIRO DERECHA RP	0	No existe pista exclusiva de giro a derecha en rama principal
	1	Existe al menos una pista exclusiva de giro a derecha en rama principal
PISTA EXCLUSIVA GIRO DERECHA RS	0	No existe pista exclusiva de giro a derecha en rama secundaria
	1	Existe al menos una pista exclusiva de giro a derecha en rama secundaria
PISTA EXCLUSIVA GIRO IZQUIERDA RP	0	No existe pista exclusiva de giro a izquierda en rama principal
	1	Existe al menos una pista exclusiva de giro a izquierda en rama principal
PISTA EXCLUSIVA GIRO IZQUIERDA RS	0	No existe pista exclusiva de giro a izquierda en rama secundaria
	1	Existe al menos una pista exclusiva de giro a izquierda en rama secundaria
Nº RAMAS ACCESO	1 - 4	Número de ramas de acceso a intersección
Nº PISTAS RP	2 - 8	Número total de pistas de la rama principal en ambos sentidos
Nº PISTAS RS	0 - 6	Número total de pistas de la rama secundaria en ambos sentidos
PRESENCIA MEDIANA RP	0	No existe mediana en la rama principal o es menor a 0.8 [m] de ancho
	1	Existe mediana en la rama principal y tiene al menos 0.8 [m] de ancho
PRESENCIA MEDIANA RS	0	No existe mediana en la rama secundaria o es menor a 0.8 [m] de ancho
	1	Existe mediana en la rama secundaria y tiene al menos 0.8 [m] de ancho

**Tabla 3.5.** Análisis descriptivo de variables explicativas de infraestructura seleccionadas.

Variable	Media	Desviación Estándar	Varianza de la muestra	Valor mínimo	Valor máximo
PISTA EXCLUSIVA GIRO DERECHA RP	0.08	0.28	0.08	0	1
PISTA EXCLUSIVA GIRO DERECHA RS	0.12	0.33	0.11	0	1
PISTA EXCLUSIVA GIRO IZQUIERDA RP	0.25	0.43	0.19	0	1
PISTA EXCLUSIVA GIRO IZQUIERDA RS	0.13	0.33	0.11	0	1
Nº RAMAS ACCESO	2.74	0.69	0.47	1	4
Nº PISTAS RP	4.29	1.76	3.11	2	8
Nº PISTAS RS	2.27	0.66	0.44	1	5
PRESENCIA MEDIANA RP	0.54	0.50	0.25	0	1
PRESENCIA MEDIANA RS	0.09	0.29	0.09	0	1

**Tabla 3.6.** Variables explicativas de operación seleccionadas.

Variable	Valor	Descripción
GIRO A LA IZQUIERDA RP	0	No existen giros a la izquierda en la rama principal
	1	Existe al menos un giro a la izquierda en la rama principal
GIRO A LA IZQUIERDA RS	0	No existen giros a la izquierda en la rama secundaria
	1	Existe al menos un giro a la izquierda en la rama secundaria
DENSIDAD PEATONAL	1	Densidad peatonal baja en la intersección
	2	Densidad peatonal media en la intersección
	3	Densidad peatonal alta en la intersección
TMD INTER	Continuo	Flujo total en la intersección completa [veq/día]
TMD RP	Continuo	Flujo total de la rama principal completa [veq/día]
TMD RS	Continuo	Flujo total de la rama secundaria completa [veq/día]
TMD GI	Continuo	Flujo total que gira a izquierda en la intersección completa [veq/día]
TMD GD	Continuo	Flujo total que gira a derecha en la intersección completa [veq/día]

**Tabla 3.7.** Análisis descriptivo de variables explicativas de operación seleccionadas.

Variable	Media	Desviación Estándar	Varianza de la muestra	Valor mínimo	Valor máximo
GIRO A LA IZQUIERDA RP	0.50	0.50	0.25	0	1
GIRO A LA IZQUIERDA RS	0.74	0.44	0.19	0	1
DENSIDAD PEATONAL	1.67	0.75	0.56	1	3
TMD INTER	31853.53	15576.36	242622876.85	6249	75918
TMD RP	26207.77	15213.05	231436779.75	4014	69342
TMD RS	5580.60	4620.52	21349169.62	0	32489
TMD GI	2661.60	2831.04	8014805.42	0	17977
TMD GD	2757.96	2622.49	6877433.26	0	14276

**Tabla 3.8.** Variables explicativas de entorno seleccionadas.

Variable	Valor	Descripción
ZONA COM/MIX	0	Zona principalmente del tipo residencial
	1	Zona principalmente del tipo comercial o mixta (residencial y comercial)
N° PARADEROS	0 - 3	Numero de paraderos en radio de 50 [m] en torno al centro de la intersección
N° CENTROS DE SALUD	0 - 1	Número de centros de salud en radio de 130 [m] en torno al centro de la intersección

**Tabla 3.9.** Análisis descriptivo de variables explicativas de entorno seleccionadas.

Variable	Media	Desviación Estándar	Varianza de la muestra	Valor mínimo	Valor máximo
ZONA COM/MIX	0.63	0.48	0.23	0	1
N° PARADEROS	1.06	0.89	0.79	0	3
N° CENTROS DE SALUD	0.05	0.21	0.05	0	1

### 3.7. Agrupación de intersecciones

Se realizó una agrupación de intersecciones, considerando algunas de las variables seleccionadas anteriormente en relación a infraestructura, operación y entorno de la intersección, con el objetivo de agruparlas e identificar si se logra un mejor ajuste de los modelos, debido a que la experiencia internacional indica que es recomendable realizar desagregaciones de los datos de accidentes, ya que esto permite lidiar con la correlación de las observaciones (SECTRA, 2014). Además, como menciona Hauer *et al.* (2002), el método EB es aplicable para entidades similares, lo cual se espera lograr mediante la agrupación. Para lograr esto se realizó una agrupación de intersecciones mediante el método de machine learning, K-means, en el software JASP 0.14.1.0 (libre acceso), para posteriormente volver a estimar los modelos de predicción.

El algoritmo de agrupación (*clustering*) *K-means*, como explica Wagstaff *et al.* (2001), es un método del *machine learning* comúnmente utilizado para dividir automáticamente un conjunto de datos en *k* grupos (*clusters*). El algoritmo, internamente, procede seleccionando *k* centros de conglomerados iniciales y luego se refinan iterativamente.

Respecto al número de *clusters*, *K*, se decide la utilización de 2 *clusters* a modo de facilitar la posterior modelación e iteración de los resultados considerando que para cada *cluster* y el set completo de intersecciones (sin agrupar) se modelaron tres casos (accidentes totales, severos y leves), es decir, 9 casos a modelar.

En cuanto a las variables utilizadas para la agrupación, se probó la agrupación con dos sets de variables. El primero contemplaba cinco variables pertenecientes a las tres categorías (infraestructura, operación y entorno) y el segundo las mismas cinco variables más las variables

dependientes de accidentes totales, severos y leves. Posteriormente en los resultados se observó que los modelos de predicción de accidentes se ajustaban de mejor manera con la primera opción, es decir sin considerar la variable dependiente, por lo tanto, solo se mantuvo la agrupación obtenida de dicho set de variables. Esto se debió a que al realizar la agrupación con la variable dependiente de accidentes se generaban *clusters* de intersecciones con cantidad de accidentes similares, es decir, se perdía la sobredispersión en los datos, el cual fue el primer supuesto para utilizar el modelo NB, entonces se vio afectada la bondad de ajuste. En la Tabla 3.10 se encuentran las variables utilizadas y su descripción.

**Tabla 3.10.** Variables explicativas seleccionadas para agrupación.

Variable	Valor	Descripción
TMD RP	Continuo	Flujo total de la rama principal completa [veq/día]
TMD RS	Continuo	Flujo total de la rama secundaria completa [veq/día]
ANGULO INTERSECCIÓN	0° - 90°	Ángulo de la intersección formado entre ambos ejes de las calles que conforman la intersección
TIPO INTER	0	Intersección en T
	1	Intersección en X
ZONA COM/MIX	0	Zona principalmente del tipo residencial
	1	Zona principalmente del tipo comercial o mixta (residencial y comercial)

El algoritmo K-means se ajusta a un modelo de agrupación, y en la Tabla 3.11 se muestran los valores de ajuste para el modelo con K igual a dos *clusters*. Cabe mencionar que, a mayor K o mayor número de *clusters*, los ajustes logrados son mejores, pero el proceso posterior de modelación de accidentes se vuelve más tedioso debido a la cantidad de casos. En los parámetros de entrenamiento de JASP para K-means se utilizó el algoritmo Hartigan-Wong con un máximo de 25 iteraciones, los cuales vienen predeterminados en el software.

**Tabla 3.11.** Ajuste de modelo de agrupación K-means en software JASP.

Clusters	N	R <sup>2</sup>	AIC	BIC	Silueta
2	253	0.275	933.00	968.33	0.33

En la Tabla 3.12 se muestran la cantidad de intersecciones (*size*) en ambos *clusters*, la variabilidad dentro de cada *cluster* en términos de la suma interna de cuadrados y los centroides. Los centroides indican el peso de las variables utilizadas en cada *cluster* al centrar y escalar (normalizar) las variables. Se observa que al dividir la suma interna de cuadrados en la suma total de cuadrados ( $347/1260=0.275$ ) se obtiene el valor del coeficiente de determinación  $R^2$  mostrado anteriormente. En definitiva, el algoritmo *K-means* entrega 2 *clusters* representativos de la muestra considerando las variables utilizadas. El Cluster 1 tendrá un total de 85 intersecciones, mientras que el Cluster 2 tendrá 168 intersecciones.

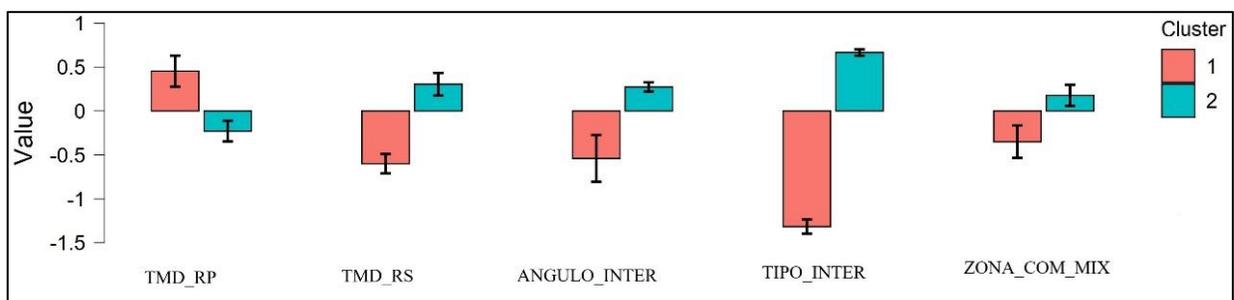
**Tabla 3.12.** Información de agrupación K-means en software JASP.

Clusters	1	2
Tamaño (cantidad de intersecciones)	85	168
Proporción explicada por la heterogeneidad del cluster	0.447	0.553
Dentro de la suma de cuadrados	407.904	505.094
Puntuación de silueta	0.227	0.388
Centroide TMD_RP	0.453	-0.229
Centroide TMD_RS	-0.600	0.304
Centroide ANGULO_INTERSECCION	-0.539	0.273
Centroide TIPO_INTER	-1.316	0.666
Centroide ZONA_COM_MIX	-0.350	0.177

*Nota.* La suma interna de cuadrados del modelo de *clusters* es 347.

*Nota.* La suma total de cuadrados del modelo de *clusters* es 1260.

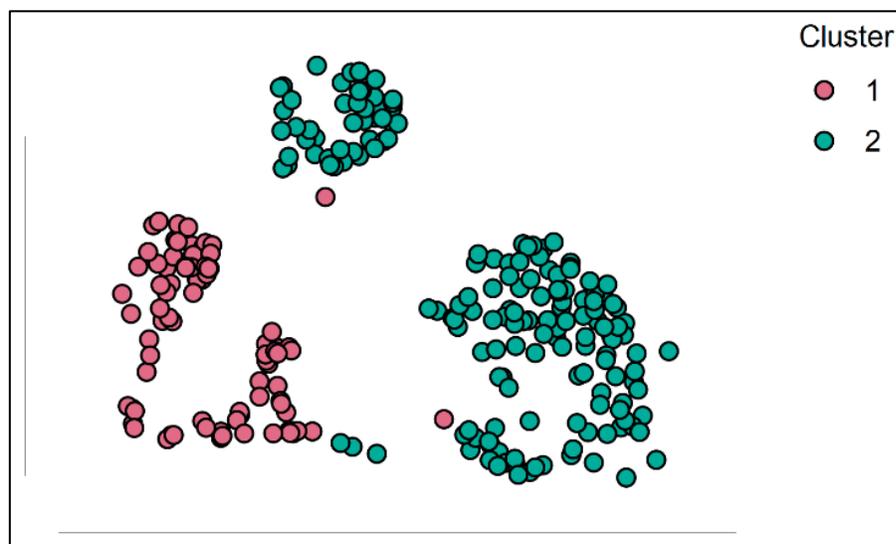
En la Figura 3.2 se muestran, de forma gráfica, los centroides de todos los parámetros predictores en cada uno de los *clusters*.



**Figura 3.2.** Medias (centroides) de predictores por *cluster* en software JASP.

De la Figura 3.2 se observan los centroides de las variables que fueron consideradas en el algoritmo de agrupación K-Medios. Los valores presentes en el gráfico representan el peso de cada variable correspondiente para cada uno de los *cluster* que resulta de escalar y centralizar las variables (normalizar), donde un valor negativo hace referencia a una media baja, mientras que un positivo indica una media mayor. Se observa que la variable de tipo de intersección es la que presenta mayor peso, donde el Cluster 1 presenta una media bastante baja en comparación al Cluster 2, lo cual se refleja en la información de la Tabla 3.13, donde se observa que el Cluster 1 tiene una media de 0.05, mientras que el Cluster 2 tiene una media de 0.98, es decir, mayoritariamente son intersecciones de tipo T y X, respectivamente.

Para tener una idea de cómo se agrupan los datos, JASP utiliza el algoritmo de incrustación de vecinos estocásticos con distribución t, cuyo resultado gráfico se observa en la Figura 3.3. En este gráfico se da una intuición de cómo se organizan los datos en un espacio de características de alta dimensión, proyectándose de forma no lineal puntos de datos similares en el espacio a puntos cercanos entre si en el plano. (JASP Team, 2020).



**Figura 3.3.** *Clusters* en el espacio de características, t-SNE en software JASP.

En la Tabla 3.13 se presenta el análisis descriptivo de algunas variables de interés. De esta se observa que en cuanto a la media de cantidad de accidentes totales, severos y leves ambos *clusters* presentan valores similares. Respecto a flujos, el Cluster 1 presenta ligeramente una media mayor

de flujo total en la intersección y en la rama principal respecto al Cluster 2, mientras que en la rama secundaria el Cluster 2 presenta media mayor de flujos que el Cluster 1. También se destaca que el Cluster 1 presenta mayor diferencia de magnitud de flujos en sus ramas (principal y secundaria) que el Cluster 2, en el cual esta diferencia de magnitud no es tan grande. Respecto al tipo de intersección, se observa que el Cluster 1 está conformado principalmente por intersecciones tipo T y el Cluster 2 por intersecciones tipo X. El tipo de zona del Cluster 1 es principalmente residencial, mientras que el Cluster 2 es tipo comercial/mixto.

**Tabla 3.13.** Análisis descriptivo de variables de interés por *cluster*.

Variable	Cluster	Media	Desviación Estándar	Varianza de la muestra	Valor mínimo	Valor máximo
ACC_TOTALES	1	11.44	10.03	100.53	0	44
	2	11.74	8.45	71.39	0	42
ACC_SEVEROS	1	1.81	1.80	3.23	0	8
	2	2.03	2.08	4.33	0	9
ACC_LEVES	1	9.62	8.80	77.48	0	36
	2	9.71	7.02	49.29	0	33
TMD_INTER	1	36036.17	14915.75	2.23e+8	9838	70050
	2	29737.32	15517.67	2.29e+8	6249	75918
TMD_RP	1	33105.17	15133.08	2.29e+8	8461	69342
	2	22718.02	14056.23	1.98e+8	4014	59596
TMD_RS	1	2806.12	2851.37	8.13e+6	0	11782
	2	6984.35	4714.41	2.22e+7	0	32489
TIPO_INTER	1	0.05	0.21	0.05	0	1
	2	0.98	0.13	0.02	0	1
ZONA_COM_MIX	1	0.46	0.50	0.25	0	1
	2	0.71	0.45	0.21	0	1

Finalmente, en la Figura 3.4 se observa un mapa con la distribución espacial de las intersecciones en la zona de estudio simbolizadas por *clusters* (Para ver el detalle por comuna ir a **Anexo C**). En ella se logra ver la presencia de ambos *clusters* en la mayoría de las comunas en estudio, solo exceptuando Hualqui y Tomé donde se encuentran intersecciones solo del Cluster 2. Además, se observa que mayoritariamente las intersecciones del Cluster 2 se encuentran en zonas céntricas, debido a que son principalmente intersecciones tipo X que suelen encontrarse en las zonas céntricas

de las comunas. En cambio, las intersecciones del Cluster 1 se encuentran mayoritariamente periféricas, debido a que son principalmente intersecciones tipo T.



