INGENIERÍA
UNIVERSIDAD CATÓLICA DE SALTA

ESCUELA
DE NEGOCIOS
UNIVERSIDAD CATÓLICA DE SALTA

Universidad Nacional de Salta
FACULTAD DE
INGENIERIA

Facultad de Ciencias Exactas
UNIVERSIDAD NACIONAL DE SALTA

FACULTAD DE
INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

FACULTAD DE TECNOLOGÍA
Y CIENCIAS APLICADAS
UNIVERSIDAD NACIONAL DE CATAMARCA

Facultad de Ciencias
Exactas y Tecnologías

IV JORNADAS INTERNACIONALES DE ESTADÍSTICA APLICADA

Descubre cómo la estadística resuelve desafíos y potencia las investigaciones

2023

Universidad Nacional de Salta
IV Jornadas Internacionales de Estadística Aplicada: Actas de trabajo de investigación JIEA 2021 / compilación de Angélica Noemí Arenas; Héctor Iván Rodríguez; Gisella Carla Mautino ; editado por Angélica Noemí Arenas ; Gisella Carla Mautino ; ilustrado por María Josefina Méndez ; Enrique Ariel Morales del Valle ; prefacio de Héctor Iván Rodríguez ; prólogo de Blanca del Valle Arenas Ramírez. - 1a ed. - Salta: Universidad Nacional de Salta, 2023.
Libro digital, DOCX

Archivo Digital: online
ISBN 978-987-633-601-7

1. Ingeniería de Sistemas. 2. Análisis Comparativo. 3. Matemática Estadística. I. Arenas, Angélica Noemí, comp. II. Rodríguez, Héctor Iván, comp. III. Mautino, Gisella Carla, comp. IV. Méndez, María Josefina, ilus. V. Morales del Valle, Enrique Ariel, ilus. VI. Arenas Ramírez, Blanca del Valle, prolog. VII. Título.
CDD 519.5071

UNIVERSIDAD NACIONAL DE SALTA

FACULTAD DE INGENIERÍA

2023

IV JORNADAS INTERNACIONALES DE ESTADÍSTICA APLICADA

INTRODUCCIÓN - LAS JORNADAS COMO UNA HERRAMIENTA DE TRANSFORMACIÓN

Las primeras Jornadas nacieron, con carácter local, en el evento del 5 y 6 de diciembre del año 2018 para dar respuesta a la necesidad de la capacitación del estudiante de ingeniería en la formación por competencias y respondiendo a la estrategia de enseñanza de la estadística basada en la aplicación práctica. En esta maniobra el estudiante puede palpar la estadística como una importante herramienta científica para resolver, decidir e investigar en distintos y variados ámbitos y disciplinas. En esa oportunidad las Facultades de Ingeniería de la UNSa, UNJu, UCASAL y la Facultad de Ciencias Exactas de la UNSa trabajaron en conjunto para organizar las ***I Jornadas de Estadística Aplicada***, con la visión que el evento contribuiría para promover el desarrollo de la región y el fortalecimiento de sus egresados ya que fue concebido con la particularidad de unir a disertantes del sector empresarial, gubernamental y académico (alumnos, docentes e investigadores), en un modelo de integración planteado por Sábato y Botana, que especifica que las complejas relaciones entre el **sector científico**, **estructura productiva** y **gubernamental**, se facilitan por el camino de la movilidad ocupacional, fomentando el intercambio colaborativo del personal entre las instituciones y compartir experiencias comunes, acercando así a los sectores. Con esta visión, se realizaron las *Jornadas*. La particularidad de las disertaciones fue justamente promover la movilidad y compartir experiencias mediante la exposición secuencial de estudiantes, docentes, investigadores, profesionales de empresas y de instituciones del estado. Esta modalidad fortalece el vínculo entre los participantes, incentiva al alumno, aporta a la formación por competencias y solidifica la vinculación de la Universidad con el sector productivo, gubernamental y el medio.

