REVISTA
SINERGIAS

Publicación Semestral de la Secretaría de Posgrado de la Universidad Nacional Guillermo Brown, en colaboración con UNaB Editora.

NÚMERO #01
JULIO DE 2024

Métodos computacionales aplicados a la Política Pública en Argentina

sinergias.unab.edu.ar

Caracterización de la fatalidad vial en NEA a partir de modelos de Machine Learning

De Brián Covaro

Brián Covaro

Caracterización de la fatalidad vial en NEA a partir de modelos de Machine Learning

Introducción

El objetivo principal de este trabajo es explorar y analizar las particularidades de la fatalidad vial en NEA[1] para el período 2019 – 2020. A partir de datos oficiales sobre siniestros viales de este bienio, se construyeron distintos modelos de Machine Learning, basados en algoritmos distintos, que buscaron caracterizar y contribuir a explicar los patrones que presentan los siniestros viales tipificados como fatales. Entre 2008 y 2020, en promedio, 5.500 personas pierden la vida en siniestros viales en Argentina. Es la principal causa de muerte en rango etario de 15 a 35 años y tercera causa de muerte por "causas externas".

La elección del período bajo estudio (2019 -2020) responde a la intención de trabajar con los datos a "año completo", lo más actualizados posible que posee el sistema de registro (SIGISVI)[2]. La elección de la región (NEA), se basa en, primero, la disponibilidad de los registros en dicho sistema; y segundo, en que NEA es la región con el mayor nivel de siniestralidad y mortalidad del período bajo estudio.[3] La preparación de datos y la aplicación de modelos se realizó enteramente con el software RStudio y QGIS.

[1] Lamentablemente, este trabajo no puede contar con los datos de siniestralidad vial de la provincia de Misiones por problemas en la carga de los mismos por parte de la jurisdicción.

[2] Sistema Integral de Gestión de la Información de Seguridad Vial. El SIGISVI es el sistema de carga y registro de siniestros viales que posee la Agencia Nacional de Seguridad Vial (ANSV, en adelante). Este sistema es ofrecido a las jurisdicciones para el correcto tratamiento de la información en este ámbito. A la fecha, 16 jurisdicciones cargan sus datos en este sistema.

[3] ANSV. Observatorio Vial. 2019. Anuario 2019. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_anuario_estadistico_2019_final.pdf.
ANSV. Observatorio Vial. 2020. Informe Anual 2020. Datos preliminares. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_informe_anual_2020_al_4_de_agosto_2021.pdf

Tratamiento de la información y preparación de los datos

Tal como se expuso, la información y los datos con los que se trabaja son los registros de siniestros viales del último bienio de las provincias de Corrientes, Formosa y Chaco.

La unidad de análisis es el siniestro vial. Un siniestro de tránsito es un suceso que ocurre cuando un vehículo entra en contacto con otro vehículo, peatón, animal u otra obstrucción estacionaria, como un poste, un edificio, un árbol, entre otros, en la vía pública.

El objeto de estudio es la fatalidad en los siniestros. Un siniestro vial es tipificado como fatal cuando a partir de ese evento, una o más personas resultan fallecidas (víctimas fatales). Una víctima fatal por siniestro de tránsito es aquella persona que fallece de inmediato o dentro de los 30 días posteriores al hecho, como consecuencia de un traumatismo causado por el siniestro vial (se exceptúan los suicidios).

Figura 1. Tipificación de Siniestros Viales

Tipo de siniestro	Estado	Detalle
Siniestro Simple	Ileso	Persona sin traumatismo alguno
Siniestro con lesionados	Herido Leve/ Herido Grave	Leve: Persona con al menos un traumatismo que exige atención médica mínima o nula (como esguinces, hematomas, heridas superficiales y rasguños) Grave: Persona con al menos un traumatismo que exige la hospitalización durante al menos 24 horas o una atención especializada, como fracturas, conmoción, shock grave y laceraciones importantes.
Siniestro Fatal	Fallecido	Persona muere de inmediato o en un plazo de 24 hs. después del siniestro debido al traumatismo causado por el mismo

Los datos de los siniestros (datos registrados en SIGISVI) se relevan en el FEU (Formulario Estadístico Único). En este formulario se relevan los datos de todo siniestro acaecido en un territorio determinado y con una fecha específica. Los datos que releva tienen el status de datos censales, dado que cada jurisdicción toma al sistema como el único registro oficial.[4]

El FEU releva indicadores de las 3 poblaciones intervinientes en un siniestro vial: siniestros, vehículos y personas. Los siniestros son eventos únicos que tienen asociados vehículo o vehículos y persona o personas intervinientes. Cada población tiene sus atributos definidos que contribuyen a conformar la tipificación posterior del hecho.

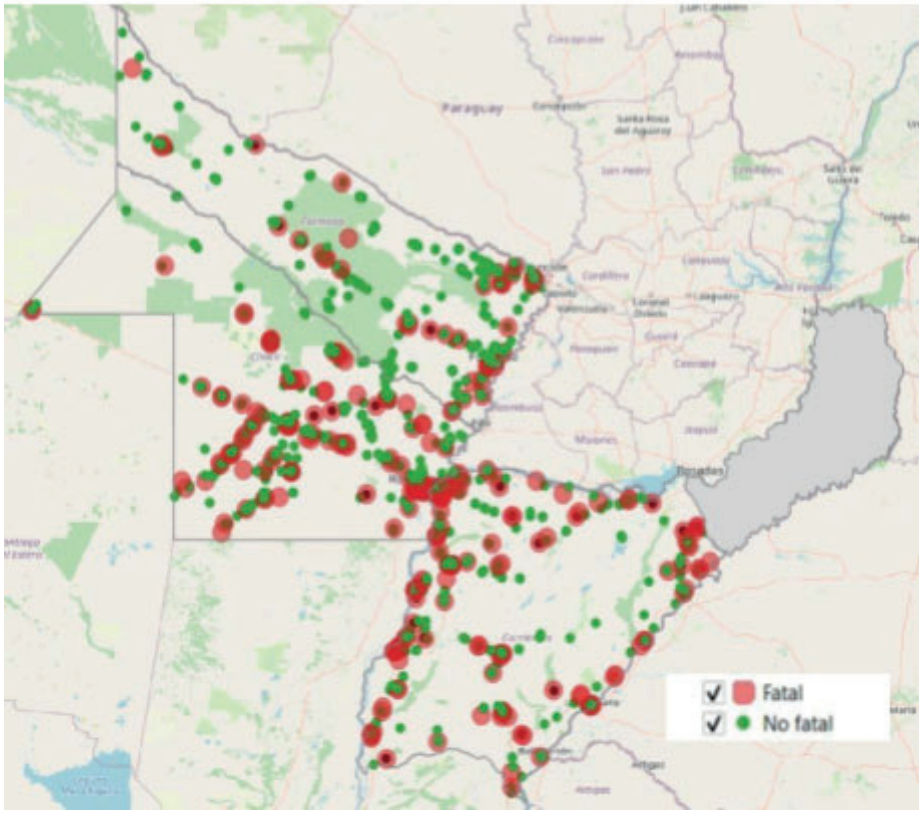
[4] Este atributo de los datos de SIGISVI es importante porque los resultados equivalen a todos los eventos en cada territorio, y no son resultado de ningún diseño muestral. Esto se torna fundamental para diagnósticos espacio temporales como también para la robustez de los estadísticos que se utilicen porque no es necesario poner atención en la significación estadística.