**Figura 3.4.** Mapa emplazamiento de intersecciones por *cluster* en zona de estudio.

# **CAPÍTULO 4**

## **RESULTADOS Y DISCUSIÓN**

## **Capítulo 4. Resultados y discusión**

En este capítulo se presentan los resultados de los modelos de predicción de frecuencia (conteo) de accidentes de tránsito en intersecciones semaforizadas respecto a accidentes totales, severos y leves en un periodo de 5 años a partir de variables explicativas de infraestructura, operación y entorno. Dicha estimación se realizó para la base completa de intersecciones y se contrasta con la subdivisión en dos *clusters* de intersecciones obtenidos de un proceso de agrupación con K-medios. Dichos modelos provienen de una regresión Binomial Negativa (NB) en complemento con la metodología Empírico Bayesiana (EB) que calcula los accidentes esperados mediante el uso conjunto de los accidentes históricos de las entidades y los accidentes estimados por el modelo NB para una entidad similar.

A continuación, se presentan los resultados, análisis y discusión para los modelos de cada fenómeno de accidentes, es decir, accidentes totales, severos y leves para la base completa, Cluster 1 y Cluster 2. Base completa se refiere a todo el set de intersecciones juntos sin ningún tipo de agrupación, mientras que el Cluster 1 y Cluster 2 son la agrupación obtenida de la aplicación de K-medios. Accidentes estimados son los obtenidos del modelo de predicción, mientras que los accidentes esperados se obtienen al aplicar el método EB sobre el modelo de predicción.

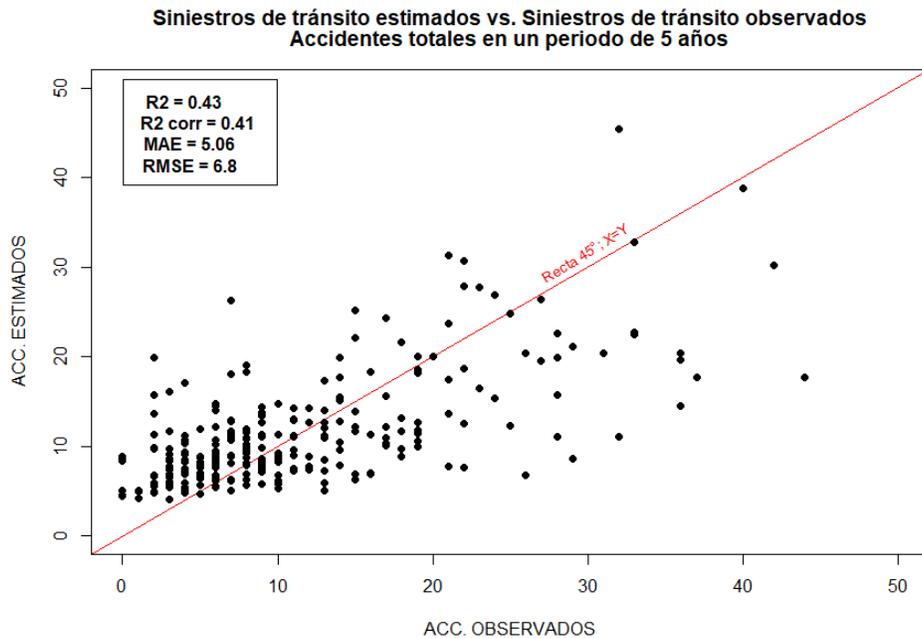
### **4.1. Resultados respecto a accidentes totales**

En la Tabla 4.1, Tabla 4.2 y Tabla 4.3 se presentan los resultados de los modelos respecto a accidentes totales de base completa, Cluster 1 y Cluster 2, respectivamente. En ellas se encuentran la estimación de parámetros, errores estándar, valores p y medidas de bondad de ajuste de los modelos.

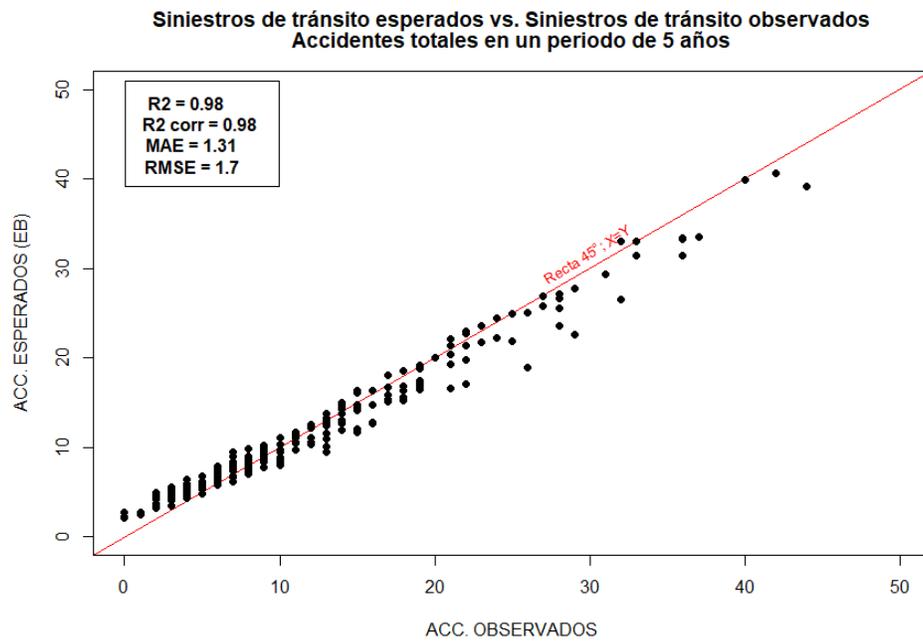
En la Figura 4.1 a Figura 4.6 se presentan los gráficos de los modelos NB y modelos NB con metodología EB respecto a accidentes totales de base completa, Cluster 1 y Cluster 2. En ellas se encuentran los coeficientes de determinación normal y corregido, y las medidas de error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE) obtenidos entre los accidentes estimados con los accidentes observados, y los accidentes esperados con los accidentes observados.

**Tabla 4.1.** Modelo de predicción de accidentes totales en base completa.

Modelo de predicción de accidentes – Accidentes Totales – Intersecciones Base completa							
$\ln(\mu) = 0.8437 + 0.0910 \cdot N\_PISTAS\_RP + 0.1297 \cdot N\_PISTAS\_RS - 0.3392 \cdot PRESENCIA\_MEDIANA\_RP + 2.3993 \cdot 10^{-5} \cdot TMD\_RP + 2.4794 \cdot 10^{-5} \cdot TMD\_RS + 0.1982 \cdot ZONA\_COM\_MIX + 0.0972 \cdot N\_PARADEROS$							
<b>Grados libertad (gl)</b>	<b>Pearson <math>\chi^2</math></b>	<b><math>\chi^2_{0.05,245}</math></b>	<b>Parámetro <math>k</math></b>	<b>Parámetro <math>\alpha</math></b>	<b>Desviación (D)</b>	<b>D/gl</b>	<b>AIC</b>
245	263	283	3.96	0.25	276	1.13	1614
Variable		Estimación	Error Est.	Valor Z	Pr (>  z )		
Intercepto		0.8437	0.1784	4.7304	2.24E-06	**	
N_PISTAS_RP		0.0910	0.0432	2.1048	0.0353	**	
N_PISTAS_RS		0.1297	0.0634	2.0466	0.0407	**	
PRESENCIA_MEDIANA_RP		-0.3392	0.1336	-2.5391	0.0111	**	
TMD_RP		2.3993E-5	3.66E-6	6.5628	5.28E-11	**	
TMD_RS		2.4794E-5	9.31E-6	2.6645	0.0077	**	
ZONA_COM_MIX		0.1982	0.0830	2.3888	0.0169	**	
N_PARADEROS		0.0972	0.0462	2.1053	0.0353	**	
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘.’ 1							



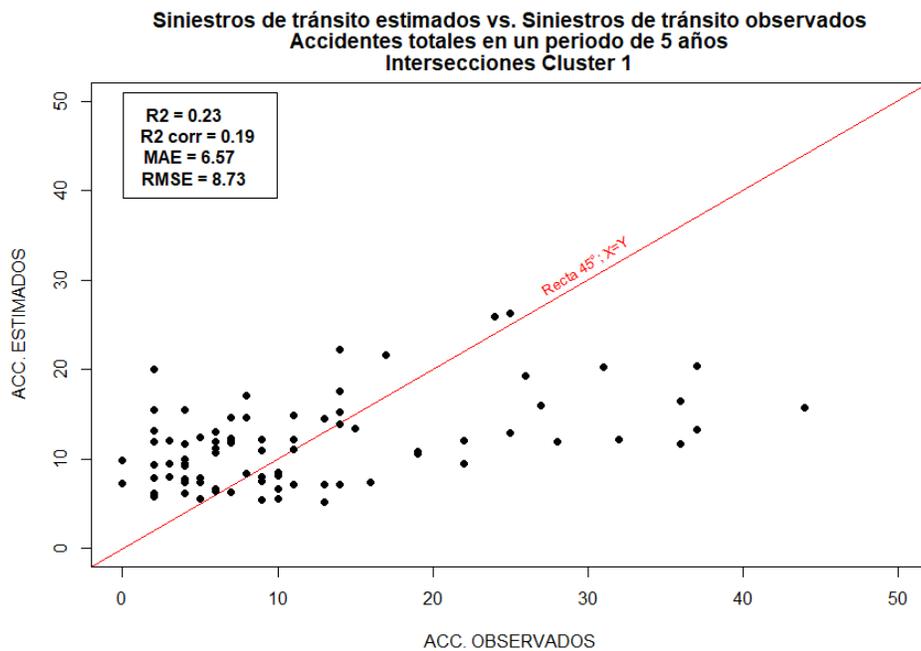
**Figura 4.1.** Representación gráfica de modelo de predicción de accidentes totales en base completa. Accidentes estimados vs. Accidentes observados.



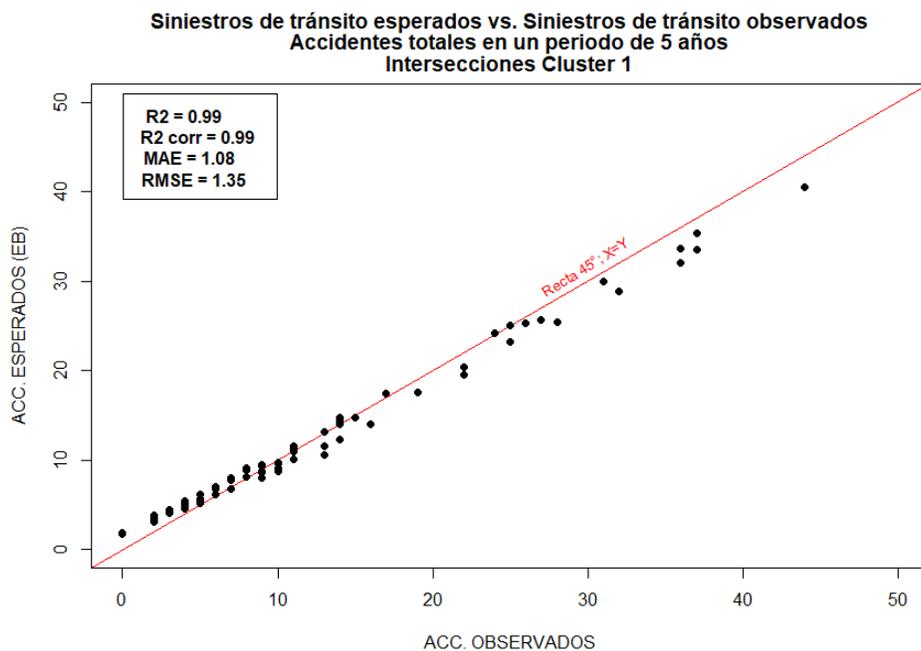
**Figura 4.2.** Representación gráfica de modelo de predicción de accidentes totales en base completa. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.2.** Modelo de predicción de accidentes totales en intersecciones del Cluster 1.

Modelo de predicción de accidentes – Accidentes Totales – Intersecciones Cluster 1							
$\ln(\mu) = 1.1721 + 0.2087 \cdot N\_PISTAS\_RS - 0.2875 \cdot PRESENCIA\_MEDIANA\_RP + 2.4403 \cdot 10^{-5} \cdot TMD\_RP + 0.0979 \cdot N\_PARADEROS$							
Grados libertad ( <i>gl</i> )	Pearson $\chi^2$	$\chi^2_{0.05,80}$	Parámetro <i>k</i>	Parámetro $\alpha$	Desviación ( <i>D</i> )	<i>D/gl</i>	<i>AIC</i>
80	83	102	3.96	0.44	91	1.14	572
Variable	Estimación	Error Est.	Valor Z	<i>Pr</i> (>  z )			
Intercepto	1.1721	0.3040	3.8556	0.0001			**
N_PISTAS_RS	0.2087	0.1079	1.9339	0.0531			*
PRESENCIA_MEDIANA_RP	-0.2875	0.2076	-1.3849	0.1661			
TMD_RP	2.4403E-05	6.4384E-06	3.7903	0.0002			**
N_PARADEROS	0.0979	0.0869	1.1261	0.2601			
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1							



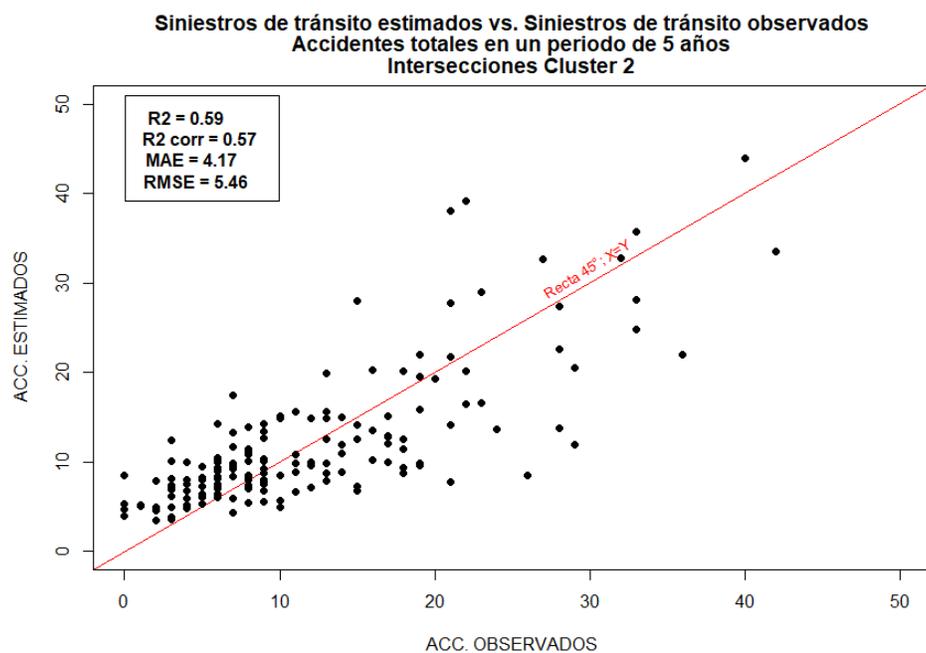
**Figura 4.3.** Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.



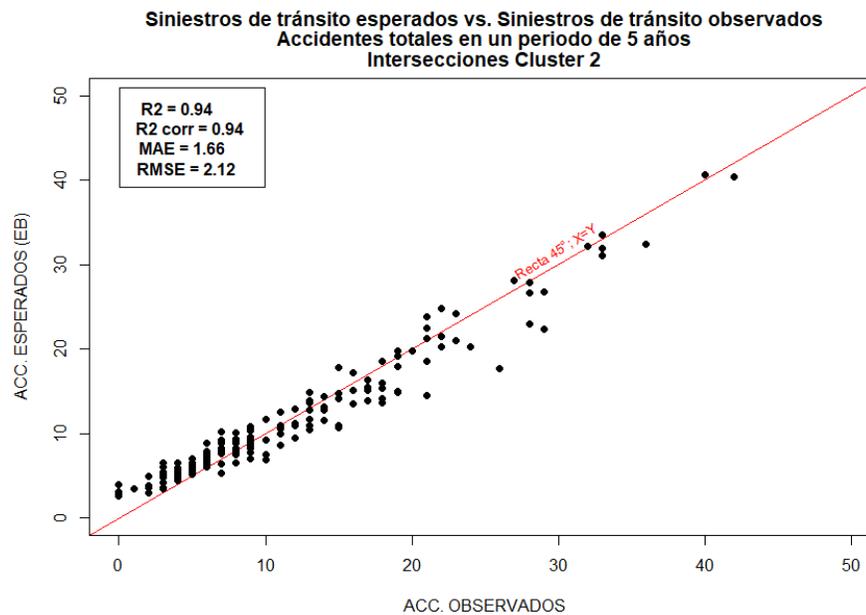
**Figura 4.4.** Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.3.** Modelo de predicción de accidentes totales en intersecciones del Cluster 2.