En diciembre de 2019 se sumaron a las Jornadas las Facultades Ciencias Exactas y Tecnológicas de la UNSe y la Facultad de Ciencias Químicas de la Universidad de Asunción del Paraguay, con participación de expositores de Chile, Bolivia, Colombia Paraguay, México y por supuesto de Argentina, adquiriendo así carácter internacional. El evento fue declarado de Interés Científico y Tecnológico por El Ministerio de Educación Ciencia y Tecnología del Gobierno de la Provincia de Salta, y avalado por resolución de los respectivos Consejos directivos de la UNSa, UNJu, UNSe, y Ucasal. En diciembre de 2020, bajo Pandemia y de manera Virtual se desarrollaron las III Jornadas Internacionales de Estadística Aplicada (III JIEA), donde se integraron formalmente a la organización las Facultades de Ingeniería de la Universidad Nacional de Salta (UNSa), Universidad Católica de Salta (UCASAL), Universidad Nacional de Jujuy (UNJu), Facultad de Tecnología y Ciencias Aplicadas de la Universidad Nacional de Catamarca (UNCA), Facultad de Ciencias Exactas y Tecnologías de la Universidad Nacional de Santiago del Estero (UNSE), Facultad de Ciencias Exactas de la UNSa, Facultad de la Escuela de Negocios de la Universidad Católica de Salta (UCASAL) y Facultad de Ciencias Químicas de la Universidad Nacional de Asunción del Paraguay.

La gran diversidad de casos presentados por empresas, profesionales, docentes, investigadores y alumnos que mostraban el importante uso de las estadísticas fue creciendo

y se incorporaron las Facultades y Universidades la Universidad Autónoma de Nayarit de México, para organizar el 9 y 10 de diciembre de 2021 las ***IV Jornadas Internacionales de Estadística Aplicada***, que se realizaron de manera presencial en el Anfiteatro de la Facultad de Ingeniería de la Universidad Nacional de Jujuy y que se trasmitieron virtualmente por zoom y streaming.

Las primeras Jornadas juegan un importante papel de promoción para la adquisición de competencias para estudiantes y se han convertido en un motor de impulso para nuevas experiencias compartidas entre investigadores, estudiantes, docentes y profesionales de empresas. La estadística aplicada en situaciones que implican una decisión en un marco de incertidumbre, así como en problemáticas de varias disciplinas, nos convoca nuevamente para alcanzar nuevos desafíos en un mundo que se transforma rápidamente hacia un cambio digital. Así también, la Pandemia nos ha dejado valiosas experiencias que promovieron cambios en el trabajo de las personas, en los distintos entornos laborales que se abordan en los trabajos del presente libro.

Héctor Iván Rodríguez

Doctor Ingeniero
Profesor Asociado en Probabilidad y Estadística y Estadística Experimental en la Universidad Nacional de Salta
Profesor Adjunto en Investigación de Marketing del MBA de la Escuela de Negocios de la Universidad Católica de Salta
ivan@ing.unsa.edu.ar

PRÓLOGO - IV JORNADAS INTERNACIONALES DE ESTADÍSTICA APLICADA

"Experiencias adquiridas en un año de Pandemia COVID-19"

Recientemente se ha publicado un informe sobre el estado de la ingeniería en España, llevado a cabo por el Observatorio de la ingeniería, cuyos resultados reafirman mi convicción de su pertenencia al conjunto de ciencias humanísticas.

Desde los inicios de mi formación como ingeniera en Construcciones y Civil en la Universidad Nacional de Salta hasta hace relativamente pocos años, creí que las “ciencias duras”, de las que forman parte la ingeniería y las matemáticas y la estadística como una disciplina más dentro de ella, se distinguían de las “ciencias humanísticas”.

Pero la reflexión junto a la práctica profesional comprometida, me han llevado a cambiar este marco conceptual tradicional.

Comenzaré por reivindicar el enorme servicio de estas disciplinas en la vida de las personas, y el tema central de estas Jornadas "Experiencias adquiridas en un año de Pandemia COVID-19", me permite establecer un marco para hacer esta reivindicación y reconocimiento del papel de la Estadística, tema central de las mismas.

El anuncio de la Pandemia nos obligó a todos a asumir el aislamiento, con la esperanza de que éste era la única vía para salvar vidas. Nos obligó a parar nuestro ritmo vital, para protegernos al abrigo de nuestras casas y, desde allí, encontrar vías de canalizar nuestra energía, que tenía que sumarse al esfuerzo colectivo de varias generaciones. Alguien dijo “pero no era solo una cuestión de fe”, sino la cesión colectiva a unos planes de la Organización Mundial de la Salud, que cada día presentaba resultados de modelos estadísticos, los cuales indicaban que “*a menor contacto menor velocidad del contagio, hasta lograr que los sistemas de salud no se vieran colapsados*”. Y así fue, como lo hicimos posible y real.

Los televisores de nuestros hogares se convirtieron en las ventanas a la vida, mostrando algunas imágenes de la tragedia colectiva y a la vez que ofrecían resultados y ejemplos de modelos de una comunidad científica, trabajando contrarreloj, para dar resultados de modelos y simulaciones que alimentaron la esperanza y dieron sustento a la idea que estábamos “venciendo al virus entre todos”.