Figura 2. Poblaciones intervinientes en Siniestros Viales

Población	Siniestros	Vehículos	Personas
Indicadores de:	Cuándo y en qué contextos ocurren los siniestros viales. Estados de las vías donde ocurren	Detalle del parque automotor involucrado en siniestros	Perfil de víctimas fatales y heridos Principales grupos de riesgo de siniestros
Variables que contienen:	• Fecha y hora • Ubicación exacta • Categoría de siniestro • Tipo de siniestro • Tipo y estado de la vía • Estado de la calzada • Clima	• Tipos de vehículos involucrados en el siniestro • Datos de cada vehículo involucrado • Tipo de usuario	• Categoría de la víctima • Género • Edad • Condición de la víctima • Uso de elementos de seguridad

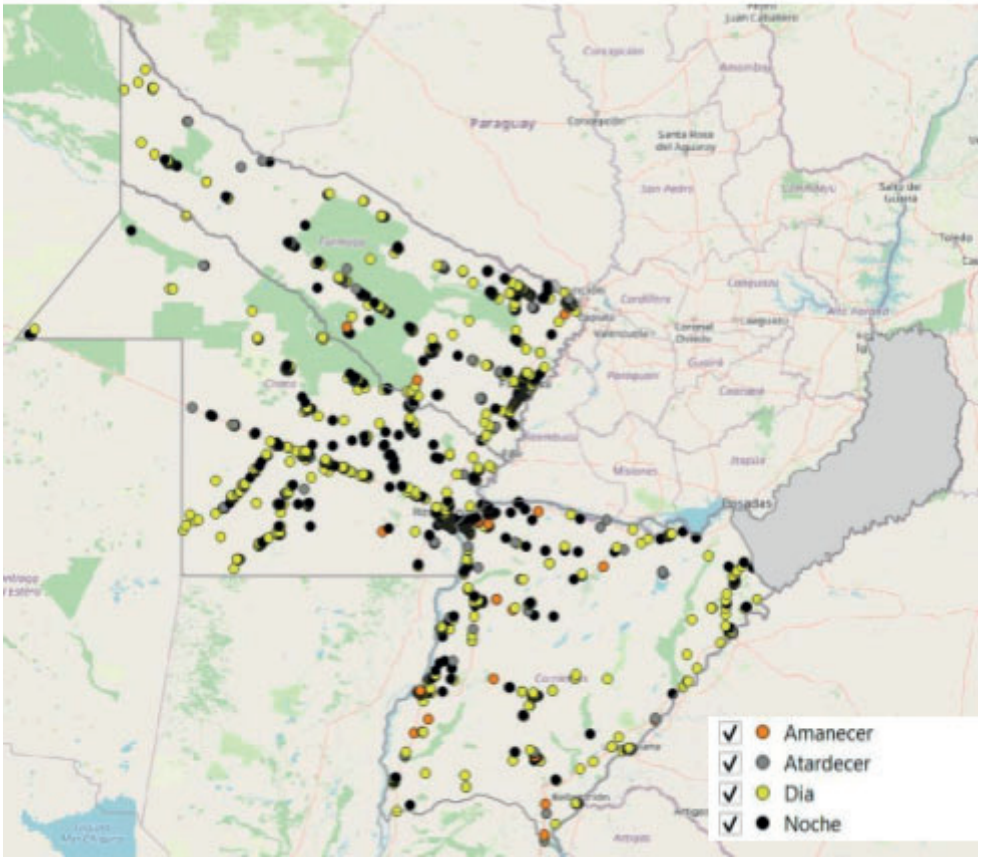
En este sentido, en el FEU se clasifican desde la ubicación geográfica exacta del siniestro, la hora, el tipo, hasta la cantidad y el tipo de vehículos intervinientes, como así también, la o las personas involucradas, con todos sus atributos.

Figura 3. Categoría de Siniestro NEA (SIGISVI - 2019 – 2020)



Las 3 poblaciones asociadas a un evento vial se vinculan mediante claves, como es usual en este tipo de diseños de sistemas de carga. Esto posibilita analizar cada población por separado como también cada siniestro con todos los elementos intervinientes.

Figura 4. Horario del Siniestro NEA (SIGISVI - 2019 – 2020)

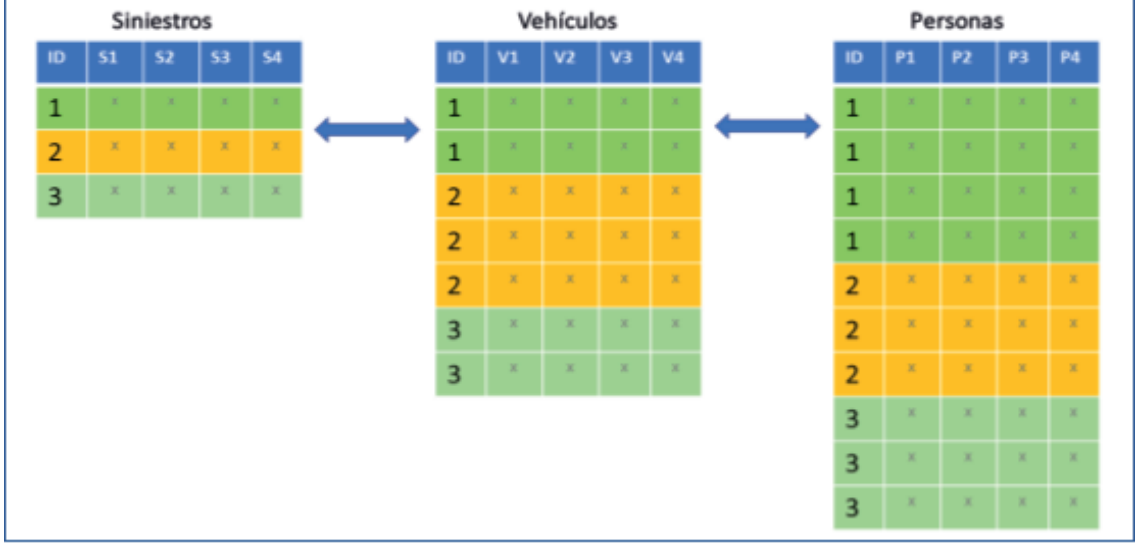


Usualmente, en las estadísticas de siniestralidad vial, los indicadores y variables de las poblaciones que conforman su universo conceptual y empírico, se vinculan en los análisis. De este modo, se puede obtener información sobre qué tipo de siniestros es el preponderante en una zona determinada, o en una época del año u hora del día; como también qué tipo de vehículos se asocian más a siniestros con lesionados, o si algún rango etario se vincula con algunos tipos de siniestros o vehículos.

Estas vinculaciones entre indicadores, además de alinearse a las normativas internacionales de registros de siniestros viales, muestran que los siniestros viales son hechos multicausales, donde intervienen y traccionan multiplicidad de factores. Un siniestro vial puede ser explicado, y parcialmente, por factores ambientales, estructurales, mecánicos, subjetivos, etc.