Modelo de predicción de accidentes – Accidentes Totales – Intersecciones Cluster 2							
$\ln(\mu) = 0.8400 + 0.2343 \cdot N\_PISTAS\_RS - 0.4379 \cdot PISTAS\_EXCLUSIVA\_GIRO\_DERECHA\_RP - 0.4662 \cdot PRESENCIA\_MEDIANA\_RS + 2.8416 \cdot 10^{-5} \cdot TMD\_RP + 1.9714 \cdot 10^{-5} \cdot TMD\_RS + 0.3039 \cdot ZONA\_COM\_MIX$							
Grados libertad ( <i>gl</i> )	Pearson $\chi^2$	$\chi^2_{0.05,161}$	Parámetro <i>k</i>	Parámetro $\alpha$	Desviación ( <i>D</i> )	<i>D/gl</i>	<i>AIC</i>
161	183	192	7.61	0.13	189	1.17	1021
Variable			Estimación	Error Est.	Valor Z	<i>Pr</i> (>  z )	
Intercepto			0.8400	0.1825	4.6020	4.1839E-6 **	
N_PISTAS_RS			0.2343	0.0820	2.8592	0.0042 **	
PISTA_EXCLUSIVA_GIRO_DERECHA_RP			-0.4379	0.1968	-2.2255	0.0260 **	
PRESENCIA_MEDIANA_RS			-0.4662	0.1954	-2.3856	0.0171 **	
TMD_RP			2.8416E-5	2.5394E-6	11.1903	4.5472E-29 **	
TMD_RS			1.9714E-5	9.2959E-6	2.1207	0.0339 **	
ZONA_COM_MIX			0.3039	0.0888	3.4216	0.0006 **	
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘.’ 1							



**Figura 4.5.** Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados



**Figura 4.6.** Representación gráfica de modelo de predicción de accidentes totales en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.

A partir de las tablas 4.1 a 4.3 se observa la presencia de sobredispersión (varianza mayor a la media) de los datos de accidentes totales para los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) al observar las medidas de dispersión  $\alpha$  (0.25, 0.44 y 0.13) y  $D/gl$  (1.13, 1.14 y 1.17) que, como menciona Alcaide (2015), al ser mayor a 0 y mayor a 1, respectivamente, indican sobredispersión en los datos, por lo tanto, se justifica la elección del modelo NB. Si  $\alpha$  fuera 0, sería un caso de equidispersión (varianza igual a la media).

En términos de bondad de ajuste, en las tablas 4.1 a 4.3 se observa que en los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) el valor del estadístico *Pearson*  $\chi^2$  (263, 83 y 183) es menor que el  $\chi^2$  crítico (283, 102 y 192) al 95% de nivel de confianza para los grados de libertad (245, 80 y 161) respectivos. Por lo tanto, se puede concluir que a un nivel de confianza del 95%, los tres modelos proporcionan un ajuste adecuado al conjunto de datos de accidentes totales utilizados en el estudio. En términos de valores p, se observa que el modelo para base completa y el Cluster 2 resultan significativos al 95% de nivel de confianza, mientras que el modelo para el Cluster 1 no, debido a que las variables N\_PISTAS\_RS, PRESENCIA\_MEDIANA\_RP y N\_PARADEROS presentan un valor p mayor a 0.1, es decir no es estadísticamente significativo.

En cuanto a la aplicación del método EB (figuras 4.1 a 4.6), los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) presentan una mejora significativa. El coeficiente de determinación corregido  $R^2_{\text{corr}}$  con los accidentes estimados (41%, 19% y 59%) aumenta significativamente en los accidentes esperados al aplicar el método EB (98%, 99%, 94%), respectivamente. A su vez, los errores MAE y RMSE disminuyen su valor en los accidentes esperados resultantes de la aplicación de la metodología EB.

Finalmente, a partir de los resultados para accidentes totales se observa que el Cluster 1 no logra resultados tan buenos como ocurre en la modelación de base completa y del Cluster 2. El Cluster 1 presenta el  $R^2_{\text{corr}}$  en accidentes estimados más bajo de los tres casos de intersecciones modeladas. Esto se debe principalmente al tamaño de la muestra, el cual en la modelación de base completa es de 253 entidades, en el Cluster 2 de 168 entidades y en el Cluster 1 de 85 entidades. Es por esto que, tal como menciona Alcaide (2015), el Cluster 1 al presentar un tamaño de muestra baja no se logra una estimación adecuada de los parámetros y su predicción empeora. En definitiva, la agrupación en *clusters* no resulta en una mejora significativa para la modelación, ya que el Cluster 2 no presenta un gran aumento del ajuste respecto a la base completa, y el Cluster 1 presenta una disminución importante en el ajuste respecto a la base completa.

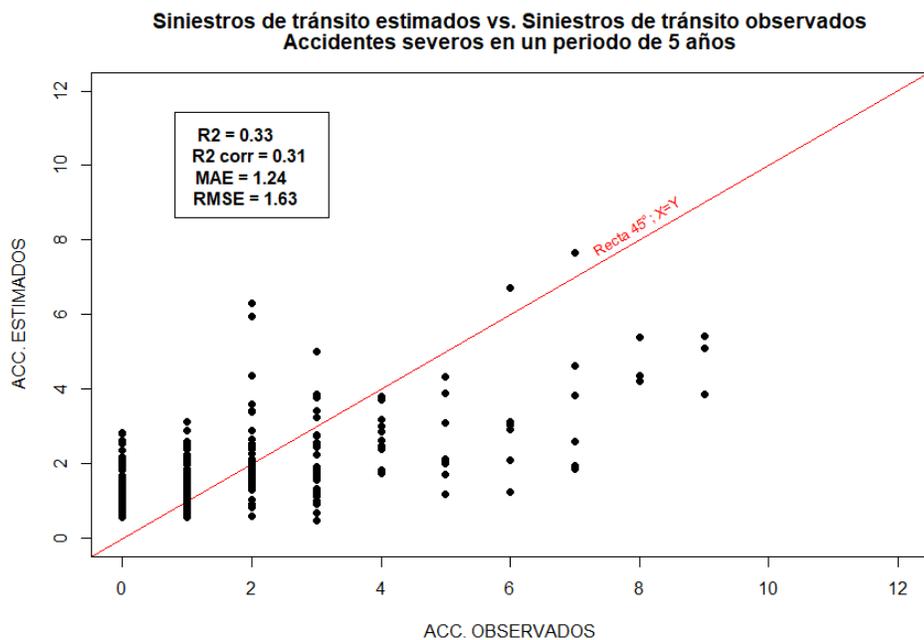
#### **4.2. Resultados respecto a accidentes severos**

En la Tabla 4.4, Tabla 4.5 y Tabla 4.6 se presentan los resultados de los modelos respecto a accidentes severos de base completa, Cluster 1 y Cluster 2, respectivamente. En ellas se encuentran la estimación de parámetros, errores estándar, valores p y medidas de bondad de ajuste de los modelos.

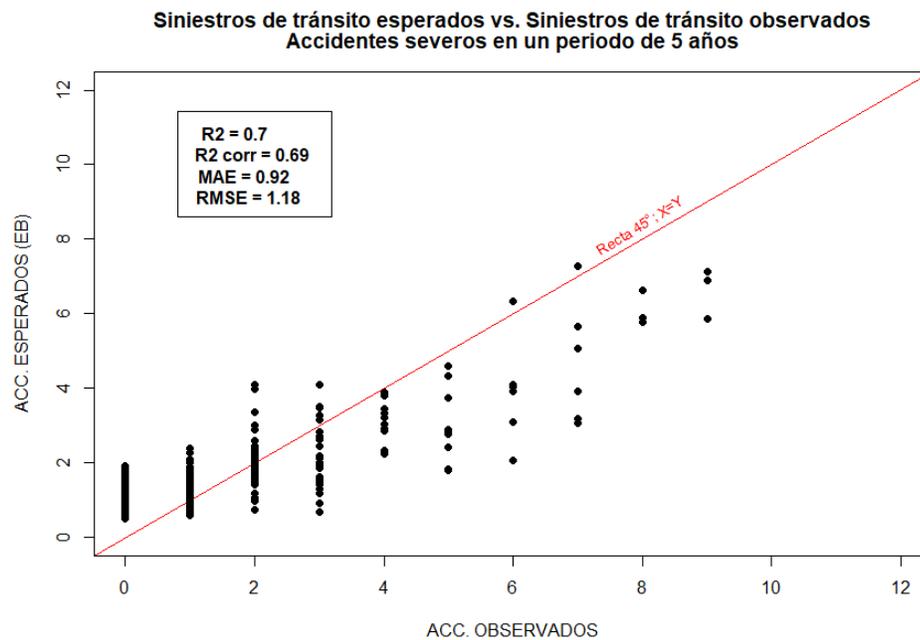
En la Figura 4.7 a Figura 4.12 se presentan los gráficos de los modelos NB y modelos NB con metodología EB respecto a accidentes severos de base completa, Cluster 1 y Cluster 2. En ellas se encuentran los coeficientes de determinación normal y corregido, y las medidas de error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE) obtenidos entre los accidentes estimados con los accidentes observados, y los accidentes esperados con los accidentes observados.

**Tabla 4.4.** Modelo de predicción de accidentes severos en base completa.

Modelo de predicción de accidentes – Accidentes Severos – Intersecciones base completa							
$\ln(\mu) = -0.922 + 0.1516 \cdot N\_PISTAS\_RP - 0.4993 \cdot PISTA\_EXCLUSIVA\_GIRO\_IZQUIERDA\_RS - 0.3525 \cdot PRESENCIA\_MEDIANA\_RP + 2.23 \cdot 10^{-5} \cdot TMD\_INTER + 4.58 \cdot 10^{-5} \cdot TMD\_GD + 0.1352 \cdot DENSIDAD\_PEATONAL$							
Grados libertad (gl)	Pearson $\chi^2$	$\chi^2_{0.05,246}$	Parámetro $k$	Parámetro $\alpha$	Desviación (D)	D/gl	AIC
246	267	284	6.01	0.17	293	1.19	881
Variable			Estimación	Error Est.	Valor Z	Pr (>  z )	
Intercepto			-0.9222	0.2335	-3.9489	7.85E-05	**
N_PISTAS_RP			0.1516	0.0603	2.5139	0.0119	**
PISTA_EXCLUSIVA_GIRO_IZQUIERDA_RS			-0.4993	0.1824	-2.7377	0.0062	**
PRESENCIA_MEDIANA_RP			-0.3525	0.2013	-1.7512	0.0799	*
TMD_INTER			2.23E-05	4.65E-06	4.8013	1.58E-06	**
TMD_GD			4.58E-05	2.08E-05	2.2017	0.0277	**
DENSIDAD_PEATONAL			0.1352	0.0773	1.7488	0.0803	*
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1							



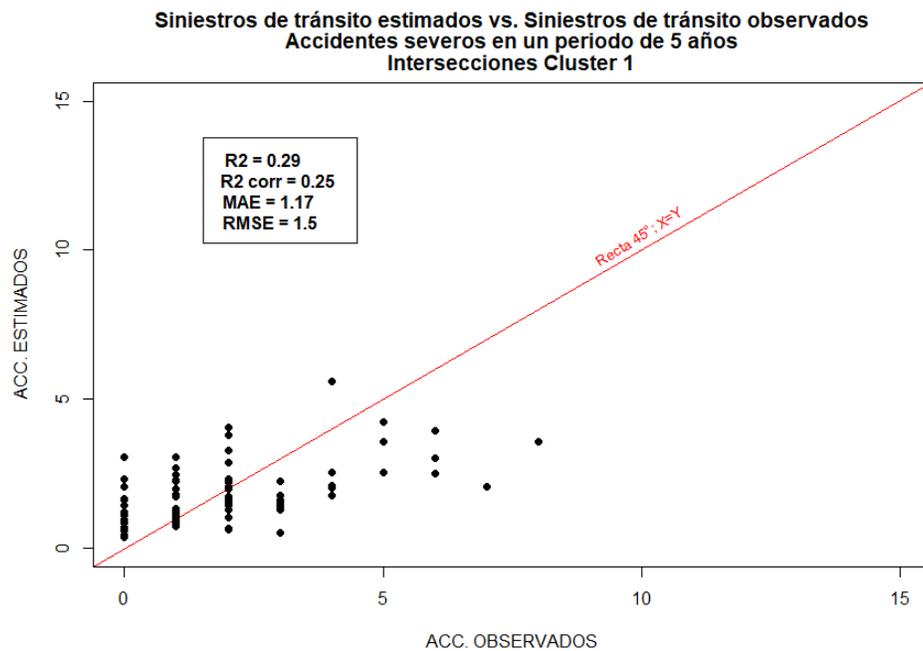
**Figura 4.7.** Representación gráfica de modelo de predicción de accidentes severos en todas intersecciones. Accidentes estimados vs. Accidentes observados.



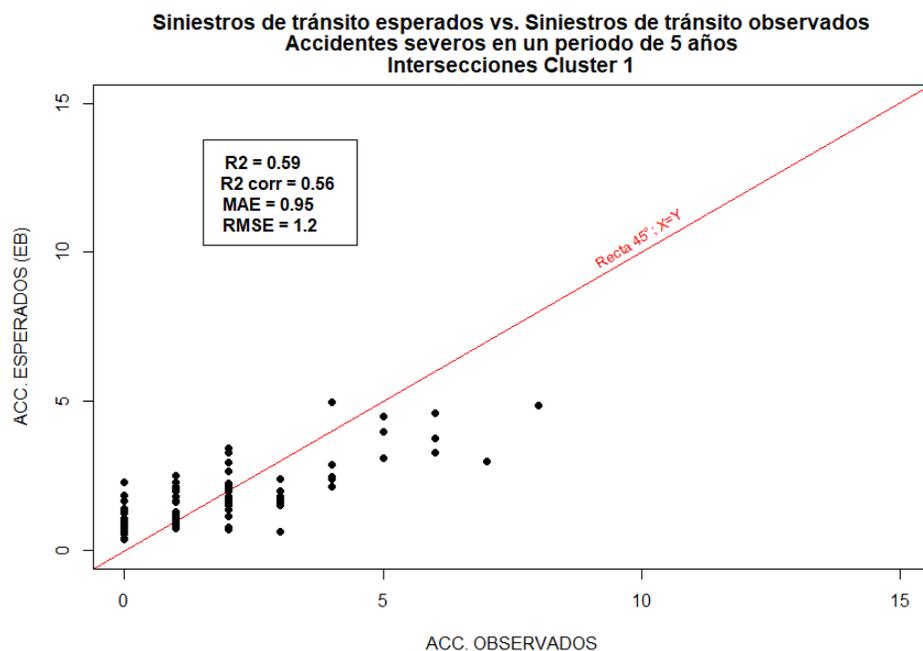
**Figura 4.8.** Representación gráfica de modelo de predicción de accidentes severos en todas intersecciones. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.5.** Modelo de predicción de accidentes severos en intersecciones del Cluster 1.

Modelo de predicción de accidentes – Accidentes Severos – Intersecciones Cluster 1							
$\ln(\mu) = 1.1074 - 0.4828 \cdot N\_RAMAS\_ACCESO - 0.5570 \cdot PISTA\_EXCLUSIVA\_GIRO\_IZQUIERDA\_RS + 2.4970 \cdot 10^{-5} \cdot TMD\_RP + 0.0001 \cdot TMD\_GD - 0.5084 \cdot ZONA\_COM\_MIX$							
<b>Grados libertad (gl)</b>	<b>Pearson <math>\chi^2</math></b>	<b><math>\chi^2_{0.05,79}</math></b>	<b>Parámetro <math>k</math></b>	<b>Parámetro <math>\alpha</math></b>	<b>Desviación (D)</b>	<b>D/gl</b>	<b>AIC</b>
79	86	101	8.87	0.11	95	1.20	290
<b>Variable</b>			<b>Estimación</b>	<b>Error Est.</b>	<b>Valor Z</b>	<b>Pr (&gt;  z )</b>	
Intercepto			1.1074	0.6010	1.8427	0.0654	*
N_RAMAS_ACCESO			-0.4828	0.2064	-2.3391	0.0193	**
PISTA_EXCLUSIVA_GIRO_IZQUIERDA_RS			-0.5570	0.2332	-2.3887	0.0169	**
TMD_RP			2.4970E-5	6.5871E-6	3.7908	0.0002	**
TMD_GD			0.0001	3.2670E-5	3.1922	0.0014	**
ZONA_COM_MIX			-0.5084	0.2046	-2.4853	0.0129	**
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘.’ 1							



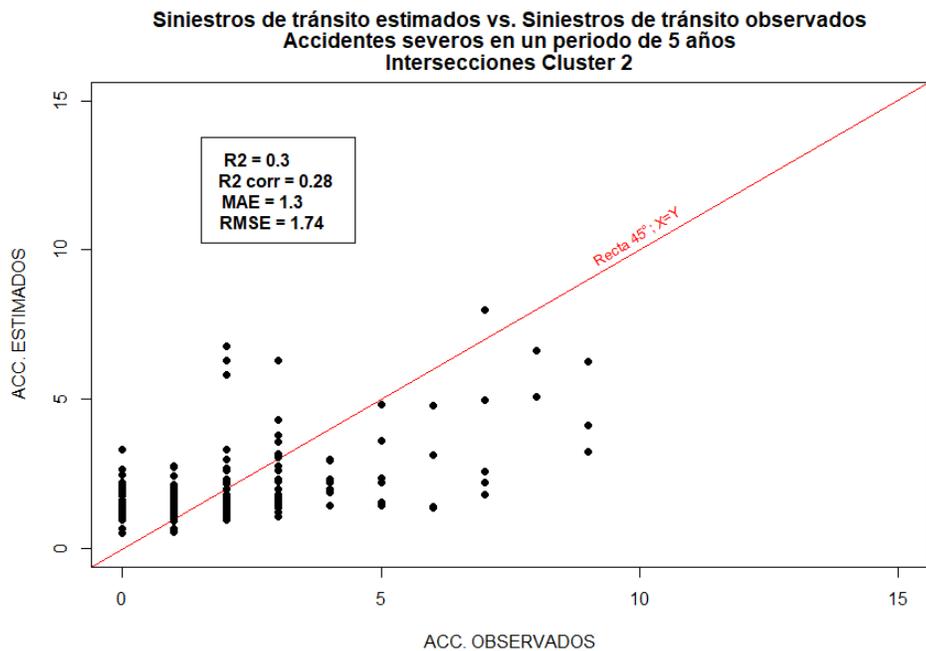
**Figura 4.9.** Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.