Ello contribuyó a familiarizar a muchas personas, con palabras que vienen de la ciencia estadística, cuyo ADN es contar, medir y ponderar. En las conversaciones entre los ciudadanos intervenían palabras clave como “tasa de contagio”, “tasa de mortalidad”, “subregistro de los datos” y cuando empezaron las pruebas y las vacunas los términos de “realización de test aleatorios”, “tasa de positivos”, “tasa de negativos”, “tasa de vacunación de la población”, etc. resultó que así, casi sin saberlo, cada uno de nosotros éramos estadísticos.

Agradezco enormemente al equipo gestor de las IV JIEA, el haberme honrado con su invitación a escribir el prólogo de este libro, por tres motivos principales: en primer lugar, porque realicé mis estudios de formación de grado en esta casa de estudios y en ella adquirí las herramientas y las competencias para emprender un camino profesional fecundo dentro y fuera de Argentina. Y así entonces, como graduada agradecida de todo lo recibido de esta casa, puedo decir que lo de que *"nadie es profeta en su tierra*", depende…. no es este el caso.

En segundo lugar, porque, más tarde, ya radicada en España encontré en la Estadística aplicada, unas herramientas para desarrollar una etapa profesional volcada a ayudar a resolver un problema humano vital: la movilidad y el transporte de forma más segura.

En tercer lugar, porque el marco conceptual de estas jornadas y la invitación a escribir el prólogo del libro me ha permitido reflexionar sobre el papel de esta ciencia en el tiempo de Pandemia del COVID-19 y cómo nos afectó, como colectivo y también a nivel personal.

Así que el tema central de las IV jornadas de Estadística Aplicada, es sumamente oportuno y justo, porque la estadística en ese momento estuvo más presente que nunca y jugó un papel de enorme importancia, como lo tiene en tantos otros campos que afectan nuestras vidas: la salud, la economía, la toma de decisiones, la convivencia, o la relación con la tecnología entre otros tantos, y de lo que no somos plenamente conscientes. La cantidad de trabajos presentados en ella, dan cuenta de su importancia, y felicito a todos los ponentes por el esfuerzo significativo de trabajar en las condiciones que impuso la Pandemia, por la resiliencia desarrollada para no sucumbir ante la incertidumbre interna y externa, por la originalidad de los enfoques de los estudios presentados, y por poner en valor lo antes dicho: "la omnipresencia de la estadística en nuestras vidas".

No quiero terminar esta breve contribución al texto de las actas de tan importantes Jornadas. sin destacar el valor de su continuidad, es la edición número cuatro, la calidad de su contenido y organización, así como su contribución al prestigio de la Universidad, que desde un rincón del Norte Argentino contribuye de manera singular a la formación y desarrollo profesional y humano de tantos jóvenes y a hacer mejor la sociedad y la economía de nuestra tierra.

Blanca Arenas Ramírez

Doctora Ingeniero de Caminos, Canales y Puertos.
Profesora Titular de Universidad. ETSI Industriales. Universidad Politécnica de Madrid UPM)
Directora de la Unidad de Estudios de transporte e impacto medioambiental de los vehículos del INSIA. Instituto Universitario de Investigación del Automóvil Francisco Aparicio Izquierdo (INSIA). Campus Sur de la UPM; Carretera de Valencia km. 7. CP:28031. MADRID.
Telf.:+34 910677273; e-mail: blanca.arenas@upm.es

www.insia-upm.es.

Introducción

En la edición anterior del Libro de las Jornadas, nos pareció atinado presentar las instituciones participantes, pues de manera animada, los colegas de distintas disciplinas nos unimos para difundir trabajos que involucraron a las universidades, empresas, otras organizaciones y gobiernos. El resultado nos alumbró más, pues desde la visión y de los objetivos de las instituciones permitió contemplar el camino recorrido y el aporte de éstas en la sociedad.

En la presente edición se consideró importante identificar aquellos representantes de la flora y fauna de los países participantes que fueran imponentes a la vista y también para llamar la atención sobre la importancia de preservar y mantener los recursos naturales de nuestras comunidades.

En las páginas que siguen, los colegas de las instituciones nos hicieron llegar sus aportes y de esa manera mostrar el interés sobre la naturaleza de nuestros países.

Jaguar o yaguareté - Universidad Nacional de Salta

Fotografía: Vista de un jaguar creada con inteligencia artificial.