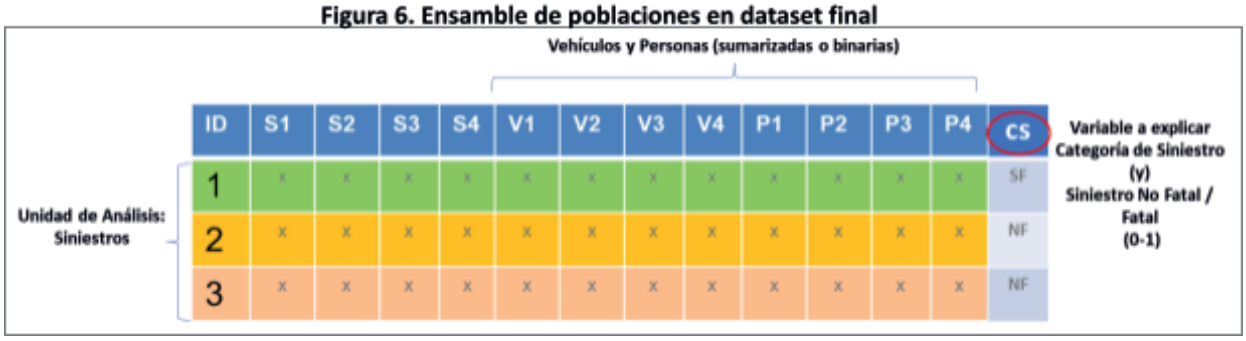
Por esta razón, que da cuenta de la complejidad de estos eventos, al momento de pensar un modelo explicativo de la fatalidad vial, es necesario contar con información asociada de las 3 poblaciones intervinientes.

Figura 5. FEU - Vinculación de las Poblaciones de Siniestro Vial



El proceso de preparación de datos vinculó las 3 poblaciones mediante sus claves, tomando al siniestro como unidad de análisis. Esto último se explica porque lo que buscamos caracterizar es un atributo de los siniestros. La variable objetivo o variable a explicar o *target*, es una variable que es un atributo del siniestro pero que se calcula a partir de un atributo de las personas. El atributo del siniestro (Fatal/ No fatal), deviene del estado de las personas intervinientes en el siniestro[5].

Figura 6. Ensamble de poblaciones en dataset final



Como los datasets tienen jerarquía distinta, al ser el siniestro la unidad de análisis, los indicadores de las personas y vehículos, se sumarizan, en caso de ser variables de intervalos, o se binarizan, en caso de ser variables categóricas. La construcción de la variable a explicar (Siniestro Fatal/ No fatal), se genera a partir del estado fallecido en el lugar del hecho o fallecido post hecho[6]. Al haber una persona (o más de una) con este status, el siniestro es clasificado como Fatal. De esta forma, queda construido el dataset con las variables de las 3 poblaciones para comenzar la exploración y el modelado.

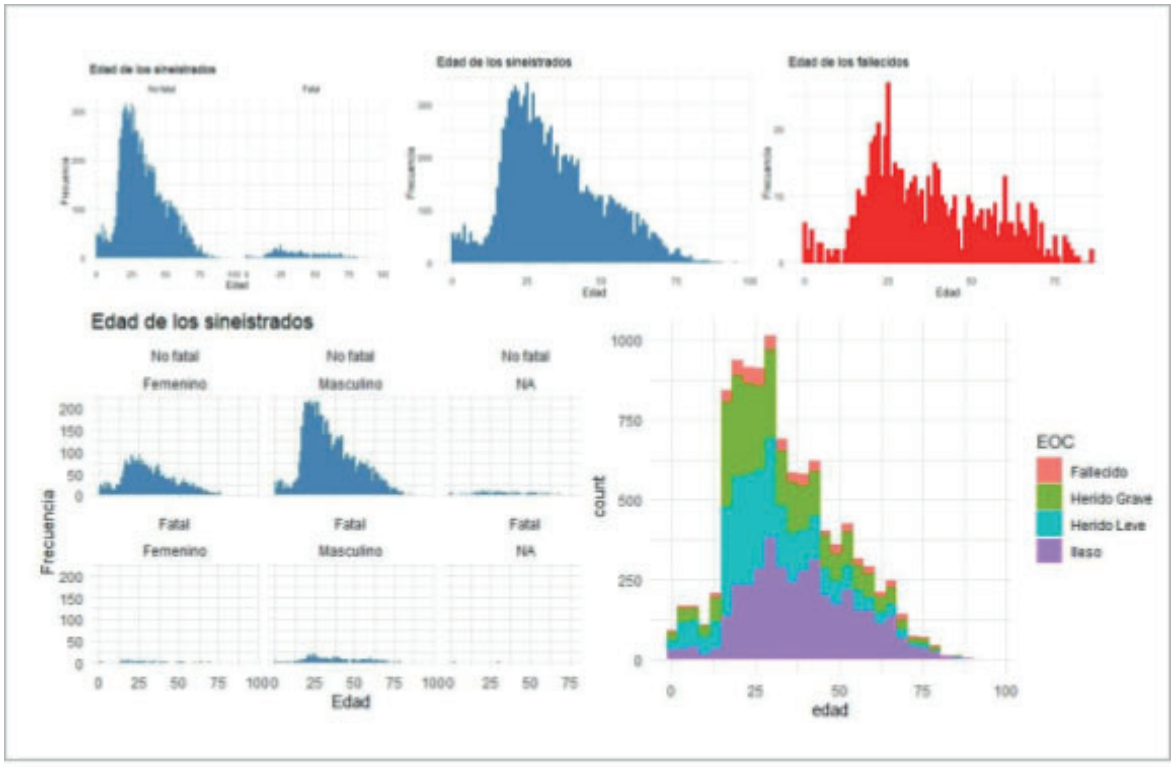
[5] Tal como se expuso en el cuadro N°1, el estado de un interviniente post. Siniestro puede ser ileso, lesionado (leve o grave) o fallecido.

[6] Los fallecidos en seguridad vial son los fallecidos en el lugar del hecho o los fallecidos "a 30 días", es decir, toda persona que haya fallecido por causa de un siniestro vial, desde la fecha del siniestro hasta 30 días. El SIGISVI permite hacer ese seguimiento y calcular este tipo de fallecidos.