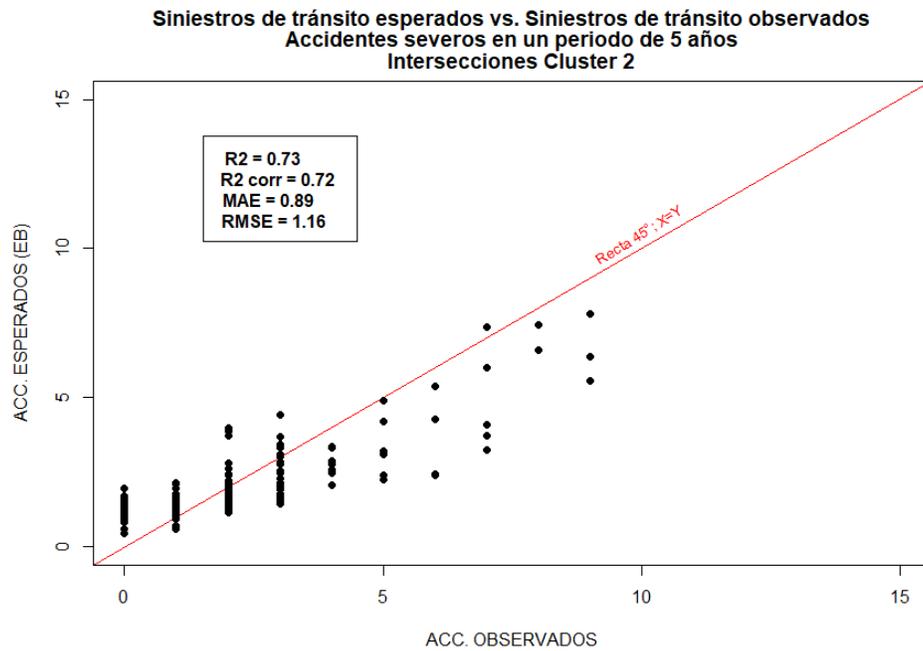
**Figura 4.10.** Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.6.** Modelo de predicción de accidentes severos en intersecciones del Cluster 2.

Modelo de predicción de accidentes – Accidentes Severos – Intersecciones Cluster 2							
$\ln(\mu) = -0.2460 + 0.3678 \cdot \text{PISTA\_EXCLUSIVA\_GIRO\_IZQUIERDA\_RP} - 0.6910 \cdot \text{PRESENCIA\_MEDIANA\_RS} + 2.66 \cdot 10^{-5} \cdot \text{TMD\_RP} + 3.02 \cdot 10^{-5} \cdot \text{TMD\_RS}$							
<b>Grados libertad (gl)</b>	<b>Pearson <math>\chi^2</math></b>	<b><math>\chi^2_{0.05,163}</math></b>	<b>Parámetro <math>k</math></b>	<b>Parámetro <math>\alpha</math></b>	<b>Desviación (D)</b>	<b>D/gl</b>	<b>AIC</b>
163	168	194	4.82	0.21	190	1.17	596
Variable			Estimación	Error Est.	Valor Z	Pr (>  z )	
Intercepto			-0.2460	0.1575	-1.5621	0.1183	
PISTA_EXCLUSIVA_GIRO_IZQUIERDA_RP			0.3678	0.1949	1.8866	0.0592	*
PRESENCIA_MEDIANA_RS			-0.6910	0.3516	-1.9657	0.0493	**
TMD_RP			2.66E-05	4.77E-06	5.5791	2.42E-08	**
TMD_RS			3.02E-05	1.39E-05	2.1638	0.0305	**
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1							



**Figura 4.11.** Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados.



**Figura 4.12.** Representación gráfica de modelo de predicción de accidentes severos en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.

A partir de las tablas 4.4 a 4.6 se observa la presencia de sobredispersión (varianza mayor a la media) de los datos de accidentes severos para los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) al observar las medidas de dispersión  $\alpha$  (0.17, 0.11 y 0.21) y  $D/gl$  (1.19, 1.20 y 1.17) que, como menciona Alcaide (2015), al ser mayor a 0 y mayor a 1, respectivamente, indican sobredispersión en los datos, por lo tanto, se justifica la elección del modelo NB. Si  $\alpha$  fuera 0, sería un caso de equidispersión (varianza igual a la media).

En términos de bondad de ajuste, en las tablas 4.4 a 4.6 se observa que en los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) el valor del estadístico *Pearson*  $\chi^2$  (267, 86 y 168) es menor que el  $\chi^2$  crítico (284, 101 y 194) al 95% de nivel de confianza para los grados de libertad (246, 79 y 163) respectivos. Por lo tanto, se puede concluir que a un nivel de confianza del 95%, los tres modelos proporcionan un ajuste adecuado al conjunto de datos de accidentes severos utilizados en el estudio. En términos de valores p, se observa que el modelo para base completa y el Cluster 2 resultan significativos al 90% de nivel de confianza debido a que presentan valores p en el rango 0.05 a 0.1, mientras que el modelo para el Cluster 1 resultó significativo al 95% de nivel de confianza.

En cuanto a la aplicación del método EB (figuras 4.7 a 4.12), los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) presentan una mejora significativa. El coeficiente de determinación corregido  $R^2_{\text{corr}}$  con los accidentes estimados (31%, 25% y 28%) aumenta significativamente en los accidentes esperados al aplicar el método EB (69%, 56%, 72%), respectivamente. A su vez, los errores MAE y RMSE disminuyen su valor en los accidentes esperados resultantes de la aplicación de la metodología EB.

Finalmente, a partir de los resultados para accidentes severos se observa que el Cluster 1 logra ajustes más pobres que la modelación de base completa y del Cluster 2. El Cluster 1 presenta el  $R^2_{\text{corr}}$  en accidentes estimados y esperados más bajo de los tres casos de intersecciones modeladas. Esto se debe principalmente al tamaño de la muestra, el cual en la modelación de base completa es de 253 entidades, en el Cluster 2 de 168 entidades y en el Cluster 1 de 85 entidades. Es por esto que, tal como menciona Alcaide (2015), el Cluster 1 al presentar un tamaño de muestra baja no se logra una estimación adecuada de los parámetros y su predicción empeora. En definitiva, la agrupación en *clusters* no resulta en una mejora para la modelación, ya que el Cluster 2 y Cluster 1 presentan una disminución en el ajuste respecto a la base completa.

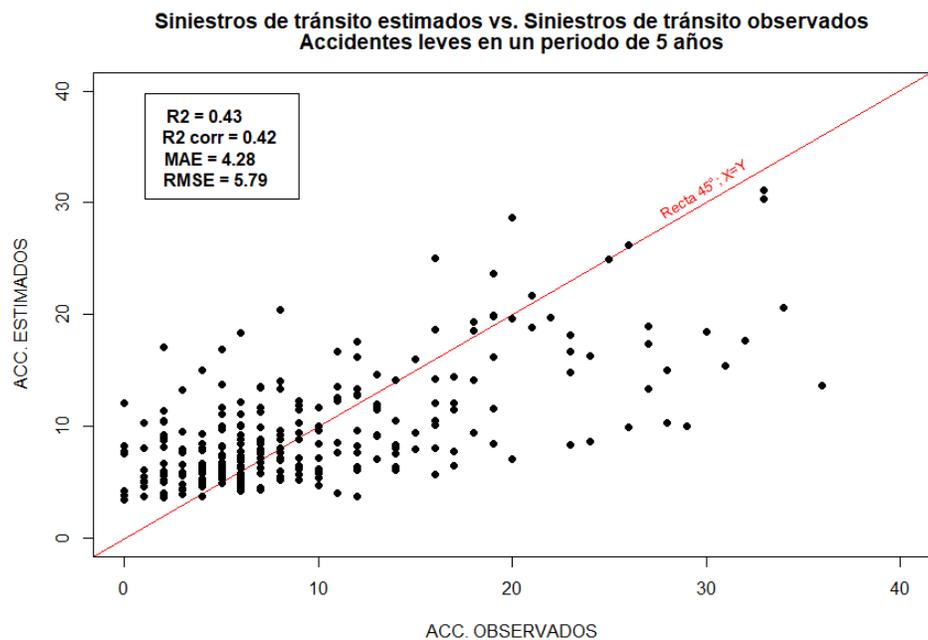
### 4.3. Resultados respecto a accidentes leves

En la Tabla 4.7, Tabla 4.8 y Tabla 4.9 se presentan los resultados de los modelos respecto a accidentes leves de base completa, Cluster 1 y Cluster 2, respectivamente. En ellas se encuentran la estimación de parámetros, errores estándar, valores p y medidas de bondad de ajuste de los modelos.

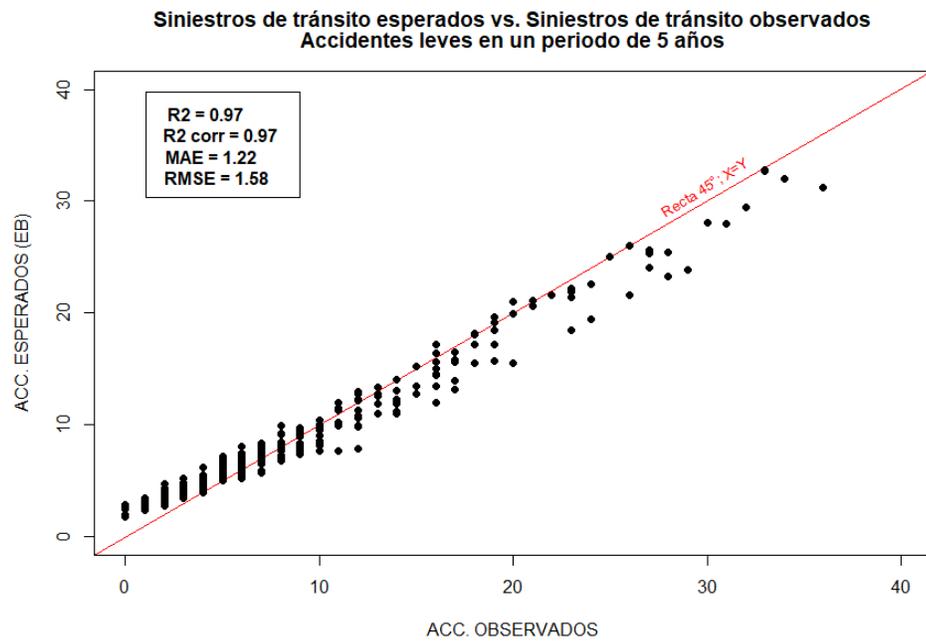
En la Figura 4.13 a Figura 4.18 se presentan los gráficos de los modelos NB y modelos NB con metodología EB respecto a accidentes leves de base completa, Cluster 1 y Cluster 2. En ellas se encuentran los coeficientes de determinación normal y corregido, y las medidas de error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE) obtenidos entre los accidentes estimados con los accidentes observados, y los accidentes esperados con los accidentes observados.

**Tabla 4.7.** Modelo de predicción de accidentes leves en base completa.

Modelo de predicción de accidentes – Accidentes Leves – Intersecciones base completa							
$\ln(\mu) = 0.6140 + 0.2561 \cdot N\_PISTAS\_RS - 0.3191 \cdot PRESENCIA\_MEDIANA\_RS + 2.34 \cdot 10^{-5} \cdot TMD\_RP + 3.24 \cdot 10^{-5} \cdot TMD\_GD + 0.2648 \cdot ZONA\_COM\_MIX + 0.1153 \cdot N\_PARADEROS$							
<b>Grados libertad (gl)</b>	<b>Pearson <math>\chi^2</math></b>	$\chi^2_{0.05,246}$	<b>Parámetro <math>k</math></b>	<b>Parámetro <math>\alpha</math></b>	<b>Desviación (D)</b>	<b>D/gl</b>	<b>AIC</b>
246	250	284	3.73	0.27	280	1.14	1539
Variable			Estimación	Error Est.	Valor Z	Pr (>  z )	
Intercepto			0.6140	0.1814	3.3845	0.0007	**
N_PISTAS_RS			0.2561	0.0769	3.3285	0.0009	**
PRESENCIA_MEDIANA_RS			-0.3191	0.1712	-1.8634	0.0624	*
TMD_RP			2.34E-05	2.71E-06	8.6185	6.78E-18	**
TMD_GD			3.24E-05	1.63E-05	1.9869	0.0469	**
ZONA_COM_MIX			0.2648	0.0848	3.1238	0.0018	**
N_PARADEROS			0.1153	0.0484	2.3816	0.0172	**
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1							



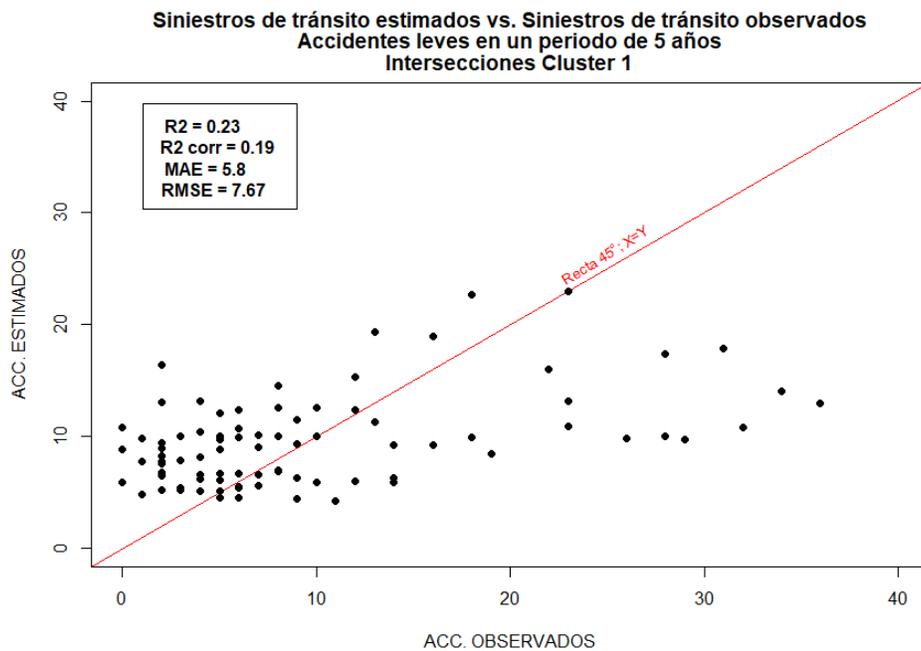
**Figura 4.13.** Representación gráfica de modelo de predicción de accidentes leves en base completa. Accidentes estimados vs. Accidentes observados.



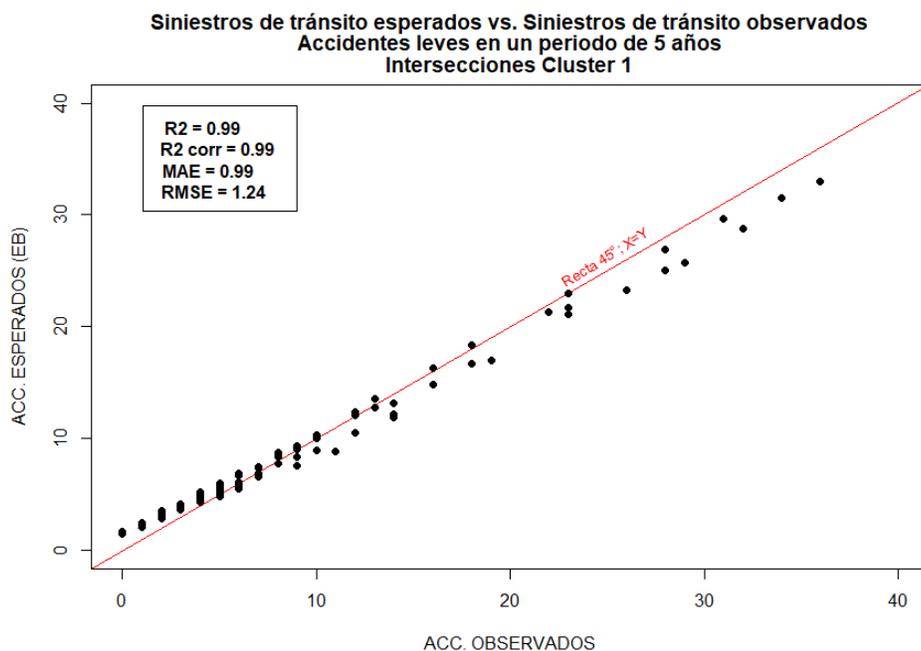
**Figura 4.14.** Representación gráfica de modelo de predicción de accidentes leves en base completa. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.8.** Modelo de predicción de accidentes leves en intersecciones del Cluster 1.