La imagen fue creada a partir de inteligencia artificial, en el ejercicio virtual se pidió: “crear una imagen de un Yaguareté con hojas verdes de fondo que se encuentre en las Yungas argentinas”. Este animal se encuentra en peligro crítico de extinción; se estima que quedan 200 ejemplares entre Salta y Jujuy, un número muy preocupante para una especie biológica de imponente belleza. Esta condición nos permite pensar en la situación actual y cómo podemos revertirla desde nuestro lugar. Quizás hablando de ello, educando, concientizando, pero sobre todo cuidando nuestro planeta. El jaguar o yaguareté como se lo denomina localmente es una especie bandera promocionada por Greenpeace en sus campañas para la conservación de las Yungas.

El estudio es reciente y menciona el número de ejemplares y el peligro que esto significa, no sólo de estos felinos sino de sus presas. Fuente consultada:

https://www.todojujuy.com/jujuy/yaguarete-creen-que-quedan-200-ejemplares-jujuy-y-salta-n142469#:~:text=Se%20estima%20que%20en%20Salta,quedan%20apenas%20unos%202000%20ejemplares.

Vicuña - Universidad Nacional de Jujuy

La vicuña es una especie de mamífero artiodáctilo perteneciente al género Vicugna (del grupo de camélidos sudamericanos) que vive en las alturas andinas de Sudamérica, En el pasado fue considerado un animal sagrado por los incas que valoraban tanto su lana hasta el punto de que solo la realeza podía vestir prendas tejidas con esta fibra obtenida a través del chaccu, un ritual que consiste en capturar a esta especie silvestre solo para trasquilar y luego devolverla a su hábitat. Esta costumbre aún se conserva y es el principal modo de aprovechar esta lana que se exporta para, una vez transformada, ocupar los escaparates más elegantes del mundo.

En la provincia de Jujuy, la Ley N° 5634 “Plan de Conservación y Manejo Sustentable de la Vicuña en Silvestría” y su Decreto Reglamentario Nº 5175 sancionados en el año 2009 habilitaron a las comunidades andinas a cosechar la fibra de las vicuñas silvestres. Esta especie nativa estuvo en peligro de extinción a mediados del siglo pasado. En la década del 90, luego del esfuerzo mancomunado de los cinco países signatarios del Convenio Internacional de la Vicuña, se logró recuperar y aumentar su número poblacional.

Fotografía 2: Una vicuña en su hábitat natural.

Taruca o venado andino, monumento natural en peligro de extinción (Catamarca)

La taruca (*Hippocamelus antisensis*) es un ciervo autóctono en la Argentina que constituye un emblema de la Región del Noroeste Argentino y que representa un Patrimonio Natural y Cultural muy valioso para la Región. Denominada también como venado andino o huemul del norte, es una de las 19 especies de mamíferos clasificadas “en peligro”, según el Libro Rojo de los mamíferos amenazados de la Argentina. En 1996, fue declarada monumento natural, con el objetivo de protegerla.

Se la puede encontrar desde Perú, el oeste de Bolivia, hasta el norte de Chile y noroeste de la Argentina. En nuestro país la encontramos en Jujuy, Salta, Catamarca, Tucumán y La Rioja. Se la encuentra en pastizales y estepas de altura (entre los 1900 hasta los 5000 metros sobre el nivel del mar) con terrenos escarpados y quebradas abruptas. En Argentina, se registran

poblaciones de tarucas en los Pastizales de Altura o Pastizales de Neblina y en los Pastizales Altoandinos.

En Catamarca, habita zonas húmedas de las Sierra de Ambato a partir de los 3000 a 3600 metros, en localidades como Los Ángeles, Concepción, El Machando, Humaya. También hay en Aconquija, desde los 2800 hasta los 4.000 metros y en las sierras de Belén.

Si bien, como animal autóctono la taruca habitó el continente desde hace siglos, su nombre y características trascendieron cuando en 2018 se decidió que su imagen fuera el centro del billete de cien pesos.

La caza y comercialización de esta especie, sus productos y subproductos se encuentra prohibida en todo el territorio nacional.

Fuente:https://www.argentina.gob.ar/sites/default/files/ficha-taruca72.pdf, https://www.pagina12.com.ar/366080-taruca-o-venado-andino-monumento-natural-en-peligro

Fotografía 3: Licencia Creative Commons, a partir de: https://commons.wikimedia.org/wiki/File:Hippocamelus_antisensis_114839310.jpg

Passiflora cincinnata - Universidad Nacional de Asunción, Paraguay

Una de las especies nativas del Paraguay es Passiflora cincinnata, conocida con el nombre común de mburucuya pytã. Es utilizada como planta ornamental debido a la belleza de sus flores que pueden verse durante todo el año.

La flor de mburucuya pytã o pasionaria también es conocida como la flor de la pasión, nombre dado por los colonizadores españoles en el siglo XVI, haciendo alusión a su forma semejante a una corona de espinas con tres clavos en la cruz que en realidad son sus pistilos.