Exploración y análisis de los datos

En el bienio 2019 – 2020, en NEA (-M)[7], se registraron 5.514 siniestros viales[8]. De éstos, un 5.9% fueron siniestros fatales. Intervinieron 9.975 vehículos y 12.551 personas, de las cuales fallecieron 714 (5.7%). 6 de cada 10 personas que se vieron involucradas en estos siniestros, tuvieron algún tipo de lesión. [7]

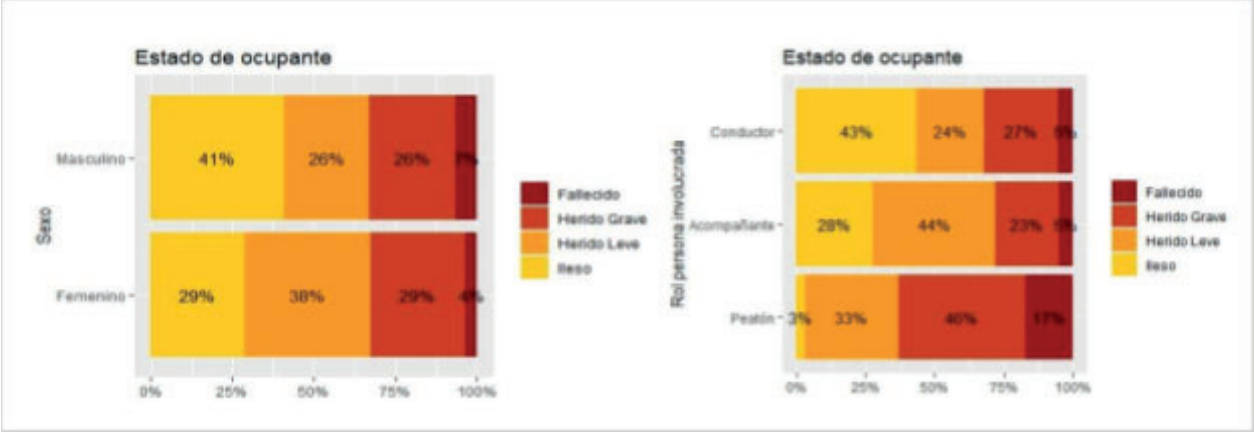
Figura 7. Descripción de las personas involucradas (siniestradas y fallecidas)



La incidencia de género en los siniestros es de 70/30 a favor de los hombres, aumentando a 80/20 en cuento a los fallecimientos. Huelga decir, y ésto se refleja en las estadísticas anuales, que estamos ante un evento predominantemente masculino. El promedio de edad de los siniestrados es de 34 años y de 37 de los fallecidos. El rango de 15 a 35 años, representa el 50% de los fallecidos. Al interior de cada género, los hombres se diferencian en sus niveles sin lesión y en fallecidos. Luego, al interior de los roles, los peatones son los que más intervienen en siniestros donde hay fallecidos.

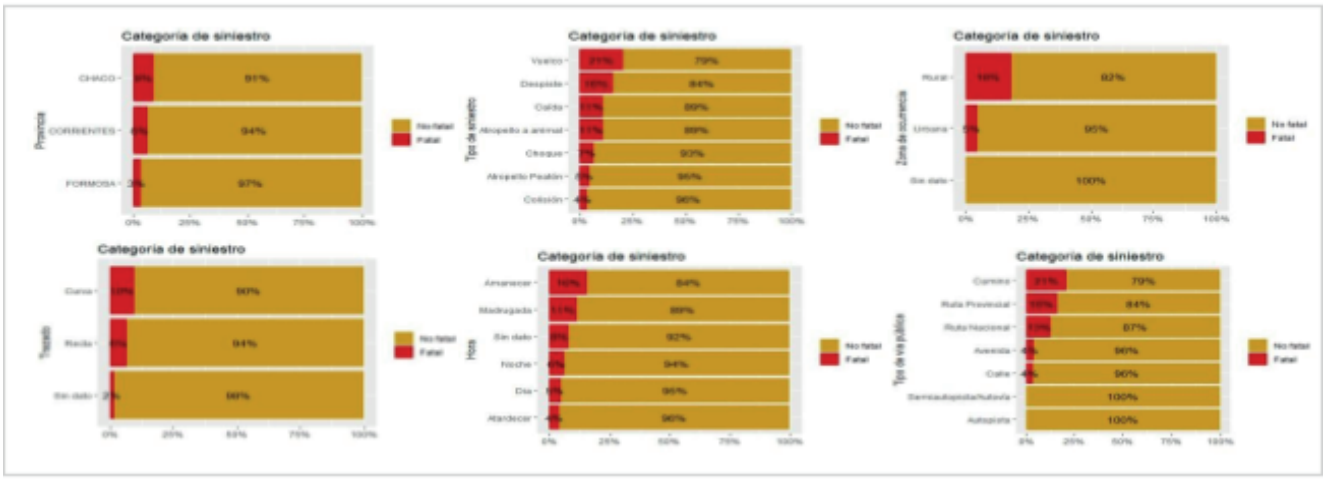
[7] Siglas que explican la ausencia de la provincia de Misiones dentro de lo que se conoce como NEA.
860% ocurrieron en 2020 y 40% en 2019, cosa que es curiosa, dado la baja de circulación que hubo en casi todo 2020 a causa del ASPO

Figura 8. Estado de ocupante según Sexo y Rol



Los siniestros presentan características puntales: Chaco (9%) es la provincia con más incidencia de siniestros fatales, seguido por Corrientes (6%) y Formosa, con un tercio respecto de la primera (3%). La fatalidad es mayor en vuelcos y despistes; en curvas y en ámbitos rurales. Asociado a ésto, se observa que el tipo de vía no es el medio urbano, sino "caminos" o rutas nacionales o provinciales.

Figura 9. Categoría de siniestro según atributos de siniestros



En términos temporales (no se percibe un patrón estacional en los meses del año), la fatalidad se da más en horas del amanecer y madrugada.

El siguiente cuadro muestra el porcentaje de participación de cada tipo de vehículo en siniestros fatales, es decir, qué participación tuvieron cada uno en la fatalidad.[8] Nuevamente hay indicios de ciertos patrones en la fatalidad de los sinestros en esta región. Tres de los cuatro primeros vehículos involucrados, son vehículos no particulares, es decir, de uso privado; son vehículos de uso profesional o laboral (Maquinaria, Transportes de Carga y Transporte de pasajeros). Esto, sumado a la incidencia por el tipo de vía (rutas nacional y provincial, y "caminos"), nos empieza a dar una fotografía de factores propios de la zona (en términos estructurales y productivos) que se asocian con la fatalidad vial.

[8] Es importante comprender que lo que se intenta caracterizar no es la participación en la siniestralidad, sino la incidencia en la fatalidad (participación en siniestros que han resultado fatales). Por lo tanto, lo que buscamos en el análisis de los vehículos involucrados, es la proporción de su participación en siniestros viales fatales, es decir, de todos los siniestros donde han participado, qué proporción fue fatal.