Modelo de predicción de accidentes – Accidentes Leves – Intersecciones Cluster 1							
$\ln(\mu) = 0.9446 + 0.2375 \cdot N\_PISTAS\_RS - 0.3106 \cdot PRESENCIA\_MEDIANA\_RP + 2.4337 \cdot 10^{-5} \cdot TMD\_RP + 0.1016 \cdot N\_PARADEROS$							
Grados libertad (gl)	Pearson $\chi^2$	$\chi^2_{0.05,80}$	Parámetro $k$	Parámetro $\alpha$	Desviación (D)	D/gl	AIC
80	81	102	1.99	0.50	93	1.16	550
Variable			Estimación	Error Est.	Valor Z	Pr (>  z )	
(Intercept)			0.9446	0.3264	2.8940	0.0038	**
N_PISTAS_RS			0.2375	0.1156	2.0540	0.0400	**
PRESENCIA_MEDIANA_RP			-0.3106	0.2230	-1.3933	0.1635	
TMD_RP			2.4337E-5	6.9146E-6	3.5196	0.0004	**
N_PARADEROS			0.1016	0.0934	1.0878	0.2767	
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1							



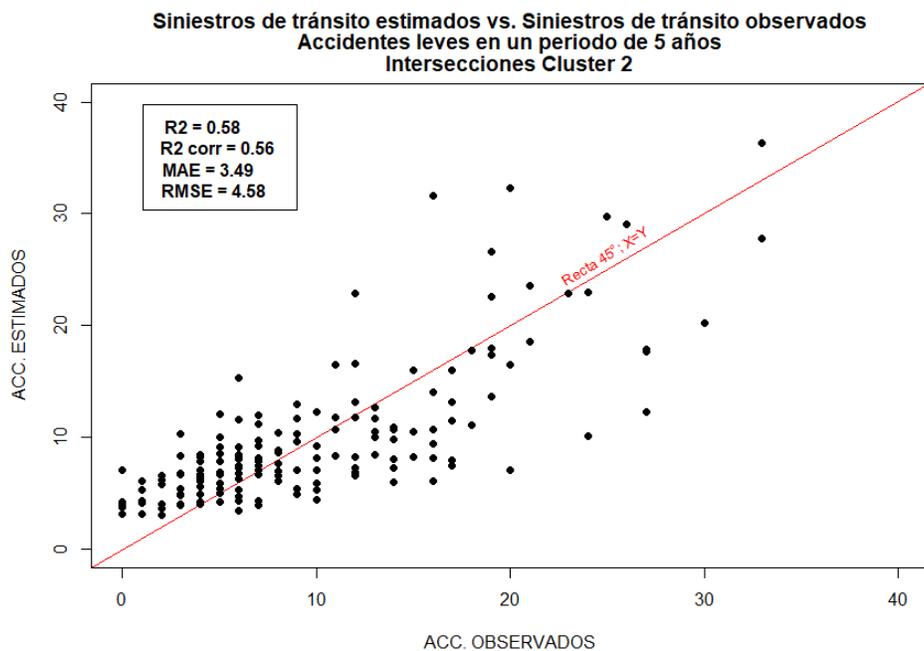
**Figura 4.15.** Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 1. Accidentes estimados vs. Accidentes observados.



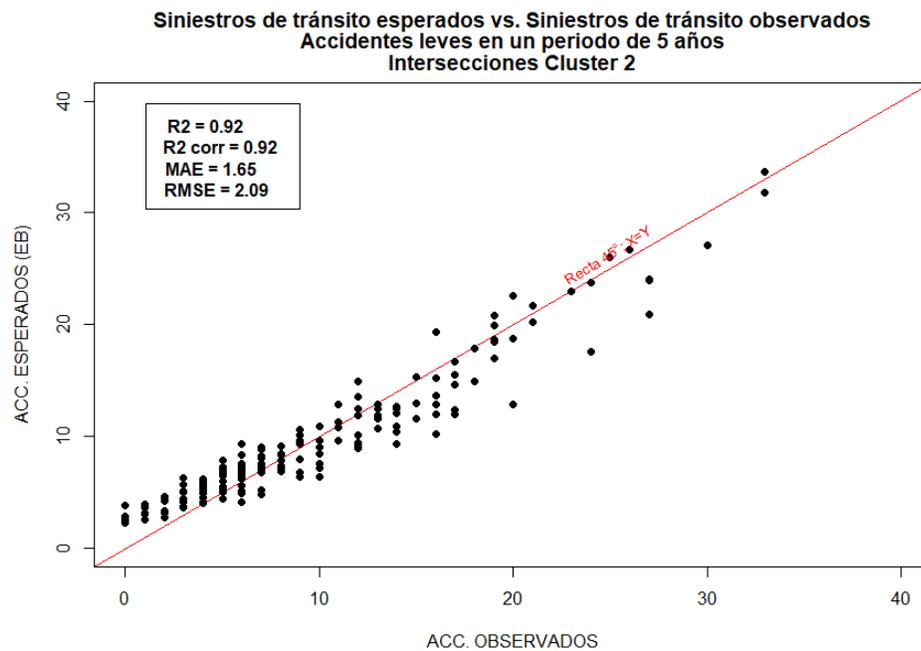
**Figura 4.16.** Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 1. Accidentes esperados (EB) vs. Accidentes observados.

**Tabla 4.9.** Modelo de predicción de accidentes leves en intersecciones del Cluster 2.

Modelo de predicción de accidentes – Accidentes Leves – Intersecciones Cluster 2							
$\ln(\mu) = 0.5829 + 0.2489 \cdot N\_PISTAS\_RS - 0.4097 \cdot PISTA\_EXCLUSIVA\_GIRO\_DERECHA\_RP - 0.4348 \cdot PRESENCIA\_MEDIANA\_RS + 2.7833 \cdot 10^{-5} \cdot TMD\_RP + 2.0326 \cdot 10^{-5} \cdot TMD\_RS + 0.3561 \cdot ZONA\_COM\_MIX$							
Grados libertad ( <i>gl</i> )	Pearson $\chi^2$	$\chi^2_{0.05,161}$	Parámetro <i>k</i>	Parámetro $\alpha$	Desviación ( <i>D</i> )	<i>D/gl</i>	<i>AIC</i>
161	182	192	8.53	0.12	190	1.18	963
Variable			Estimación	Error Est.	Valor Z	<i>Pr</i> (>  z )	
Intercepto			0.5829	0.1850	3.1517	0.0016	**
N_PISTAS_RS			0.2489	0.0822	3.0265	0.0025	**
PISTA_EXCLUSIVA_GIRO_DERECHA_RP			-0.4097	0.1990	-2.0588	0.0395	**
PRESENCIA_MEDIANA_RS			-0.4348	0.1960	-2.2189	0.0265	**
TMD_RP			2.7833E-5	2.5363E-6	10.9738	5.1097E-28	**
TMD_RS			2.0326E-05	9.2963E-6	2.1865	0.0288	**
ZONA_COM_MIX			0.3561	0.0909	3.9198	8.8638E-5	**
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘.’ 1							



**Figura 4.17.** Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 2. Accidentes estimados vs. Accidentes observados.



**Figura 4.18.** Representación gráfica de modelo de predicción de accidentes leves en intersecciones del Cluster 2. Accidentes esperados (EB) vs. Accidentes observados.

A partir de las tablas 4.7 a 4.9 se observa la presencia de sobredispersión (varianza mayor a la media) de los datos de accidentes leves para los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) al observar las medidas de dispersión  $\alpha$  (0.27, 0.5 y 0.12) y  $D/gl$  (1.14, 1.16 y 1.18) que, como menciona Alcaide (2015), al ser mayor a 0 y mayor a 1, respectivamente, indican sobredispersión en los datos, por lo tanto, se justifica la elección del modelo NB. Si  $\alpha$  fuera 0, sería un caso de equidispersión (varianza igual a la media).

En términos de bondad de ajuste, en las tablas 4.7 a 4.9 se observa que en los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) el valor del estadístico *Pearson*  $\chi^2$  (250, 81 y 182) es menor que el  $\chi^2$  crítico (284, 102 y 192) al 95% de nivel de confianza para los grados de libertad (246, 80 y 161) respectivos. Por lo tanto, se puede concluir que a un nivel de confianza del 95%, los tres modelos proporcionan un ajuste adecuado al conjunto de datos de accidentes leves utilizados en el estudio. En términos de valores p, se observa que el modelo para base completa resulta significativo al 90% de nivel de confianza, el Cluster 2 resulta significativo al 95% de nivel de confianza, mientras que el modelo para el Cluster 1 no resulta

significativo, debido a que las variables PRESENCIA\_MEDIANA\_RP y N\_PARADEROS presentan un valor p mayor a 0.1, es decir no es estadísticamente significativo.

En cuanto a la aplicación del método EB (figuras 4.13 a 4.18), los tres casos de intersecciones modeladas (base completa, Cluster 1 y Cluster 2) presentan una mejora significativa. El coeficiente de determinación corregido  $R^2_{\text{corr}}$  con los accidentes estimados (42%, 19% y 56%) aumenta significativamente en los accidentes esperados al aplicar el método EB (97%, 99%, 92%), respectivamente. A su vez, los errores MAE y RMSE disminuyen su valor en los accidentes esperados resultantes de la aplicación de la metodología EB.

Finalmente, a partir de los resultados para accidentes leves se observa que el Cluster 1 no logra resultados tan buenos como ocurre en la modelación de base completa y del Cluster 2. El Cluster 1 presenta el  $R^2_{\text{corr}}$  en accidentes estimados más bajo de los tres casos de intersecciones modeladas. Esto se debe principalmente al tamaño de la muestra, el cual en la modelación de base completa es de 253 entidades, en el Cluster 2 de 168 entidades y en el Cluster 1 de 85 entidades. Es por esto que, tal como menciona Alcaide (2015), el Cluster 1 al presentar un tamaño de muestra baja no se logra una estimación adecuada de los parámetros y su predicción empeora. En definitiva, la agrupación en *clusters* no resulta en una mejora significativa para la modelación, ya que el Cluster 2 no presenta un gran aumento del ajuste respecto a la base completa, y el Cluster 1 presenta una disminución importante en el ajuste respecto a la base completa.

#### 4.4. Comparación de modelos

Según los resultados de modelación de accidentes totales, severos y leves en los distintos casos de las tablas 4.1 a 4.9, en la Tabla 4.10 se resumen los coeficientes de determinación corregidos obtenidos en los modelos sin el método EB y con el método EB, respectivamente. Se observó que los mejores ajustes sin EB se logran al explicar la variable dependiente de accidentes totales y leves, mientras que en los accidentes severos no se logra tan buen ajuste. Como se mencionó anteriormente, esto se debe a la sobredispersión de los datos presentes en accidentes totales y leves y para los cuales el modelo NB se adecua. Cabe destacar que en el caso de modelación de accidentes severos hay gran cantidad de registros de accidentes con valores igual a cero y valores cercanos

entre sí, por lo cual no existe un gran sobredispersión en los datos, lo que conlleva en el bajo ajuste del modelo de estimación. Debido a lo anterior ocurre el fenómeno de que en la modelación de accidentes severos la mejora del método EB no sea tan significativa como en accidentes totales y leves, ya que como se observa en la ecuación 3.14, los accidentes estimados ( $\mu_i$ ) van dividiendo en el cálculo del peso ( $w_i$ ), y al ser pocos los accidentes severos, es decir su estimación son valores bajos, este peso resulta en un valor alto, lo que conlleva a que en la ecuación 3.13 los accidentes estimados ( $\mu_i$ ), que ya presentaban mal ajuste, tengan un mayor peso que los históricos ( $y_i$ ), por lo tanto los accidentes esperados (EB) no resultan con tan buen ajuste.

**Tabla 4.10.** Resumen de coeficientes de determinación corregidos ( $R^2_{\text{corr}}$ ) normal y EB en los casos modelados.

	Accidentes totales	Accidentes severos	Accidentes leves
<b>Base completa</b>	41%; 98%	31%; 69%	42%; 97%
<b>Cluster 1</b>	19%; 99%	25%; 56%	19%; 99%
<b>Cluster 2</b>	57%; 94%	28%; 72%	56%; 92%

Generalmente, los casos de modelación para base completa y Cluster 2 son los que presentan mejores ajustes ya que, como se mencionó anteriormente, estos presentan mayor cantidad de datos (intersecciones) lográndose así una mejor estimación de parámetros.

Finalmente, se observó que el método EB proporciona una mejora muy significativa en la precisión y ajuste de los modelos, aumentando los coeficientes de determinación corregidos a un nivel sobre el 90%, excepto en los casos de modelación de accidentes severos debido a lo explicado anteriormente.

#### 4.5. Interpretación, análisis y discusión respecto a las variables explicativas

Según los resultados de modelación de accidentes totales, severos y leves de las tablas 4.1 a 4.9, en la Tabla 4.11 se resumen las significancias y positividad o negatividad de los parámetros estimados de las variables y para cada uno de los casos modelados. En la Tabla 4.12 se presentan los efectos, expresados en porcentajes de aumento o disminución, de las variables explicativas en los accidentes estimados para en cada modelo.

**Tabla 4.11.** Resumen de variables explicativas resultantes de los modelos.

Variable Explicativa	Casos de modelación								
	Accidentes Totales			Accidentes Severos			Accidentes Leves		
	Base comp.	Cluster 1	Cluster 2	Base comp.	Cluster 1	Cluster 2	Base comp.	Cluster 1	Cluster 2
<b>Infraestructura</b>									
Nº PISTAS RP	** (+)			** (+)					
Nº PISTAS RS	** (+)	* (+)	** (+)				** (+)	** (+)	** (+)
PISTA EXCLUSIVA GIRO DERECHA RP			** (-)						** (-)
PISTA EXCLUSIVA GIRO IZQUIERDA RP						* (+)			
PISTA EXCLUSIVA GIRO IZQUIERDA RS				** (-)	** (-)				
Nº RAMAS ACCESO					** (-)				
PRESENCIA MEDIANA RP	** (-)	(-)		* (-)				(-)	
PRESENCIA MEDIANA RS			** (-)			** (-)	* (-)		** (-)
<b>Operación</b>									
TMD INTER				** (+)					
TMD RP	** (+)	** (+)	** (+)		** (+)	** (+)	** (+)	** (+)	** (+)
TMD RS	** (+)		** (+)			** (+)			** (+)
TMD GD				** (+)	** (+)		** (+)		
DENSIDAD PEATONAL				* (+)					
<b>Entorno</b>									
ZONA COM/MIX	** (+)		** (+)		** (-)		** (+)		** (+)
Nº PARADEROS	* (+)	(+)					* (+)	(+)	
Rangos significancia: 0 ‘***’ 0.05 ‘*’ 0.1 ‘ ’ 1 (+) Parámetro positivo. (-) Parámetro negativo									

**Tabla 4.12.** Efecto de variables explicativas en la frecuencia de accidentes.

Variable Explicativa	Casos de modelación								
	Accidentes Totales			Accidentes Severos			Accidentes Leves		
	Base comp.	Cluster 1	Cluster 2	Base comp.	Cluster 1	Cluster 2	Base comp.	Cluster 1	Cluster 2
<b>Infraestructura</b>									
N° PISTAS RP	+9.5%			+16.4%					
N° PISTAS RS	+13.9%	+23.2%	+26.4%				+29.2%	+26.8%	+28.3%
PISTA EXCLUSIVA GIRO DERECHA RP			-35.5%						-33.6%
PISTA EXCLUSIVA GIRO IZQUIERDA RP						+44.5%			
PISTA EXCLUSIVA GIRO IZQUIERDA RS				-39.3%	-42.7%				
N° RAMAS ACCESO					-38.3%				
PRESENCIA MEDIANA RP	-28.8%	-25.0%		-29.7%				-26.7%	
PRESENCIA MEDIANA RS			-37.3%			-49.9%	-27.3%		-35.3%
<b>Operación</b>									
TMD INTER (por cada +10% de TMD)				+4.5%					
TMD RP (por cada +10% de TMD)	+4.9%	+5.0%	+5.8%		+5.1%	+5.5%	+4.8%	+5.0%	+5.7%
TMD RS (por cada +10% de TMD)	+5.1%		+4.0%			+6.2%			+4.1%
TMD GD (por cada +10% de TMD)				+9.58%	+23.2%		+6.7%		
DENSIDAD PEATONAL				+14.5%					
<b>Entorno</b>									
ZONA COM/MIX	+21.9%		+35.5%		-39.9%		+30.3%		+42.8%
N° PARADEROS	+10.2%	+10.3%					+12.2%	+10.7%	

Dicho porcentaje se calculó como una tasa de cambio entre un valor inicial y final que toma la variable, mediante la expresión de la siguiente ecuación derivada de la forma funcional de los modelos:

$$\left( \frac{e^{\text{Estimación} \cdot \text{ValorFinal}} - e^{\text{Estimación} \cdot \text{ValorInicial}}}{e^{\text{Estimación} \cdot \text{ValorInicial}}} \right) \cdot 100 \quad (4.1)$$

La expresión anterior utiliza el parámetro estimado de la variable en el modelo correspondiente y dos valores (inicial y final) que toma dicha variable en el set de datos. Por ejemplo, para evaluar el aumento generado por el número de pistas de la rama secundaria, el valor inicial sería 1 (una pista) y el valor final sería 2 (dos pistas), obteniéndose el aumento de accidentes que genera el incremento de 1 pista en dicha rama.

A partir de los resultados presentados anteriormente en las tablas 4.1 a 4.12, las variables explicativas finales de los modelos se interpretan de la siguiente manera.