Es una hierba trepadora nativa del Paraguay que crece en bosques higrófilos ribereños, en sabanas inundables, en bordes de los caminos, serranías y en lugares donde abundan las rocas. Su uso en medicina popular está enfocado en las hojas para la preparación de infusiones y decocciones relajantes, también se consume para el tratamiento de la hipertensión. Sus frutos comestibles son amarillo-rojizos en la madurez, con semillas oscuras y pulpa transparente de sabor ligeramente ácido.

Los habitantes del Paraguay se identifican con esta hermosa flor del mburucuya pytã a la que consideran como la flor nacional.

Informaciones y fotografías proporcionadas por el Ing. Agr. Germán González, investigador del Departamento de Botánica de la Facultad de Ciencias Químicas de la Universidad Nacional de Asunción.

Ocelote – Universidad de México

El ocelote es una especie de mamífero carnívoro de la familia Felidae. Se encuentra ampliamente distribuido en América, principalmente en ambientes tropicales, donde se diferencia en numerosas subespecies. Puede confundirse con el margay o tigrillo.

Angélica Noemí Arenas
Doctora en Ingeniería Industrial
Profesora Asociada en Ingeniería de Planta y Gestión Ambiental de la Universidad Nacional de Salta
Profesor Titular en Proyecto de Grado de la Universidad Católica de Salta
angelica@ing.unsa.edu.ar

ÍNDICE

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Diagnóstico y análisis de necesidades diferenciales para un envejecimiento saludable de los adultos mayores de la ciudad de el Carmen – Jujuy – Argentina año 2020-2021

Carmen Graciela Gallardo[1], Noelia Fabiana Cruz[1], Norma Gladys Mogro[1], Ana María Chalabe[2]

Institución: 1: Nuestra Señora del Carmen, El Carmen, Jujuy – 2; Facultad de Humanidades y Ciencias Sociales – Universidad Nacional de Jujuy

Datos de contacto: carmengra6@yahoo.es Tel: 3885006365

RESUMEN

Las ECNT – Enfermedades Crónicas no Transmisibles - han tomado mucha relevancia en los últimos tiempos. En nuestra ciudad no se tiene dimensión de la población afectada y el grado de gravedad que conlleva esta situación la cual no permite que los Adultos Mayores logren un bienestar en esta etapa de la vida. Con este trabajo se buscó evidenciar los problemas y tratar de llegar a una conclusión para mejorar esta realidad, desde el punto de vista de la salud contribuyendo a idear un plan que beneficie a estas personas. Para ello se aplicó una encuesta estandarizada por la OPS – Organización Panamericana de la Salud – en personas de 60 años o mayores (Método STEPS Versión Panamericana).

Esta investigación constituye una experiencia metodológica en el abordaje de estos hechos. Buscando determinar indicadores que establezcan relaciones entre factores de riesgo y las ECNT, y de esta manera avanzar en la caracterización de la población municipal.

Se realizó un estudio observacional, descriptivo, de corte transversal sobre 1900 habitantes de 60 años y más de la base de datos de Atención Primaria de la Salud del Hospital Nuestra Señora del Carmen, (APS), correspondientes al año 2020 y 2021.

La encuesta recabó datos de demografía, comportamiento, mediciones físicas, bioquímicas y un módulo opcional de patologías prevalentes.

Palabras Clave: Adultos mayores, ENTC, Calidad de vida, Vejez, Salud.

INTRODUCCIÓN

Las enfermedades crónicas no transmisibles del adulto (ECNT) constituyen hoy en día el principal problema de salud del país definiéndose a las mismas como "enfermedades de etiología incierta, habitualmente multicausales, con largos períodos de incubación o latencia; largos períodos subclínicos con prolongado curso clínico, con frecuencia episódico; sin tratamiento específico y sin resolución espontánea en el tiempo".
En general incluyen enfermedades cardiovasculares, enfermedad cerebro-vascular, enfermedades neoplásicas, enfermedades respiratorias crónicas, enfermedades osteoarticulares invalidantes, enfermedades invalidantes como diabetes mellitus y otras.
La información disponible nos indica que muchas de estas enfermedades son prevenibles, y sus muertes a edades tempranas evitables. Se han identificado factores de riesgo (FR) asociados epidemiológicamente a ellas, como tabaquismo, consumo excesivo e inapropiado de alcohol, inactividad física, obesidad, hipertensión arterial, perfil lipídico alterado y dieta inadecuada.
En la Ciudad del Carmen Jujuy no se han desarrollado hasta el momento programas integrales que contemplen la detección en las enfermedades del adulto, la intervención apropiada, el seguimiento planificado, promoción de la salud, el envejecimiento activo y saludable, acompañados de una atención primaria de la salud adaptada al adulto.
De no mediar una acción preventiva que modifique el nivel y perfil de los FR, esta población alcanzará en los años por venir cifras epidémicas siguiendo la tendencia mundial. Esto indica que existe una potencialidad de daño que aún no se ha manifestado clínicamente. Podemos suponer que en muchos individuos la historia natural de las enfermedades crónicas está en etapas tempranas, subclínicas. En estos grupos, una oportuna y eficiente intervención impediría o retardaría el curso inexorable de las ECNT mejorando la salud integral y calidad de vida de las personas adultas.