Figura 10. Porcentaje de fatalidad en el tipo de vehículo involucrado	
Maquinaria	16,7%
Transporte de carga	13,4%
Otros	12,5%
Transporte de pasajeros	11,6%
Bicicleta	7,8%
Peatón	6,0%
Motocicleta	5,1%
Camioneta/ Utilitario	4,7%
Automóvil	3,6%

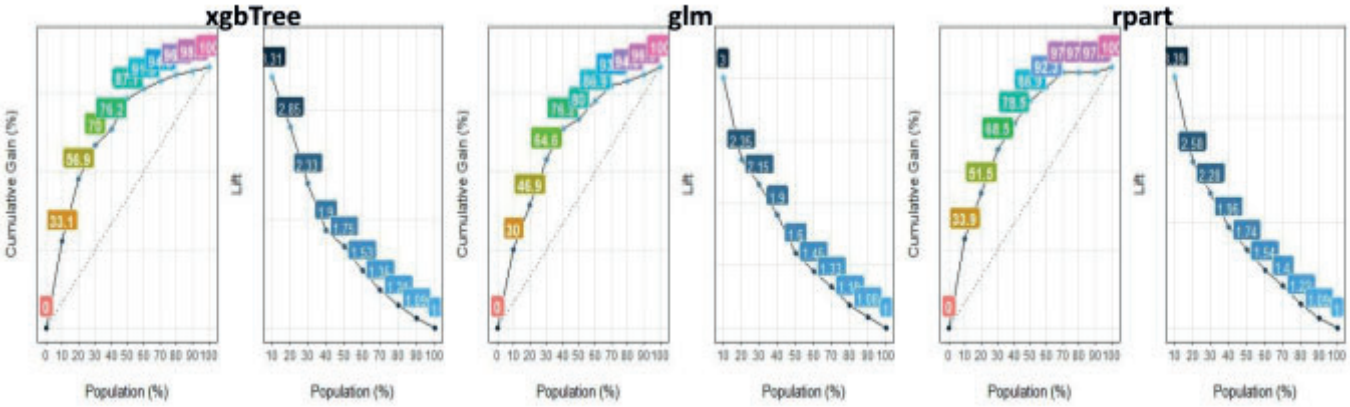
Cuando hablamos de los siniestros viales como eventos multicausales o multifactoriales, estamos hablando de cómo interactúan diversos factores que contribuyen a explicar estos eventos. Esta zona parece tener atributos y características particulares relacionadas no sólo a los ambientes naturales, sino a factores sociales y productivos. Una hipótesis plausible de ser contrastada, podría ser que la siniestralidad en general, y la fatalidad en particular, difiere significativamente (cuantitativa y cualitativamente) a medida que pasamos de una región a otra.

Modelización de la fatalidad vial

Para modelizar la siniestralidad fatal, se pusieron en práctica 3 algoritmos para decidir cuál de ellos desarrollaba la mejor solución ante este problema de investigación y este set de datos. Todos son para procesos basados en técnicas de dependencia o supervisadas. Los algoritmos usados fueron "glm", la función del modelo lineal generaliza, es su opción binomial (logit), "rpart", árbol de regresión y "XGBoost", árboles de decisión basados en el principio de *boosting.*[9]

Para ver el rendimiento de los modelos[10], analizamos las curvas de Ganancia (Gain curve). Como se puede ver, son muy parejas, siendo la curva de xgbTree la que tiene mejor performance (ganancia de 70% en los 30 p./ 87.7% de acumulación de positivos en 50 p. y lift de 1.75).

Figura 10. Curva de Ganancia de los Modelos



[9] La idea principal de esta parte del trabajo no es profundizar en los algoritmos, sino en los resultados. La elección de estos tres tuvo que ver con probar algoritmos distintos o con lógicas de clasificación distintas.

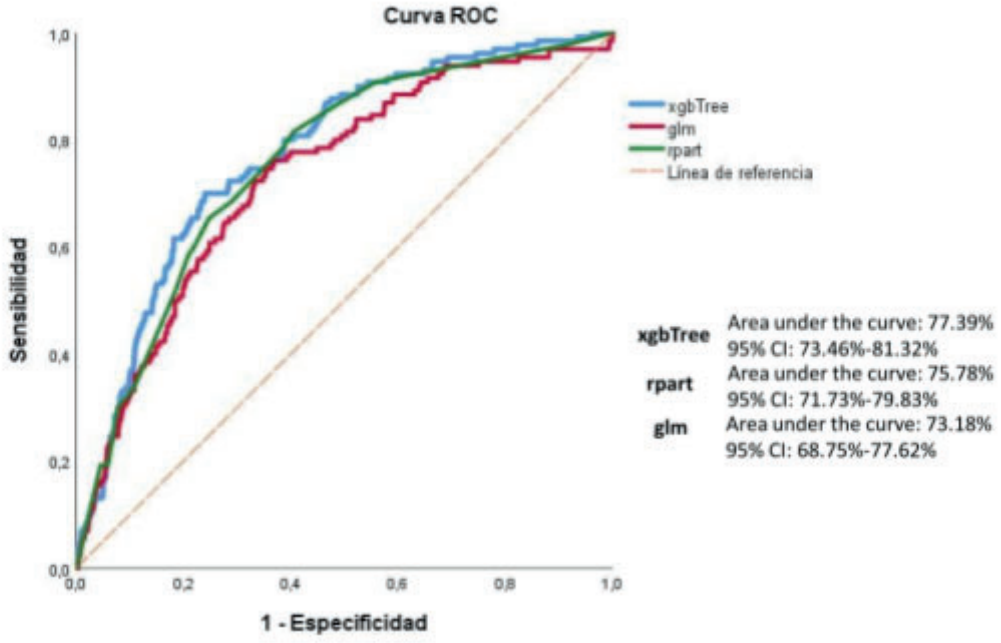
[10] Previamente se dividió el dataset en train/ test (60/40). El análisis posterior de resultados corresponde al dataset de testing.

xgbTree

Population	Gain	Lift	Score.Point
1	10	33.08	3.31	0.76873952
2	20	56.92	2.85	0.63756694
3	30	70.00	2.33	0.51295121
4	40	76.15	1.90	0.42125511
5	50	87.69	1.75	0.33519310
6	60	91.54	1.53	0.27165674
7	70	94.62	1.35	0.21468726
8	80	96.92	1.21	0.16127484
9	90	98.46	1.09	0.10764062
10	100	100.00	1.00	0.03498399

Otra manera de analizar el ajuste de los modelos es comprar su rendimiento mediante la curva ROC, analizando su AUC (Área bajo la curva). La lógica es similar a la curva de Ganancia, en cuanto al análisis de la buena o mala clasificación. La clasificación indica que el modelo es aceptable.

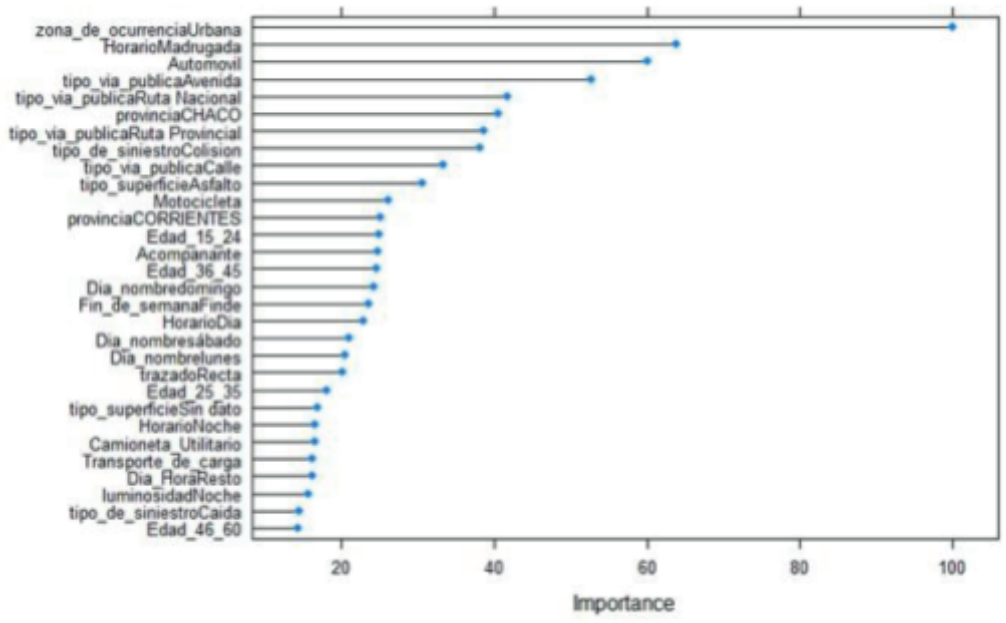
Figura 11. Curva ROC de los Modelos



El algoritmo xgbTree ofrece un gráfico que muestra el aporte (gráfico de importancia) de cada variable al ajuste final del modelo. En este caso, se muestran 30 variables que más aportan a la clasificación de la variable objetivo.[11]