#### 4.5.1. Variables explicativas de infraestructura

Las variables de números de pistas por rama principal y secundaria se encuentran principalmente en los modelos de accidentes totales y leves (ver Tabla 4.11), por lo cual no son significativos en la generación de accidentes severos. Estas variables presentan una estimación de parámetro de signo positivo en todos los casos, es decir que a mayor cantidad de pistas en alguna de las ramas se genera una disminución de la seguridad y aumenta el riesgo de ocurrencia de un siniestro de tránsito. Esto se encuentra en conformidad con los hallazgos en investigaciones anteriores de Layfield *et al.* (1996), Greibe (2003) y Dufeu (2019), donde concluyen que el número de pistas de las ramas genera aumentos en la cantidad de accidentes. Lo anterior se explicaría por las maniobras riesgosas de adelantamiento y cambio de carril que podrían realizar los usuarios e implicar un peligro ya que requiere habilidad por parte del usuario conductor y requiere que los demás usuarios en la vía también estén atentos ante tal maniobra. A partir de los datos de la Tabla 4.12 se observó que el número de pistas de la rama principal generó aumentos entre 9.5% a 16.4% en la frecuencia de accidentes por cada aumento de 1 pista en dicha rama, mientras que el número de pistas de la rama secundaria generó aumentos entre 13.9% a 29.2% por cada aumento de pista. Además, cabe

destacar que el número de pistas de la rama secundaria, que se encuentra presente en los modelos de ambos *clusters* para accidentes totales y leves, genera mayor cantidad de accidentes de tránsito en el Cluster 2 que en el Cluster 1, con diferencias de aumento de accidentes entre ambos *clusters* de 3.2% para accidentes totales y 1.5% en accidentes leves. Recordar que el Cluster 2 estaba conformado principalmente por intersecciones de tipo X, por lo cual hay más ramas de acceso que en el Cluster 1, conformado principalmente por intersecciones tipo T, lo cual explicaría que un aumento en el número de pistas genere mayor cantidad de accidentes en el Cluster 2.

Las variables que indican presencia de pista exclusiva de giro a la derecha o izquierda en las ramas principal y secundaria, resultaron significativas en cinco de los nueve modelos. La presencia de pista exclusiva de giro a la derecha en rama principal se encuentra en los modelos del Cluster 2 para accidentes totales y leves, mientras que la presencia de pista exclusiva de giro a la izquierda en la rama secundaria se encuentra los modelos de accidentes severos para base completa y Cluster 1. Estas variables resultan en una estimación de parámetro de signo negativo, lo cual indica que la presencia de pistas de viraje proporciona seguridad en la intersección, disminuyendo la frecuencia de accidentes de tránsito, a excepción de la presencia de pista exclusiva de giro izquierda en rama principal que resulto en un aumento de los accidentes severos para el Cluster 2. La disminución de accidentes se encuentra en conformidad con los resultados presentados en el estudio de SECTRA (2014), donde indican que las pistas de viraje generan un factor de reducción de accidentes. Lo anterior se explicaría debido a que la presencia de vehículos en las pistas de viraje advierte al usuario que viene en el sentido opuesto y es de esperar que esté atento a la maniobra del usuario que está en la pista de viraje. Cuando la pista exclusiva de viraje no existe, ocurre que hay una pista que es compartida por los movimientos de viraje y de continuación recta, por lo tanto, puede existir conflicto entre ambos movimientos y generarse colisiones. En el caso de la presencia en pista exclusiva de giro izquierda en rama principal que genera aumento en el conteo de accidentes se debe a que tal movimiento de giro izquierda desde la rama principal estaría generando accidentes y colisiones probablemente con peatones o ciclistas que circulan y cruzan la rama secundaria. A partir de los datos de la Tabla 4.12 se observó que la pista exclusiva de giro derecha en la rama principal genera disminución de frecuencia de accidentes entre 33.6% a 35.5%, la pista exclusiva de giro izquierda en la rama principal genero aumento de accidentes en 44.5% y la pista exclusiva de giro izquierda en la rama secundaria genero disminución de accidentes entre 39.3% y 42.7%.

El número de ramas de acceso a la intersección es una variable que aparece como variable final solo en el modelo de accidentes severos para el Cluster 1. Esta presenta una estimación de parámetro estadísticamente significativa y negativa, lo que indica que a mayor cantidad de accesos que presenta la intersección se genera una baja en la frecuencia de accidentes de tránsito. Este resultado es opuesto a lo visto en el trabajo de Huang *et al.* (2017), en el cual en número de ramas presenta una correlación positiva con los siniestros de tránsito. Lo anterior ocurre por el hecho de que una intersección con mayor cantidad de ramas no presentara gran densidad peatonal comparado con una intersección de baja cantidad de ramas que suelen emplazarse en zonas céntricas y comerciales, por lo tanto, los accidentes ocurridos no serán severos ni fatales, entonces para el caso modelado de accidentes severos para el Cluster 1 se estaría generando una disminución de los siniestros. A partir de los datos de la Tabla 4.12 se observó que el número de ramas de acceso genera una disminución de 38.3% en la frecuencia de accidentes de tránsito severos para el Cluster 1.

La presencia de mediana en la rama principal y secundaria son variables que se encuentran en todos los modelos de accidentes totales, severos y leves, exceptuando el modelo de accidentes severos para el Cluster 1. Presentan una estimación de parámetro de signo negativo, lo cual indica que en las entidades cuyas ramas de acceso presentan mediana separando los sentidos del tránsito se genera una disminución en la ocurrencia de accidentes. Dicho resultado se encuentra en conformidad con los hallazgos en investigaciones anteriores de SECTRA (2014) y Alarifi *et al.* (2018), donde concluyen que la presencia de mediana en las ramas implica un aumento en la seguridad y una disminución en la frecuencia de accidentes. Lo anterior se explicaría por la separación generada entre las calzadas de tránsito opuesto y la presencia de las soleras de la mediana que ayudan a evitar que los vehículos se crucen a la calzada de sentido contrario por algún motivo externo. Además, la presencia de mediana implica mayor espacio de maniobra para los vehículos que realizan algún movimiento de giro a la izquierda en la intersección, pudiendo ser también un espacio de seguridad para el usuario antes de completar el giro. A partir de los datos de la Tabla 4.12 se observó que la presencia de mediana en la rama principal generó disminuciones entre 25.0% a 29.7%, mientras que la presencia de mediana en la rama secundaria generó disminuciones entre 27.3% a 49.9%.

#### 4.5.2. Variables explicativas de operación

En cuanto a variables de tipo operación, las variables de flujo (TMD) son las más presentes en los modelos finales. El flujo es una variable importante en la modelación y siempre es utilizada en los modelos de predicción según lo estudiado en literatura previa y los resultados obtenidos en este trabajo. Se observa que todos los casos modelados presentan al menos una variable de flujo y todas son significativas al 95% de nivel de confianza, además presentaron los valores de test Z más altos de cada modelo. La variable flujo total en la intersección aparece en el modelo de base completa para accidentes severos. La variable flujo de la rama principal se encuentra en ocho de los nueve modelos, exceptuándose solo el modelo de base completa para accidentes severos. El flujo de la rama secundaria resulta significativo en los modelos de accidentes totales para base completa y en todos los modelos para Cluster 2 (totales, severos y leves). Otra variable de flujo que resulto significativa es el flujo del giro derecha que aparece en los modelos de accidentes severos en base completa y Cluster 1, y en el modelo de accidentes leves en base completa. En todos los modelos, las variables de flujo presentan una relación estadísticamente significativa y positiva con los accidentes, esto quiere decir que a mayor flujo se genera una disminución de la seguridad en la intersección y por ende aumenta la frecuencia de accidentes de tránsito. Este resultado se encuentra en conformidad con lo observado en literatura previa de Guo *et al.* (2010), SECTRA (2014), Chen y Xie (2016), Huang *et al.* (2017), Dufeu (2019) e Islam *et al.* (2019), donde los resultados indican factores de aumento de accidente a medida que el flujo es mayor. Lo anterior se debe al hecho de que a mayor volumen de vehículos en la vía habrá mayor interacción entre vehículos y otros modos de transporte, por lo tanto, hay mayor riesgo de conflicto y colisión entre ellos ya que las maniobras de adelantamiento, cambio de carril o viraje, que puede realizar el usuario, tendrán mayor dificultad de realizarse por la presencia de otros modos en la vía. El flujo de giro derecha principalmente genera accidentes debido al conflicto generado con peatones o ciclistas que cruzan la rama a la cual el vehículo desea virar. A partir de los datos de la Tabla 4.12 se observó que por cada aumento de 10% en el flujo, el flujo total de la intersección genero un aumento de los accidentes estimados de 4.5%, el flujo de la rama principal generó aumentos entre 4.8% a 5.8%, el flujo de la rama secundaria generó aumentos entre 4.0% a 6.2% y el flujo de giro derecha generó aumentos entre 6.7% a 23.2%. Además, cabe destacar que el flujo de la rama principal, que se encuentra presente en los modelos de ambos *clusters* para accidentes totales, severos y leves, genera mayor cantidad de accidentes de tránsito en el Cluster 2 que en el Cluster 1, con diferencias de aumento de

accidentes entre ambos *clusters*, por cada 10% de aumento de flujo, de 0.8% para accidentes totales, 0.4% en accidentes severos y 0.7% en accidentes leves. Recordar que el Cluster 2 estaba conformado principalmente por intersecciones de tipo X, por lo cual hay más ramas de acceso que en el Cluster 1, conformado principalmente por intersecciones de tipo T, lo cual explicaría que un aumento en el flujo genere mayores accidentes en el Cluster 2.

En cuanto a la densidad peatonal, variable de tipo operación, esta solo fue encontrada significativa al 90% de nivel de confianza en el modelo de base completa para accidentes severos. Esta presenta una estimación de parámetro positiva, lo que quiere decir que a mayor densidad peatonal habrá mayor frecuencia de accidente de tránsito. Dicho resultado es consistente con los modelos obtenidos en el estudio de Huang *et al.* (2017) y Dufeu (2019), donde indican que la densidad peatonal está asociada positivamente con la generación de accidentes de tránsito. Esto se debe a que la densidad peatonal y aglomeraciones puede generar distracciones en los conductores, además de aumentar el riesgo de atropellos, donde es probable que resulte en un accidente de tipo severo, en especial si la colisión es de un vehículo contra un peatón, pudiéndose llegar a consecuencias fatales. A partir de los datos de la Tabla 4.12 se observó que la el aumento de cada categoría de densidad peatonal generó un aumento de 14.5% en la estimación de frecuencia de accidentes severos para la base completa.

#### 4.5.3. Variables explicativas de entorno

La variable tipo de zona comercial/mixta es una variable de tipo entorno e indica que el uso de suelo del entorno de la intersección es residencial, comercial o mixta (residencial y comercial). Esta variable toma el valor 0 para zona residencial y 1 para zona comercial o mixta. La variable resultó estadísticamente significativa y positiva en los modelos de accidentes totales y leves para base completa y Cluster 2. Dicho resultado indica que las entidades en zonas comerciales o mixtas presentan un mayor riesgo en la ocurrencia de siniestros aumentando la frecuencia de accidentes estimados. Esto se encuentra en conformidad con los resultados del estudio de SECTRA (2014), donde sus modelos de predicción presentan factores positivos para la variable entorno, la cual en dicho trabajo toma el valor 1 en zonas céntricas definidas como áreas comerciales o servicios con concentración de peatones. Lo anterior se debe principalmente a que en las zonas comerciales suele existir mayor flujo vehicular y aglomeraciones de peatones, incluyendo actividades de abordaje o

descenso de pasajeros del transporte público, pudiéndose generar distracciones en el usuario conductor o en transeúntes, lo cual puede terminar en un siniestro de tránsito. En el caso específico del modelo de accidente severo para Cluster 1, la zona comercial genera una disminución de los accidentes severos, lo cual se debe a que, en las zonas comerciales de alto flujo de peatones, generalmente céntricas, los vehículos no suelen ir a gran velocidad, por lo tanto, no se generan accidentes severos. A partir de los datos de la Tabla 4.12 se observó que el emplazamiento de una intersección en una zona de uso comercial o mixto (residencial y comercial) generó un aumento de los accidentes estimados totales y leves entre 21.9% a 42.8%, mientras que en accidentes severos se genera una disminución de los accidentes estimados de 39.9%.

Otra variable de entorno encontrada estadísticamente significativa es el número de paraderos en torno a la intersección. Esta se encuentra en los modelos de accidentes totales y leves en la base completa y Cluster 1, con una correlación positiva con la frecuencia de accidentes, por lo tanto, se concluye que no es significativa en la generación accidentes severos. Este resultado muestra que a mayor cantidad de paraderos mayor es la frecuencia de siniestros de tránsito, lo cual es consecuente con resultados en los estudios de Huang *et al.* (2017) e Islam *et al.* (2019). Lo anterior se debe a que generalmente los paraderos están cercanos a las intersecciones y por lo tanto son intersecciones concurridas en cuanto a peatones. La presencia de paradero puede hacer que el chofer del transporte público realice algún movimiento brusco y cambios de carril para detenerse en el paradero, lo cual puede provocar colisiones con los vehículos y ciclistas que estén circulando. Usualmente ocurren colisiones generadas por la apertura de puertas y la circulación de ciclistas que generalmente es por el lado derecho de la pista. Además, cabe destacar que los buses suelen tapan la visibilidad de los demás usuarios en la vía lo cual puede producir atropellos a peatones que aparecen inoportunamente. Para evitar estos siniestros se recomienda el uso de bahías de bus en los paraderos (sobrecancha de la calzada) para así evitar que la detención del transporte público en un paradero genere un impedimento en la circulación de los demás vehículos o ciclistas. A partir de los datos de la Tabla 4.12 se observó que por cada paradero presente en el área de la intersección se generó aumentos en los accidentes de tránsito entre 10.2% a 12.2%.

# **CAPÍTULO 5**

## **CONCLUSIONES**

## **Capítulo 5. Conclusiones**

El objetivo general de este estudio fue estimar modelos de predicción de frecuencia de accidentes de tránsito en intersecciones urbanas a partir de variables de infraestructura, operación y entorno. Para lograrlo se realizaron los pasos explicados en este trabajo desde el estudio del estado del arte hasta la realización de la modelación de la cantidad de accidentes en las entidades consideradas. La estimación de accidentes se realizó para la base completa de intersecciones y para la subdivisión de la base de datos en dos *clusters* de intersecciones obtenidos de un proceso de agrupación con K-medios. Los modelos se desarrollaron a partir de una regresión Binomial Negativa (NB) en complemento con la metodología Empírico Bayesiana (EB). Dichos modelos se generaron a partir de la base de datos del “Gran Concepción”. Las variables dependientes estimadas fueron los accidentes totales, severos y leves de las entidades para 5 años.

De los resultados de las modelaciones se observó que el método Empírico Bayesiano generó un aumento significativo en los ajustes de todos los modelos de predicción de accidentes, por lo cual se concluye que dicho método es eficaz en mejorar la precisión de los modelos. Por otro lado, la agrupación de las intersecciones en dos *clusters* generó principalmente aumentos en los ajustes solo para el *cluster* con mayor cantidad de datos, mientras que el otro *cluster*, que consideró menor cantidad de intersecciones, no obtuvo buenos ajustes. Por lo tanto, se recomienda para modelación de accidentes la utilización de bases de datos que presenten gran cantidad de entidades y que el número de *cluster* a utilizar sea bajo para evitar obtener *clusters* con poca cantidad de entidades y posteriormente deriven en bajos ajustes de los modelos. Además, lo anterior facilita posteriormente la aplicación de los modelos al reducirse la cantidad de casos. Por otro lado, las modelaciones por *cluster* no generó aumentos tan grandes de ajuste (en el caso del Cluster 2) e incluso resultó en disminuciones del ajuste (en el caso del Cluster 1) respecto a la base completa, por lo tanto y en definitiva la agrupación en dos *clusters* no resultó en una mejora significativa para la modelación. En cuanto al modelo Binomial Negativo utilizado, se confirma su efectividad frente a datos con sobredispersión como es el caso de los accidentes totales y leves, en los cuales se lograron los mejores ajustes del coeficiente de determinación corregido en los modelos, mientras que, en los accidentes severos, cuyo set de datos no presenta sobredispersión, los ajustes fueron bajos.

En cuanto a la interpretación de variables, las variables explicativas de infraestructura que resultaron más significativas en los modelos fueron el número de pistas de la rama secundaria, la presencia de mediana en la rama principal y la presencia de mediana en la rama secundaria. El número de pistas de la rama secundaria resultó en una correlación positiva en la generación de accidentes de tránsito, por lo tanto, se concluye que a mayor cantidad de pistas existe menor seguridad en la intersección. Por otro lado, la presencia de mediana en la rama principal y en la rama secundaria resultaron en una correlación negativa en la generación de accidentes de tránsito, por lo tanto, se concluye que la presencia de mediana es una variable que aumenta la seguridad en la intersección. Las variables explicativas de operación que resultaron más significativas en los modelos fueron el flujo total de la rama principal, flujo total de la rama secundaria y flujo total de giro derecha. Todas ellas resultaron en una correlación positiva en la generación de accidentes de tránsito, por lo tanto, se concluye que a mayor TMD existe menor seguridad en la intersección. La variable explicativa de entorno que resultó más significativa en los modelos fue la presencia de tipo de zona (uso de suelo) comercial o mixta en los alrededores de la intersección. Este resultado principalmente en una correlación positiva en la generación de accidentes de tránsito totales y leves, mientras que en la generación de accidentes severos para el Cluster 1 se generó una disminución de accidentes, por lo tanto, se concluye que las zonas comerciales son intersecciones menos seguras, vulnerables a la ocurrencia de siniestros de tránsito de gravedad leve pero menos vulnerables a accidentes severos.