METODOLOGÍA

La vigilancia se considera como una función esencial de la salud pública. Es una acción reconocida como parte de las responsabilidades para la preservación de la salud de las comunidades. Al mismo tiempo, es una herramienta para llevar a cabo la salud pública basada en evidencias, la toma de decisiones y el monitoreo del éxito de las intervenciones en salud, con información coherente entre áreas y dentro de las mismas. Los programas destinados a enfermedades crónicas se orientan hacia registros poblacionales permitiendo establecer patrones de enfermedad a nivel comunitario, el análisis y la diseminación de la información obliga a homogeneizar la información para hacerla comparable. De esta manera, una de las principales actividades a desarrollar es establecer una línea basal de información que permita un seguimiento posterior, estableciendo tendencias temporales y geográficas para las enfermedades crónicas en los adultos mayores de la Ciudad del Carmen. Los factores de riesgos frecuentes y evitables son la base de la mayor parte de las enfermedades crónicas. Estos factores de riesgo de enfermedades crónicas son la principal causa de la carga de mortalidad y de discapacidad. La Organización Mundial de la Salud (OMS) propone un modelo progresivo de vigilancia de enfermedades crónicas no transmisibles: El STEPS PANAMERICANO: es un método simple estandarizado para recolectar, analizar y diseminar información sobre factores de riesgo a otras áreas y es el método que se aplicó, utilizando el protocolo y las preguntas estandarizadas, se usó la información obtenida para monitorear las tendencias locales. El método facilita la recolección de la información en forma rutinaria y continua. STEPS es un proceso secuencial. Comienza con la recopilación de información fundamental sobre los factores de riesgo por cuestionario, (STEPS 1); a continuación, pasa a unas mediciones físicas sencillas (STEPS 2) y, después, a una recogida más compleja de muestras de sangre para su análisis bioquímico (STEPS 3)

Definición operacional de las variables: Demografía, Sexo, Edad, Educación, Estado Civil, Actividad Laboral, Nivel de ingresos, Consumo de Tabaco, Dieta, Actividad Física, Tensión

Arterial, Diabetes, Perfil lipídico, Historia familiar, Mediciones físicas: altura, peso, cintura, contorno de cadera, presión arterial.

Mediciones bioquímicas: Glucemia, Colesterol, HDL-Colesterol, Triglicéridos, Creatinina.

Información adicional: estudios médicos complementarios realizados para prevención de cáncer: Colonoscopia, PAP, Mamografía, Estudios de Próstata.

Descripción del ámbito de estudio.

El ámbito de estudio se desarrolló en un hospital del interior de la Provincia de Jujuy cuya misión consiste en proporcionar a la población una asistencia médico-sanitaria completa, curativa y preventiva, y cuyos servicios externos irradian hasta el ámbito familiar. Se trata de un Hospital cabecera de Área Programática del que depende un Centro Regional de Referencia y seis Puestos de Salud; atiende una población de casi 20.000 habitantes. El 38.8 % de familias tienen necesidades básicas insatisfechas; 34.1 % de la población está por debajo de la línea de pobreza y 25.4 % por debajo de línea de indigencia. La población de 60 años o más es alrededor de 2000 personas. La Institución cuenta con indicadores que expresan sensiblemente las condiciones de vulnerabilidad social y su relación con las situaciones de crisis.

Tipo de estudio y diseño.

El estudio es observacional, descriptivo, de corte transversal.

Universo o población objetivo:

Personas de 60 años o más de ambos sexos que viven en la ciudad de El Carmen Provincia de Jujuy, en los años 2020 y 2021.

Unidad de análisis, criterios de inclusión y exclusión:

Unidad de análisis: Cada persona de 60 o más años de la ciudad de El Carmen Provincia de Jujuy.