[11] Si bien no deja de ser una síntesis útil, en este tipo de algoritmos basados en el boosteo (*boosting*, o combinación los resultados de varios clasificadores débiles para obtener un clasificador robusto), hay que interpretarlos con prudencia. En este caso, la partición del este tipo de árboles, también trabaja con el peso de cada categoría en cuanto a lo que aporta para la partición de los nodos, es decir, que una categoría con mucha "presencia" (peso), aporte (está más presente) en los nodos, no siendo necesariamente útil para la clasificación de la variable objetivo. En este caso, se puede ver que la categoría "urbana" tiene la mayor "importancia", pero esa importancia es su presencia en los datos y en la partición.

Figura 11. Plot de Importancia de xgbTree



Habiendo resultado aceptable el ajuste final, y como el objetivo principal de este modelo es caracterizar[12] la fatalidad vial, analizamos la distribución de los predictores seleccionados. Como expusimos al principio, las variables seleccionadas (42) para modelizar, corresponden a distintas dimensiones del corpus conceptual de la seguridad vial: estructurales, ambientales, mecánicas, productivas y subjetivas. Para ordenar la lectura de los resultados, dada la cantidad de los mismos, se "ordenaron" según estas dimensiones.

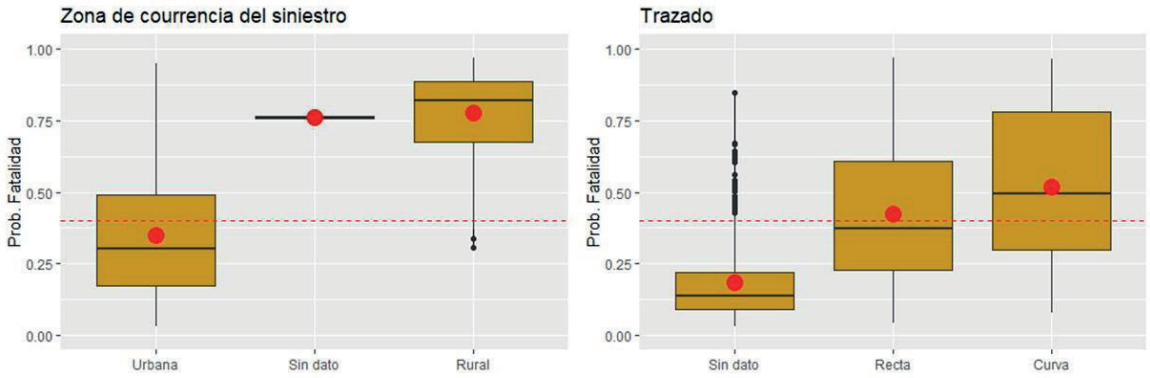
Lo primero que observamos es la preminencia no urbana de la fatalidad. Los gráficos[13] son muy claros en cuanto al promedio de la probabilidad de fatalidad y la agrupación de casos en "rural" como zona de ocurrencia del siniestro[14].

[12] Estos modelos cumplen todos los requisitos para ser modelos predictivos, pero no es el objetivo de este trabajo. Entendemos que lo predictivo en un modelo de estas características, se da al momento de poner en producción al modelo, clasificando datos "nuevos" o datos no supervisados. Esa instancia no es posible en este momento del proceso, por eso optamos por la caracterización de los eventos (siniestros fatales).

[13] Los gráficos tienen el promedio de probabilidad de fatalidad (punto rojo) y los boxplot de la distribución de los casos de las categorías de la variable predictora (la línea negra de cada caja es la media de la distribución). La línea roja punteada, es el promedio de la probabilidad observada del socre.

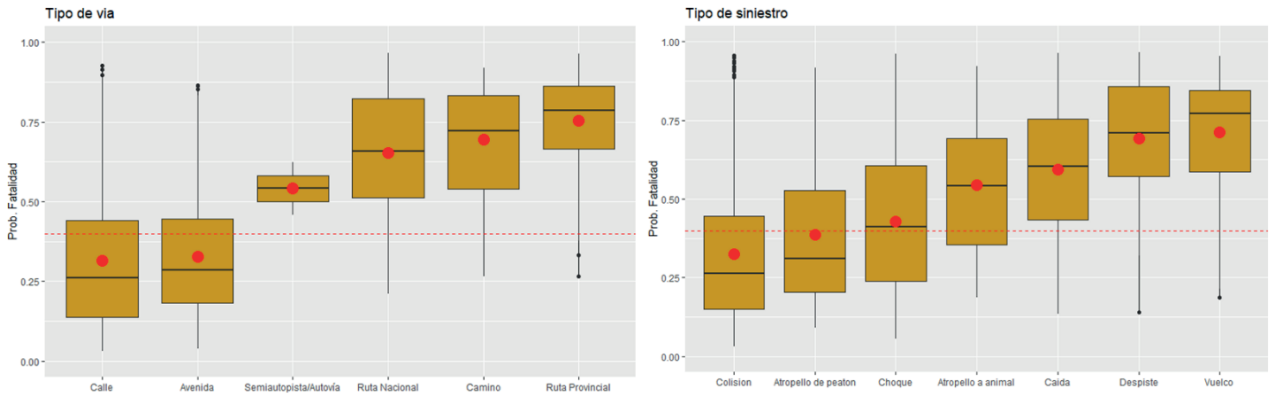
[14] Recordemos que lo que buscamos caracterizar es la fatalidad, no la siniestralidad (91% urbana en NEA). Evidentemente, la siniestralidad es "más fatal" (aunque inmensamente menor en incidencia) en zonas rurales (casi 5 veces más de fatalidad) que urbanas (más siniestralidad). Es una relación entre eventos posibles (fatales) dentro de eventos observables (siniestros).

Figura 12. Probabilidad de fatalidad según zona de ocurrencia y trazado



Relacionado con esto, los tipos de vías con mayor probabilidad de fatalidad son las preeminentes en medios rurales: Rutas nacionales y provinciales y "caminos" rurales. En tipo de siniestro, es muy clara la probabilidad de fatalidad si se dan vuelcos, despistes, caídas o atropellos[15]. La curva es la zona del trazado que se evidencia como más fatal.