En cuanto a la aplicación de los modelos estimados para intersecciones del Gran Concepción, se recomienda el uso de ellos por las autoridades de seguridad vial para la predicción de frecuencia de siniestros como una medida de evaluación del nivel de seguridad de las intersecciones urbanas semaforizadas. El estudio también sugiere que es recomendable la presencia de infraestructura como medianas y pistas de viraje exclusivo que aumentan la seguridad de la intersección y disminuyen el riesgo de siniestros. Las intersecciones con flujos altos son más inseguras, teniendo mayor riesgo de accidentes, por lo tanto, es recomendable la implementación de elementos de seguridad, como vallas peatonales y ciclo pistas con separadores, a modo de proteger a los usuarios más vulnerables de la vía (peatones y ciclistas). El flujo de giro derecha resultó principalmente generador de accidentes severos, debido a la posible colisión de vehículos con peatones o ciclistas, por lo tanto, es recomendable la instalación de señaléticas que adviertan al usuario conductor de la

presencia de ciclistas y peatones, o que recomienden la cedencia del paso a alguno de los involucrados.

Entre las limitaciones del trabajo se encontró el bajo ajuste logrado en los modelos para accidentes severos y en los modelos del Cluster 1. Además, hay que considerar que siempre existen aspectos que no están siendo considerados o captados por las variables explicativas, ya que en todos los modelos existe un valor de intercepto distinto de cero, por lo tanto, los fenómenos de accidentes de tránsito modelados no se explicaron en su totalidad por las variables utilizadas.

Existen varios desarrollos futuros posibles de realizar para complementar y mejorar este estudio. Para la modelación específica de los accidentes severos, que presentan datos de baja dispersión y cercanos a cero, es recomendable la utilización de un modelo que funcione mejor en datos sin sobredispersión, como es el caso de los modelos Poisson, y metodologías que mejoren la estimación en presencia de muchos datos iguales a cero, como son los modelos cero inflados. Al momento de desagregar la base de datos, como por ejemplo mediante la utilización de *clusters*, se recomienda tener la precaución de que los *clusters* no presenten baja cantidad de datos, evitando una mala estimación de parámetros. Por otro lado, las variables que no han resultado significativas en los modelos no significan que no sean capaces de explicar el fenómeno de accidente, por lo tanto, es recomendable realizar una redefinición de ellas y probarlas en la modelación. A su vez resultaría interesante la recopilación de mayor información acerca de cada accidente y nuevas variables enfocadas en los usuarios involucrados en los siniestros y las condiciones climáticas al momento de ocurrido el accidente para así lograr captar ese porcentaje de accidentes que no fue explicado por las variables explicativas utilizadas. En ese sentido, se recomienda obtener el registro de los modos involucrados en el siniestro (veh-veh, veh-peatón, veh-ciclista, etc.) con los cuales se podrían modelar nuevos casos de accidentes más específicos. Además, se recomienda la recopilación de variables de interés que sean capaces de captar la responsabilidad del usuario como la velocidad a la que se trasladaban, estado de los usuarios (presencia de alcohol o drogas en sangre, horas de sueño, etc.) y variables climáticas (temperatura ambiente, porcentaje de humedad ambiente, velocidad del viento, presencia de lluvia, etc.) ya que estas podrían influir en la adherencia de los vehículos al pavimento, en la concentración y visibilidad del conductor, siendo factores de riesgo y pudiendo aumentar la ocurrencia de siniestros de tránsito.

**Bibliografía**

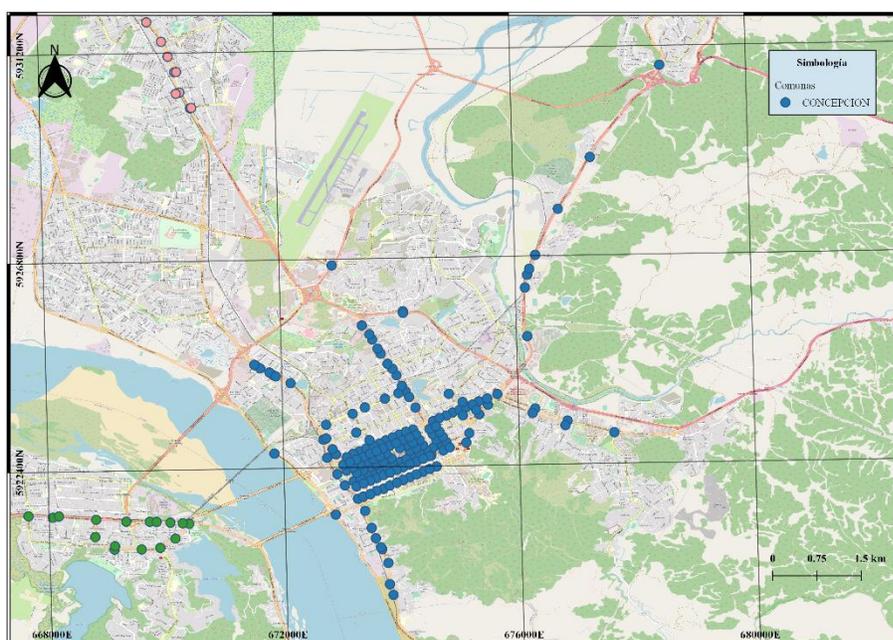
- Alarifí, S., Abdel-Aty, M., y Lee, J. (2018). A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. *Accident Analysis and Prevention*, 263-273.
- Alcaide, M. (2015). *Trabajo Fin de Grado: Modelo de regresión Binomial Negativa*. Sevilla: Universidad de Sevilla.
- Beyer, J. (1 de Abril de 2021). *Congestión vehicular en Concepción*. Obtenido de Blog Investigación: <https://www.ucsc.cl/blogs-academicos/congestion-vehicular-en-concepcion/>
- Camarillo, J., Saavedra, A., y Ramos, C. (2013). Recomendaciones para Seleccionar Índices para la Validación de Modelos. *TecnoLógicas*, 109-122.
- Cameron, C., y Trivedi, P. (1998). Regression analysis of count data. *Cambridge University Press*.
- Chen, C., y Xie, Y. (2016). Modeling the effects of AADT on predicting multiple-vehicle crashes at urban and suburban signalized intersections. *Accident Analysis and Prevention*, 72-83.
- Cheng, W., y Washington, S. (2005). Experimental evaluation of hotspot identification methods. *Accident Analysis and Prevention*, 870-881.
- Chías, L. (2004). *Atlas de la seguridad vial en México*. Ciudad de México: Instituto de Geografía, UNAM, Secretaría de Salud, Consejo Nacional para la Prevención de accidentes.
- Comisión Nacional de Seguridad de Tránsito. (2017). *Libro del Nuevo Conductor Automovilistas*. Santiago: Ministerio de Transportes y Telecomunicaciones.
- Comisión Nacional de Seguridad de Tránsito. (2019). *Evolución de siniestros de tránsito, consecuencias e indicadores (Período 1972-2018)*. Santiago: Ministerio de Transporte y Telecomunicaciones.
- Comisión Nacional de Seguridad de Tránsito. (16 de Diciembre de 2020). *Principios*. Obtenido de <https://www.conaset.cl/principios/>
- Cunto, F., Sobreira, L., y Ferreira, S. (2015). Assessing the Transferability of the Highway Safety Manual Predictive Method for Urban Roads in Fortaleza City, Brazil. *Transportation Engineering*.
- Dufeu Fuller, R. A. (2019). *Modelación de Factores Determinantes de Accidentes en Intersecciones Urbanas*. Concepción: Universidad de Concepción.
- Elvik, R. (2008). Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record*, 72-75.
- Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso*. Barcelona: Universitat Oberta de Catalunya.
- Ghandour, A., Hammoud, H., y Al-Hajj, S. (2020). Analyzing Factors Associated with Fatal Road Crashes: A Machine Learning Approach. *Environmental Research and Public Health*.
- Gobernación Provincia de Concepción. (20 de Febrero de 2021). *Gobernación Concepción*. Obtenido de Conozca la Provincia: <http://www.gobernacionconcepcion.gov.cl/geografia/>
- Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis and Prevention*, 273-285.
- Hao, Y., Xu, L., Qi, B., Wang, T., y Zhao, W. (2019). A Machine Learning Approach for Highway Intersection Risk Caused by Harmful Lane-Changing Behaviors. *ASCE*.
- Hauer, E., Harwood, D., Council, F., y Griffith, M. (2002). Estimating Safety by the Empirical Bayes Method: A tutorial. *Transportation Research Record*.

- Huang, H., Zhou, H., Wang, J., Chang, F., y Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic Methods in Accident Research*, 10-21.
- Islam, R., Barua, S., Akter, S., Hadiuzzaman, y Haque, N. (2019). Impacts of nongeometric attributes on crash prediction at urban signalized intersections of developing countries. *Transportation Safety & Security*.
- JASP Team. (2020). *JASP (Version 0.14.1) [Computer Software]*.
- Kononov, J., Durso, C., Reeves, D., y Allery, B. (2012). Relationship Between Traffic Density, Speed, and Safety and Its Implications for Setting Variable Speed Limits on Freeways. *Transportation Research Record*, Record 2280, 1-9.
- Layfield, R., Summersgill, I., Hall, R., y Chatterjee, K. (1996). Accidents at urban priority crossroads and staggered junctions. *Transport Research Laboratory*.
- Lee, A., Lin, W.-H., Gill, G., y Cheng, W. (2018). An enhanced empirical bayesian method for identifying road hotspots and predicting number of crashes. *Journal of Transportation Safety & Security*.
- Li, P., Abdel-Aty, M., Cai, Q., y Yuan, C. (2020). The application of novel connected vehicles emulated data on real-time crash potential prediction for arterials. *Accident Analysis and Prevention*.
- Lin, C., Wu, D., Liu, H., Xia, X., y Bhattarai, N. (2020). Factor Identification and Prediction for Teen Driver, Crash Severity Using Machine Learning: A Case Study. *Applied Sciences*.
- Lord, D., y Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research*, Part A 44, 291-305.
- Mafi, S., AbdelRazig, Y., y Doczy, R. (2018). Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. *Transportation Research Record*, 1-13.
- Miaou, S.-P., y Lord, D. (2003). Modeling Traffic Crash-Flow Relationships for Intersections. *Transportation Research Record*.
- Mohammadian, A., Sadri, A., Nadimi, N., y Shams, M. (2020). A comparison between Artificial Neural Network and Hybrid Intelligent Genetic Algorithm in predicting the severity of fixed object crashes among elderly drivers . *Accident Analysis and Prevention*.
- Organización Mundial de la Salud. (7 de Diciembre de 2018). *Accidentes de Tránsito*. Obtenido de <https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries>
- Pooyan, A., Haque, M., y Washington, S. (2020). Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis and Prevention*.
- Rizaldi, A., Dixit, V., Pande, A., y Adelwin, R. (2017). Predicting casualty-accident count by highway design standards compliance. *Transportation Science and Technology*.
- RStudio Team. (2020). *RStudio: Integrated Development for R*. Boston, MA: RStudio, PBC.
- Sawalha, Z., y Sayed, T. (2006). Traffic accident modeling: some statistical issues. *NRC Research Press*.
- Subsecretaría de Transportes. (2014). *Metodología para la evaluación social de la reducción de accidentes urbanos*. Santiago: Ministerio de Transporte y Telecomunicaciones.
- Tasic, I., Porter, R., y Brewer, S. (2016). Applications of Generalized Additive and Bayesian Hierarchical Models for Areal Safety Analysis. *Transportation Research Board*, Record 2601, 99-109.
- Wagstaff, K., Cardie, C., Rogers, S., y Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 577-584.

- Wang, J.-H., Abdel-Aty, M., y Wang, L. (2017). Examination of the reliability of the crash modification factors using empirical Bayes method with resampling technique. *Accident Analysis and Prevention*, 96-105.
- Wang, K., Ivan, J., Ravishanker, N., y Jackson, E. (2017). Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis and Prevention*, 6-19.
- Yuan, J., Abdel-Aty, M., Gong, Y., y Cai, Q. (2019). Real-Time Crash Risk Prediction using Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record*.

Anexos**Anexo A. Zona de estudio**

**Figura A. 1.** Mapa emplazamiento de intersecciones en comuna de Chiguayante.



**Figura A. 2.** Mapa emplazamiento de intersecciones en comuna de Concepción.

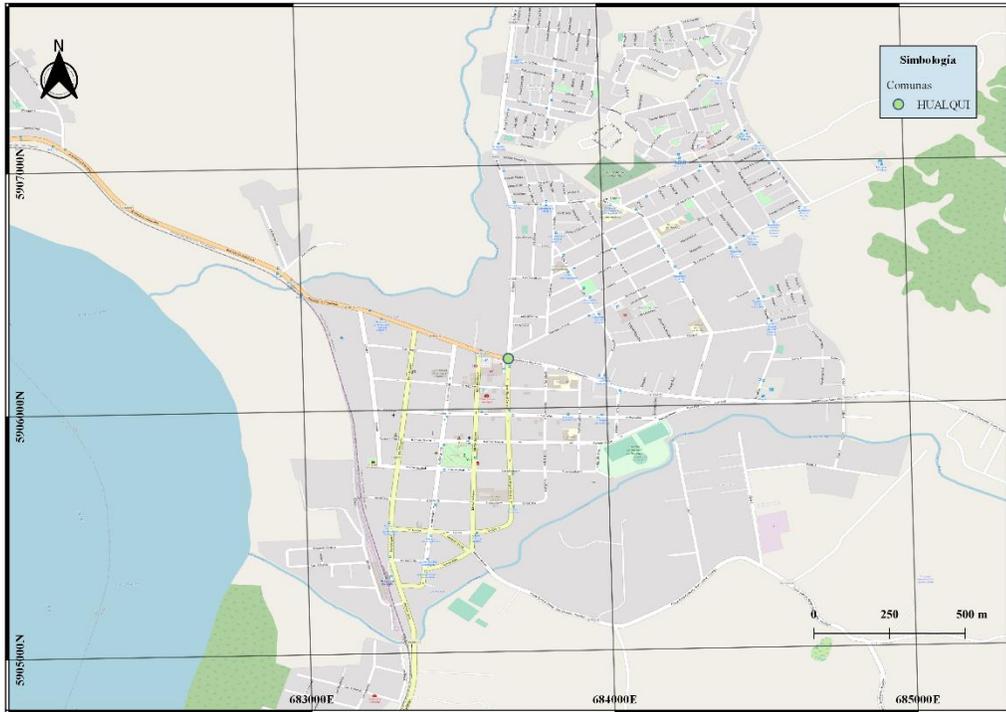


Figura A. 3. Mapa emplazamiento de intersecciones en comuna de Hualqui.

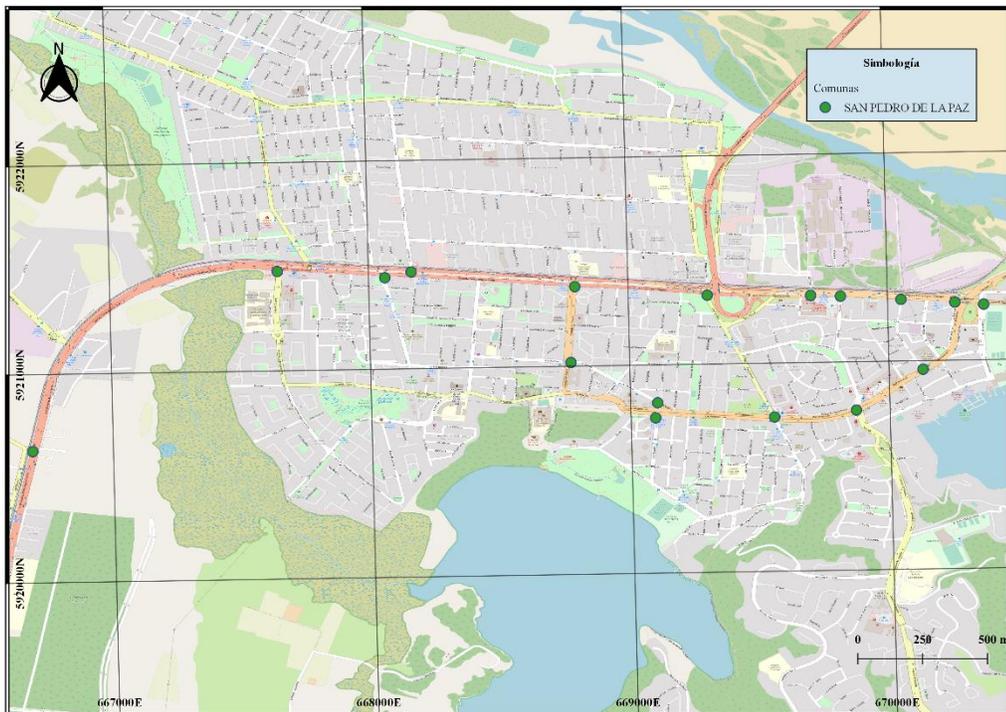


Figura A. 4. Mapa emplazamiento de intersecciones en comuna de San Pedro de la Paz.



Figura A. 5. Mapa emplazamiento de intersecciones en comuna de Talcahuano.



Figura A. 6. Mapa emplazamiento de intersecciones en comuna de Tomé.