Criterios de inclusión: Los criterios de inclusión fueron los siguientes:

- personas de ambos sexos
- edad entre 60 a más años
- residentes en la ciudad de El Carmen

Criterio de exclusión: No participaron aquellas personas que:

- no sean residentes en la ciudad de El Carmen
- aquellos que no expresen su consentimiento informado.
- Menores de 60 años.

Población accesible. Muestra. Selección y tamaño de la muestra. Análisis de sesgos.

Población accesible: Se realizó la encuesta a los habitantes de 60 años o más de la ciudad de El Carmen, durante los años 2020 y 2021.

La selección de la muestra se realizó en cuatro etapas:

Se realizó un filtro del total de registros para determinar los excluidos según criterios de exclusión ya enunciados. Con los registros restantes se confeccionó un registro definitivo.

Se realizó un MUESTREO SIMPLE ALEATORIO. Se consideró un N= 2.000 teniendo en cuenta que la población censada por Atención Primaria de la Salud (APS) durante el año 2019 fue de 1960 y que el muestreo tendrá una duración de un año.

El tamaño muestral n calculado es de 360 pacientes (IC 95%), y n= 2000 individuos, para compensar los rechazos de participación (10 %).

Sesgos: Para evitar los sesgos por selección se realizó la elección de registros a analizar mediante la asignación de número aleatorios con el software EPI DAT 3.0.

Selección de técnica e instrumento de recolección de datos. Fuentes primarias y secundarias. Prueba piloto del instrumento

Como fuente primaria de datos se procedió a realizar la encuesta STEPS.

Para las Mediciones Bioquímicas se procedió a la toma de muestras de sangre en ayunas, por el Profesional Bioquímico, las cuales fueron procesadas en el Laboratorio del Hospital Ntra. Señora del Carmen, validando los resultados al cumplir los controles de calidad establecidos (internos y externos).

Prueba piloto: se realizó una experiencia piloto en quince personas para verificar la interpretación correcta de cada ítem a considerar.

Plan de análisis de los resultados.

Una vez obtenida la información se procedió a:

- Clasificar y ordenar el material y datos relevados.
- Sistematizar los datos y relevar las categorías.
- Interpretar la información en el marco de cada uno de los procesos.

El análisis estadístico se genera mediante el empleo de técnicas de cálculo matemático con soporte de software estadístico, presentación tabular y gráfica e interpretación de modo sucesivo y lógico de tres tipos de medidas: Medidas de ocurrencia, Medidas de asociación y Medidas de significación estadística.

Se comienza con análisis univariados, construcción de Tablas de frecuencias y se complementa con análisis de subgrupos por ejemplo según género, edad, etc. En esta investigación se tratan diversas variables, se las categoriza en una etapa previa al análisis y se observa la distribución por sí solas, para entender cómo se comportan. El Análisis bivariado entre variables que se desean estudiar se realiza en forma tabular y gráfica por subgrupos de frecuencias encontradas. *Análisis Multivariado:* Se desea conocer la relación entre variables juntas. Según la característica de las variables, se recurren a las siguientes expresiones: 1. Medidas de tendencia central y dispersión (media, mediana, moda, Cuartiles, mínimo, máximo y Desvío Estándar), 2. Frecuencias (absoluta o relativa, proporciones) y 3. Medidas de Asociación y Significancia. El procesamiento estadístico de los datos se realiza con el paquete informático EPI-INFO 7.1.5.2, Excel 2010. Todas las Tablas y Gráficos del trabajo fueron de elaboración propia, realizados con registros de las encuestas recabadas en el periodo 2020-2021.

Consideraciones éticas:

Toda la información relevada se trató garantizando el carácter confidencial y anónimo de los datos, de manera que no sea posible identificar personas físicas ni jurídicas. Para ello se respetó lo establecido en la Ley Nacional 25.326 Artículos 8 y 9.

Se recabó el consentimiento informado de los encuestados.

DESARROLLO

Información Demográfica

Tabla 1– Distribución de la población según grupos de edad y sexo

Grupos de Edad	Mujeres		Hombres		Ambos Sexos	
	n	%	n	%	n	%
60-64	62	17,9%	40	11,6%	102	29,5%
65-69	76	22,0%	44	12,7%	120	34,7%
70-74	42	12,1%	26	7,5%	68	19,7%
75-79	14	4,0%	14	4,0%	28	8,1%
80-84	20	5,8%	2	0,6%	22	6,4%
85-89			6	1,7%	6	1,7%
Total general	**214**	**0,6**	**132**	**0,4**	**346**	**1,0**

Fuente: Encuesta STEPS 2021

En cuanto al nivel de educación la población registró los siguientes datos: sin escolarización 5.7%, primaria incompleta 10.4%, primaria completa 53.2%, secundaria incompleta 10.9%, secundaria completa 12.7%, terciaria 2.9 % y universitaria 4.2%

En cuanto al estado civil de las personas encuestadas el 50.3 % de ambos sexos están casados

Con respecto a la situación laboral el 68.3 % no trabaja, siendo el 54.3 % de ellos jubilados.