Figura 13. Probabilidad de fatalidad según Tipo de vía y siniestro

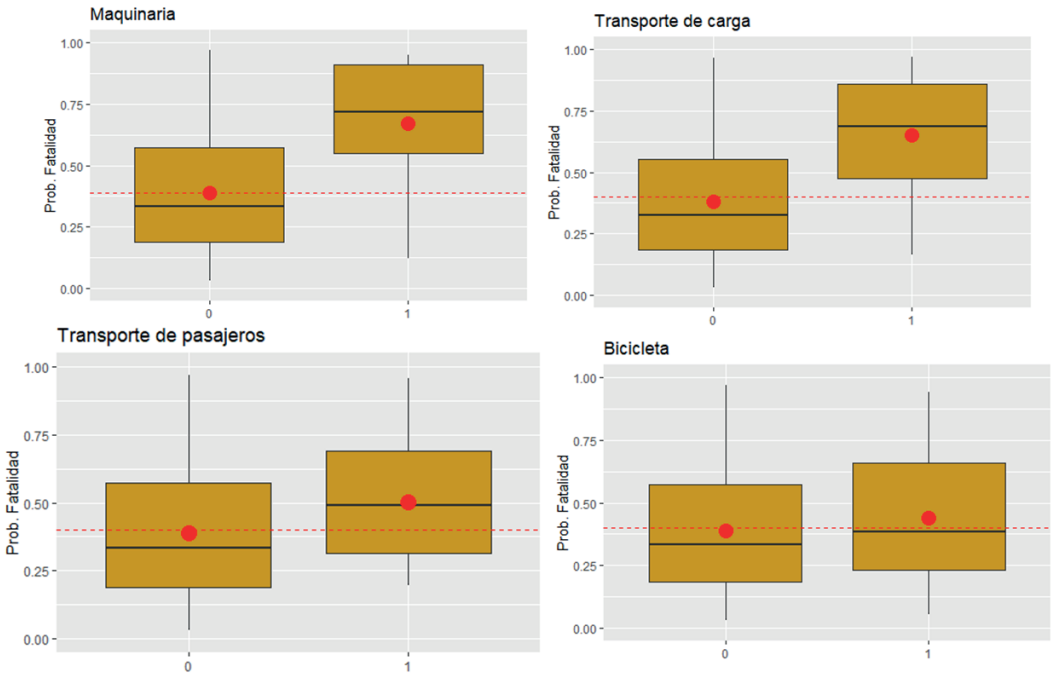


Una parte importante del análisis de la fatalidad vial, es el papel de los vehículos (dimensión mecánica). En este caso, confirmamos la tendencia de la "ruralidad" en la fatalidad en esta región, porque los tipos de vehículos con una probabilidad de fatalidad significativa son: Maquinarias, Transportes de carga, Transporte de pasajeros. Bicicleta es el único tipo de vehículo en que arriesgamos como "urbana" que aparece como significativo en su probabilidad de fatalidad. Los restantes tipos de vehículos (automóvil, motocicletas, camionetas, etc.) no aportaron de forma significativa a la probabilidad de intervenir en un siniestro fatal (similares a la observada o un poco por encima, como motocicletas y automóviles).

También parece haber un patrón temporal en la caracterización de la fatalidad. Principalmente la fatalidad se da los fines de semana, mayormente de noche. De los días, específicamente el domingo. El horario que presenta mayor diferencia a favor de la fatalidad es el amanecer y la madrugada.

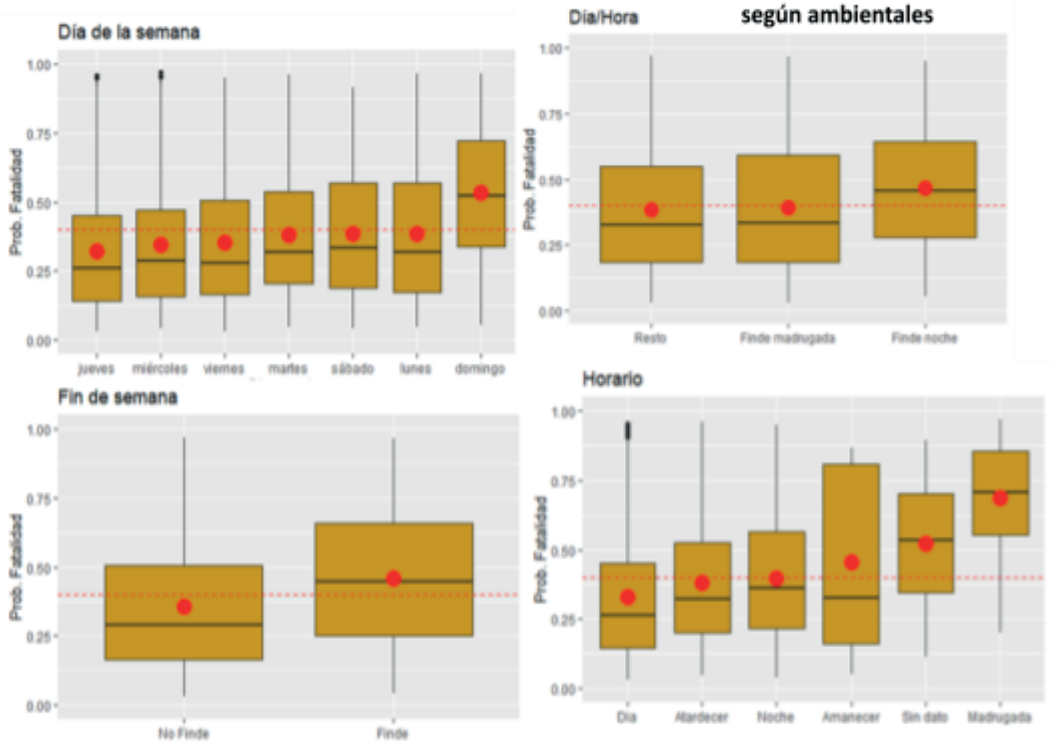
[15] Como se observa, colisión es el tipo de siniestro con mayor incidencia neta, pero el de menor fatalidad relativa.

Figura 14. Probabilidad de fatalidad según atributos de los vehículos



Tenemos evidencia como para ensayar la hermenéutica de estos resultados. La fatalidad vial en NEA (-M), por lo menos en los datos de siniestros viales del bienio 2019 – 2020, parece estar asociada, en mayor medida, al ámbito rural más que al urbano, a los vehículos de porte y de uso profesional, más que a los vehículos particulares. Esto tiene sentido cuando observamos la traza vial de la región, la misma está atravesada por 12 rutas nacionales con un importante TMDA.[17]