## Anexo B. Base de datos

**Tabla B. 1.** Variables explicativas de infraestructura presentes en BD completa.

Variable	Valor	Descripción
TIPO INTER	0	Intersección en T
	1	Intersección en X
PISTA EXCLUSIVA GIRO DERECHA RP	0	No existe pista exclusiva de giro a derecha en rama principal
	1	Existe al menos una pista exclusiva de giro a derecha en rama principal
PISTA EXCLUSIVA GIRO DERECHA RS	0	No existe pista exclusiva de giro a derecha en rama secundaria
	1	Existe al menos una pista exclusiva de giro a derecha en rama secundaria
PISTA EXCLUSIVA GIRO IZQUIERDA RP	0	No existe pista exclusiva de giro a izquierda en rama principal
	1	Existe al menos una pista exclusiva de giro a izquierda en rama principal
PISTA EXCLUSIVA GIRO IZQUIERDA RS	0	No existe pista exclusiva de giro a izquierda en rama secundaria
	1	Existe al menos una pista exclusiva de giro a izquierda en rama secundaria
Nº RAMAS ACCESO	1 - 4	Número de ramas de acceso a intersección
Nº PISTAS RP	2 - 8	Número total de pistas de la rama principal en ambos sentidos
Nº PISTAS RS	0 - 6	Número total de pistas de la rama secundaria en ambos sentidos
PRESENCIA MEDIANA RP	0	No existe mediana en la rama principal o es menor a 0.8 [m] de ancho
	1	Existe mediana en la rama principal y tiene al menos 0.8 [m] de ancho
PRESENCIA MEDIANA RS	0	No existe mediana en la rama secundaria o es menor a 0.8 [m] de ancho
	1	Existe mediana en la rama secundaria y tiene al menos 0.8 [m] de ancho
VALLA PEATONAL	0	No existe valla peatonal
	1	Existe valla peatonal al menos en alguna acera
ANGULO INTERSECCIÓN	0° - 90°	Ángulo de la intersección formado entre ambos ejes de las calles que conforman la intersección
Nº CALLES UNIDIRECCIONALES	0 - 2	Numero de calles unidireccionales en la intersección (sentido único)

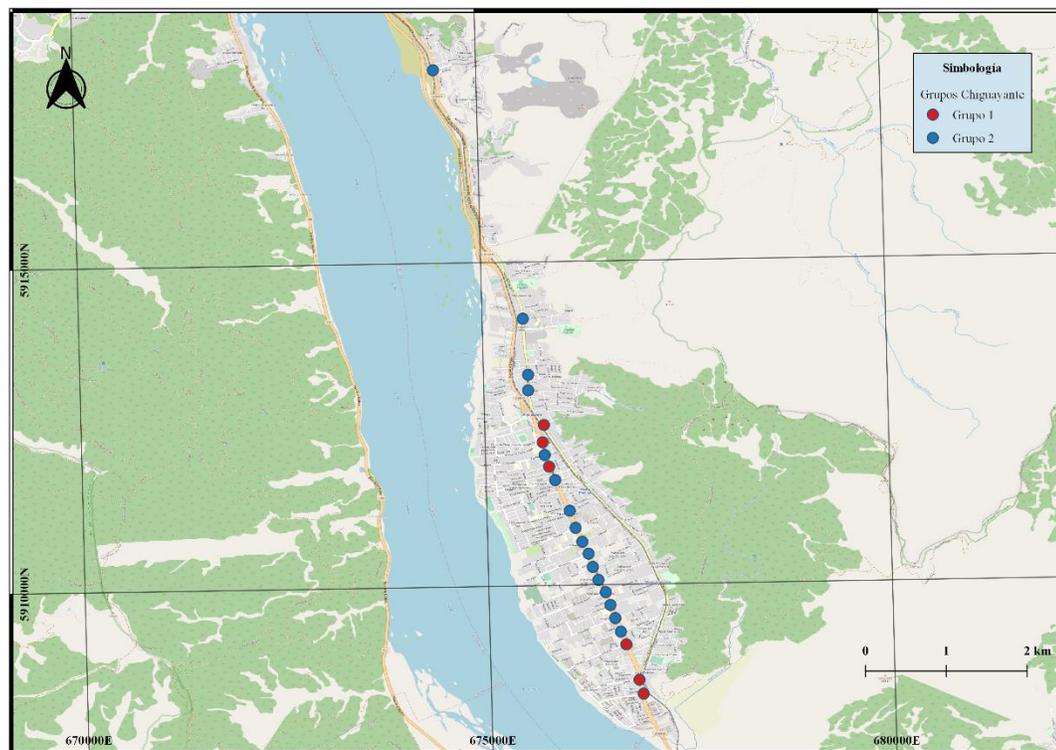
**Tabla B. 2.** Variables explicativas de operación presentes en BD completa.

Variable	Valor	Descripción
GIRO A LA IZQUIERDA RP	0	No existen giros a la izquierda en la rama principal
	1	Existe al menos un giro a la izquierda en la rama principal
GIRO A LA IZQUIERDA RS	0	No existen giros a la izquierda en la rama secundaria
	1	Existe al menos un giro a la izquierda en la rama secundaria
DENSIDAD PEATONAL	1	Densidad peatonal baja en la intersección
	2	Densidad peatonal media en la intersección
	3	Densidad peatonal alta en la intersección
N° CRUCES PEATONALES	0 - 8	Número de cruces peatonales sin considerar las medianas
N° CRUCES PEATONALES (2)	0 - 11	Número de cruces peatonales considerando las medianas
N° FASES SEMÁFORO	2 - 3	Número de fases del semáforo en la intersección
VISIBILIDAD	4.8 - 200	Distancia de visibilidad, según método MOP (2015)
VISION INTERSECCION	1	Nivel de visión malo, obstáculos importantes
	2	Nivel de visión regular, obstáculos menores
	3	Nivel de visión bueno, obstáculos casi imperceptibles
	4	Nivel de visión excelente, sin obstáculos
TMD INTER COMP	Continuo	Flujo total en la intersección completa [veq/día]
TMD RP COMP	Continuo	Flujo total de la rama principal completa [veq/día]
TMD RS COMP	Continuo	Flujo total de la rama secundaria completa [veq/día]
TMD GI COMP	Continuo	Flujo total que gira a izquierda en la intersección completa [veq/día]
TMD GD COMP	Continuo	Flujo total que gira a derecha en la intersección completa [veq/día]
TMD OP COMP	Continuo	Flujo total que se opone en la intersección completa [veq/día]
TMD VL COMP	Continuo	Flujo total de vehículos livianos en la intersección completa [veq/día]
TMD TXC COMP	Continuo	Flujo total de taxi colectivos en la intersección completa [veq/día]
TMD TXB COMP	Continuo	Flujo total de taxi buses en la intersección completa [veq/día]
TMD CAM COMP	Continuo	Flujo total de camiones en la intersección completa [veq/día]
TMD BUS COMP	Continuo	Flujo total de buses inter urbanos en la intersección completa [veq/día]
TMD NOMOTOR COMP	Continuo	Flujo total no motorizado en la intersección completa [veq/día]

**Tabla B. 3.** Variables explicativas de entorno presentes en BD completa.

Variable	Valor	Descripción
ZONA COM/MIX	0	Zona principalmente del tipo residencial
	1	Zona principalmente del tipo comercial o mixta (residencial y comercial)
N° PARADEROS	0 - 3	Numero de paraderos en radio de 50 [m] en torno al centro de la intersección
N° CENTROS DE EDUCACION	0 - 3	Número de centros de educación en radio de 130 [m] en torno al centro de la intersección
N° CENTROS DE SALUD	0 - 1	Número de centros de salud en radio de 130 [m] en torno al centro de la intersección

### Anexo C. Agrupación de intersecciones

**Figura C. 1.** Mapa emplazamiento de intersecciones por *cluster* en Chiguayante.

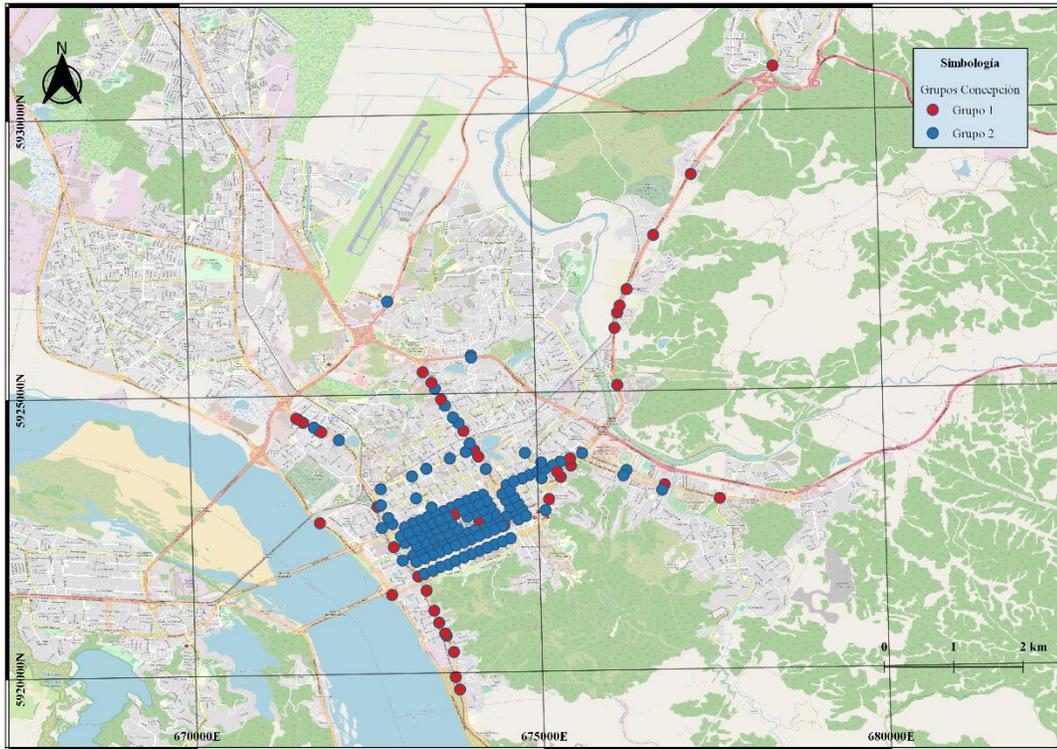


Figura C. 2. Mapa emplazamiento de intersecciones por *cluster* en Concepción.

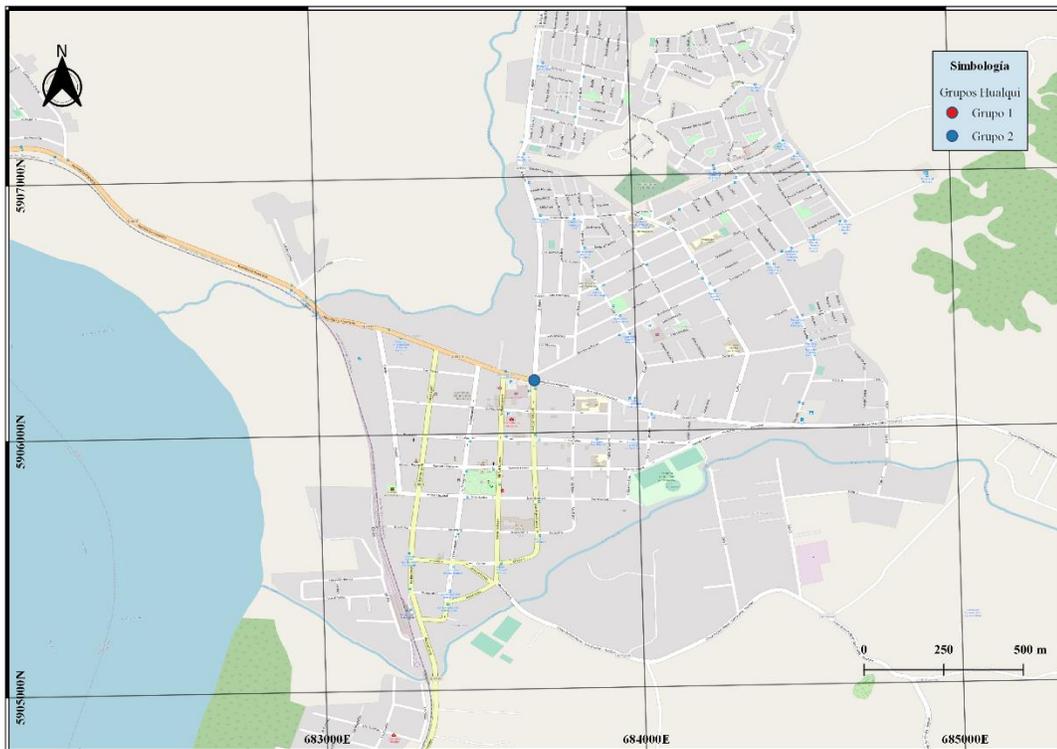


Figura C. 3. Mapa emplazamiento de intersecciones por *cluster* en Hualqui.

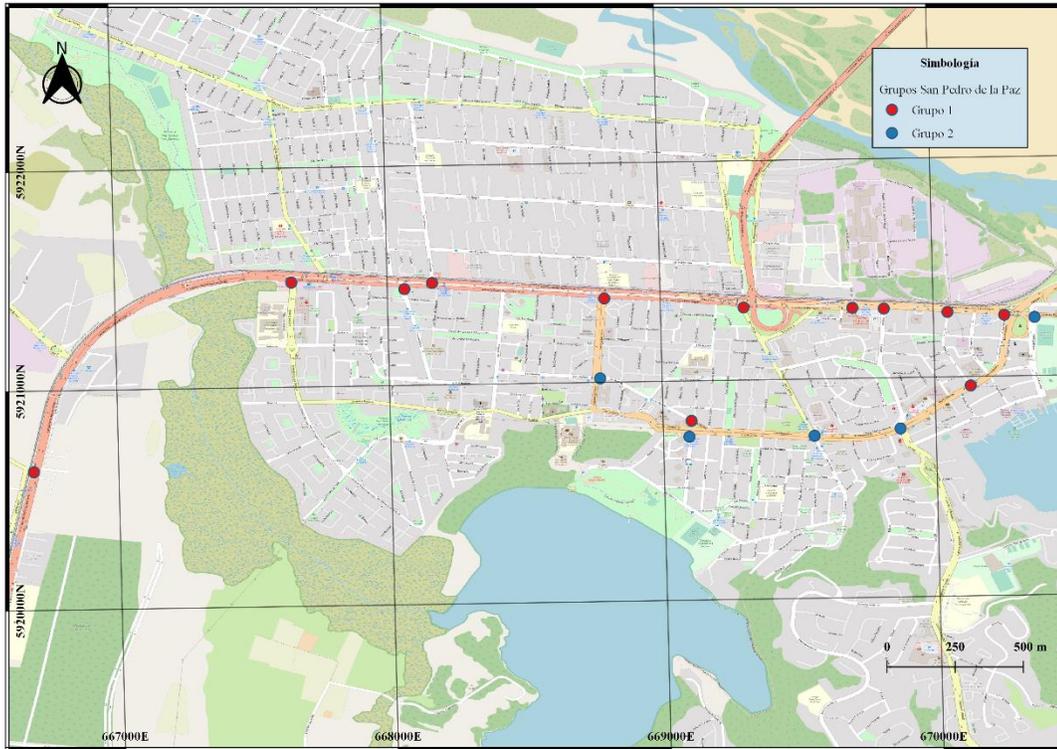


Figura C. 4. Mapa emplazamiento de intersecciones por *cluster* en San Pedro de la Paz.

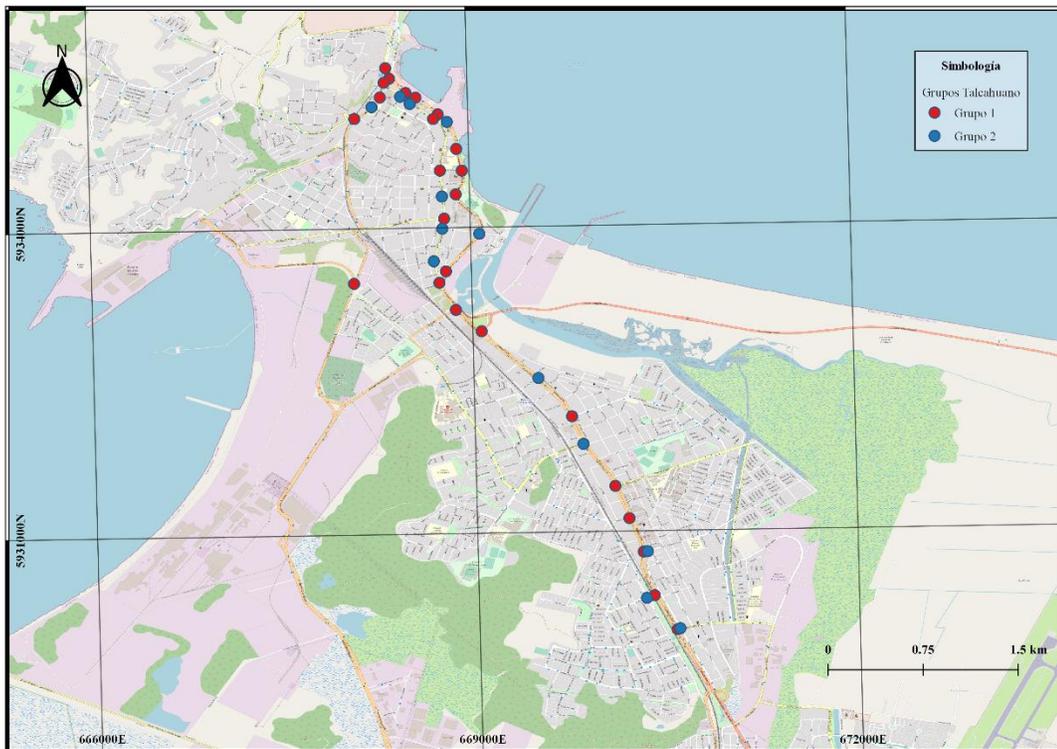


Figura C. 5. Mapa emplazamiento de intersecciones por *cluster* en Talcahuano.



Figura C. 6. Mapa emplazamiento de intersecciones por *cluster* en Tomé.