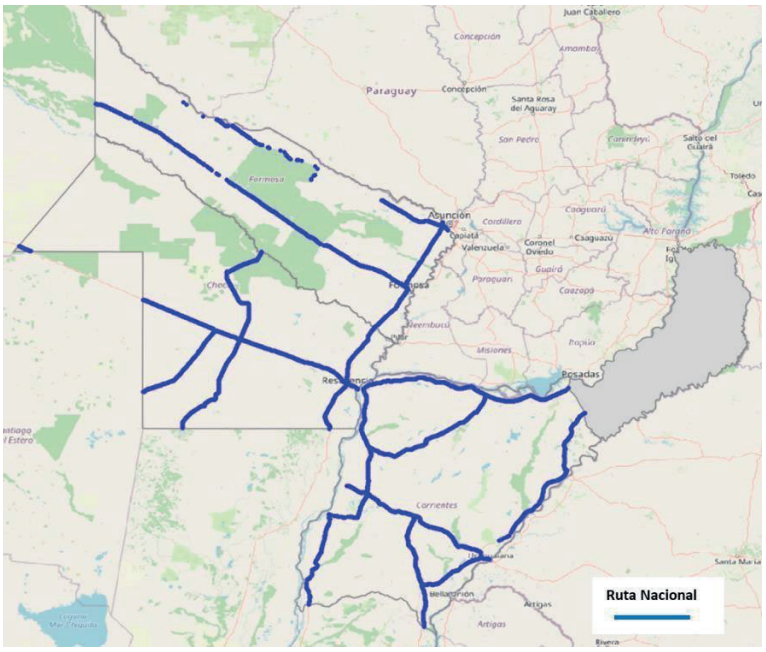
Figura 15. Probabilidad de fatalidad según ambientales



[17] TDMA: Tránsito Diario Medio Anual

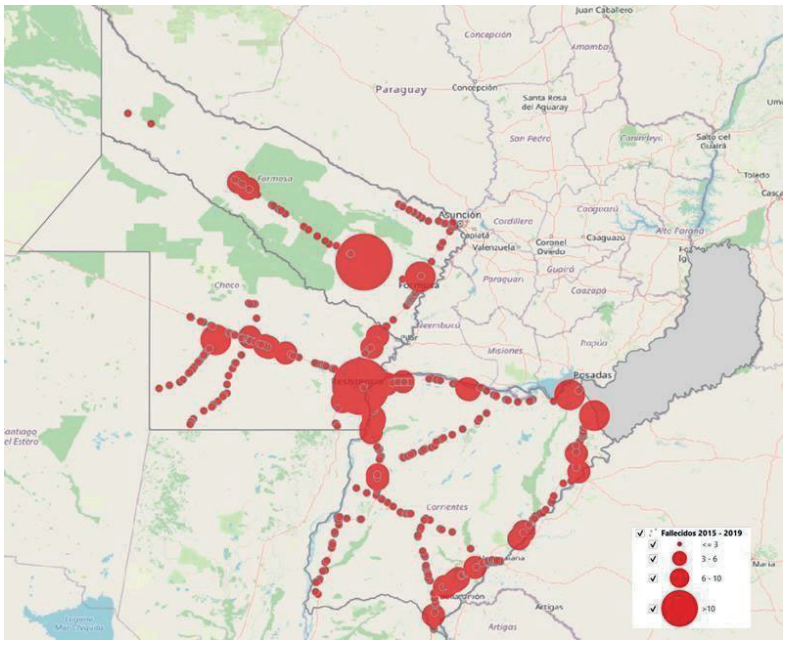
Por lo tanto, estaríamos ante un nuevo patrón de la fatalidad, la asociada al ámbito productivo. La fatalidad vial está asociada a vehículos en actividad en medio de un proceso productivo (trabajo), ya sea maquinaria (suponemos, agrícola) o transporte (de carga o pasajeros).

Figura 16. Rutas nacionales NEA (-M)



Si a esto le sumamos que los siniestros devenidos en fatales se dan en hora de madrugada o amanecer y el tipo de siniestro fatales son vuelcos o despistes, muy probablemente estemos ante siniestros fatales por causa de exceso de tiempo de conducción, es decir por cansancio. Tampoco hay que descartar el factor de visibilidad y el estado de las rutas como causantes de este tipo de siniestros (dimensión estructural).

Figura 17. Fallecidos en Rutas nacionales NEA (-M) (SIAT -2015 -2019)



Consideraciones finales

Hemos explicado y detallado los datos y los resultados del modelo de caracterización de fatalidad en NEA y los asumimos como lógicos y consistentes con los datos y los objetivos planteados. Estimamos que se logró un modelo de caracterización de la fatalidad en esta región que aporta conocimiento sustancial hacia esta problemática. En este cierre, no vamos a repetir los patrones anteriormente detallados y explicado sobre la fatalidad en esta región, la idea es no redundar. Ahora existen indicios por dónde se podría comenzar a trabajar en un área temática (seguridad vial) que no se caracteriza por tener abundancia de estudios y de datos al respecto.

Lo que se podría agregar es lo siguiente:

Primero, lamentamos no haber podido aprovechar al máximo un sistema de gestión de datos como el SIGISVI. Esto se debió no solo por la ausencia de la provincia de Misiones (cuestión que ya explicamos), sino también a que no pudimos contar con indicadores que, estimamos, hubiesen sido de gran utilidad para explicar la fatalidad vial.[16] Esta situación se debe a la cantidad de variables "sin dato" que posee el sistema. No deja de ser un llamado de atención para los encargados de la carga de datos, pero también para reflexionar sobre el diseño del FEU.

Segundo, desde el comienzo de este trabajo, explicamos que el objetivo era la caracterización y no la predicción, por no tener la chance práctica de clasificar datos nuevos. Sin embargo, creemos que este tipo de procesos y de información resultante, pueden aportar a predecir (o prevenir, traduciéndolo a políticas públicas) la fatalidad vial. Este tipo de procesos deberían (aprovechando los datos que se registran) intervenir en algún proceso regular dentro de la ANSV.[17] La ANSV otorga licencias de conducir, intervienen en la reglamentación de vehículos profesionales, es actor interviniente en legislación vial, etc. Este tipo de trabaja con datos, debería complementar y contribuir en alguna de estas tareas, con el fin de mejorar, reforzar la seguridad vial y prevenir siniestros y víctimas.

[16] El sistema cuenta con indicadores como utilización de elementos de protección (caso, cinturón, etc), seguro del vehículo, existencia de señalizaciones, estado de la ruta/calle, presencia de luz artificial, VTV del vehículo, alcoholemia del conductor, color del vehículo, entre otros. Ninguno de éstos se pudo utilizar en el modelo para caracterizar la fatalidad, desaprovechando una fortaleza que tienen estas soluciones de ML.

[17] No olvidar que uno de los objetivos principales de la ANSV es la reducción de la siniestralidad vial en general y las víctimas fatales viales en particular.

Referencias y material de consulta

- OPS. 2016. Sistema de datos Manual de seguridad vial para decisores y profesionales. http://whqlibdoc.who.int/cgibin/repository.pl?url=/publications/2008/9789275316283_spa.pdf
- ANSV. Observatorio Vial. 2019. Glosario de términos y definiciones relativas a la seguridad vial. https://www.argentina.gob.ar/sites/default/files/glosario_de_terminos_seguridad_vial.pdf
- ANSV. Observatorio Vial. 2019. Anuario 2019. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_anuario_estadistico_2019_final.pdf
- ANSV. Observatorio Vial. 2020. Informe Anual 2020. Datos preliminares. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_informe_anual_2020_al_4_de_agosto_2021.pdf
- Introduction to Boosted Trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html
- IRTAD 2019. International Transport Forum. Road Safety Annuel Report – Argentina. https://www.itf-oecd.org/sites/default/files/argentina-road-safety.pdf