El ingreso económico en el hogar ronda entre $12.000 y $24.000 mensual, lo cual se corresponde con la jubilación mínima

Consumo de Tabaco

En cuanto al consumo de tabaco se observó una prevalencia mayor en hombres que alguna vez fumaron (26,9%), siendo el 64,5 % de estos los que lo hacían a diario y el 45,1% aún lo continúan haciendo

Tabla 2 - Patrón de consumo de tabaco fumado por sexo y edad

Mujeres													
Grupos de Edad	n	Ocasional			A diario			Ex Fumador			Nunca		
		n	%	IC 95%	n	%	IC 95%	n	%	IC 95%	n	%	IC 95%
60-64	62	16	4,6%	2,4 - 6,8	6	1,7%	0,3 - 3	10	2,9%	1,1 - 4,6	46	13,3%	9,7 - 16,8
65-69	76	34	9,8%	6,6 - 12,9	20	5,8%	3,3 - 8,2	14	4,0%	1,9 - 6	42	12,1%	8,6 - 15,5
70-74	42	8	2,3%	0,7 - 3,8	2	0,6%	0,2 - 1,4	6	1,7%	0,3 - 3	34	9,8%	6,6 - 12,9
75-79	14	4	1,2%	0,05 - 2,3	2	0,6%	0,2 - 1,4	2	0,6%	0,2 - 1,4	10	2,9%	1,1 - 4,6
80-84	20	4	1,2%	0,05 - 2,3							16	4,6%	2,4 - 6,8
85-89													
Totales	214	66	19,1%	15 - 23,2	30	8,7%	5,7 - 11,6	32	9,2%	6,1 - 12,2	148	42,8%	37,5 - 48
Hombres													
Grupos de Edad	n	Ocasional			A diario			Ex Fumador			Nunca		
		n	%	IC 95%	n	%	IC 95%	n	%	IC 95%	n	%	IC 95%
60-64	40	23	6,6%	4 -9,2	14	4,0%	1,9 - 6	9	2,6%	0,9 - 4,2	17	4,9%	2,6 - 7,1
65-69	44	32	9,2%	6,1 - 12,2	20	5,8%	3,3 - 8,2	12	3,5%	1,6 - 5,4	12	3,5%	1,6 - 5,4
70-74	26	24	6,9%	4,2 - 9,5	6	1,7%	0,3 - 3	18	5,2%	2,8 - 7,5	2	0,6%	0,2 - 1,4
75-79	14	8	2,3%	0,7 - 3,8	2	0,6%	0,2 - 1,4	6	1,7%	0,3 - 3	6	1,7%	0,3 - 3
80-84	2	2	0,6%	0,2 - 1,4									
85-89	6	4	1,2%	0,05 - 2,3							2	0,6%	0,2 - 1,4
Totales	132	93	26,9%	22,2 - 31,5	42	12,1%	8,6 -15,5	45	13,0%	9,4 - 16,5	39	11,3%	7,9 - 14,6
Ambos sexos													
Grupos de Edad	n	Ocasional			A diario			Ex Fumador			Nunca		
		n	%	IC 95%	n	%	IC 95%	n	%	IC 95%	n	%	IC 95%
60-64	102	39	11,3%	7,9 -14,6	20	5,8%	3,3 - 8,2	19	5,5%	3 - 7,9	63	18,2%	14,1 - 22,2
65-69	120	66	19,1%	15 - 23,2	40	11,6%	8,2 - 15	26	7,5%	4,7 - 10,2	54	15,6%	11,7 - 19,4
70-74	68	32	9,2%	6,1 - 12,2	8	2,3%	0,7 - 3,8	24	6,9%	4,2 - 9,5	36	10,4%	7,2 - 13,6
75-79	28	12	3,5%	1,6 - 5,4	4	1,2%	0,05 - 2,3	8	2,3%	0,7 - 3,8	16	4,6%	2,4 - 6,8
80-84	22	6	1,7%	0,3 - 3							16	4,6%	2,4 - 6,8
85-89	6	4	1,2%	0,05 - 2,3							2	0,6%	0,2 - 1,4
Totales	346	159	46,0%	40,7 - 51,2	72	20,8%	16,5 -25	77	22,3%	17,9 - 26,6	187	54,0%	48,7 - 59,2

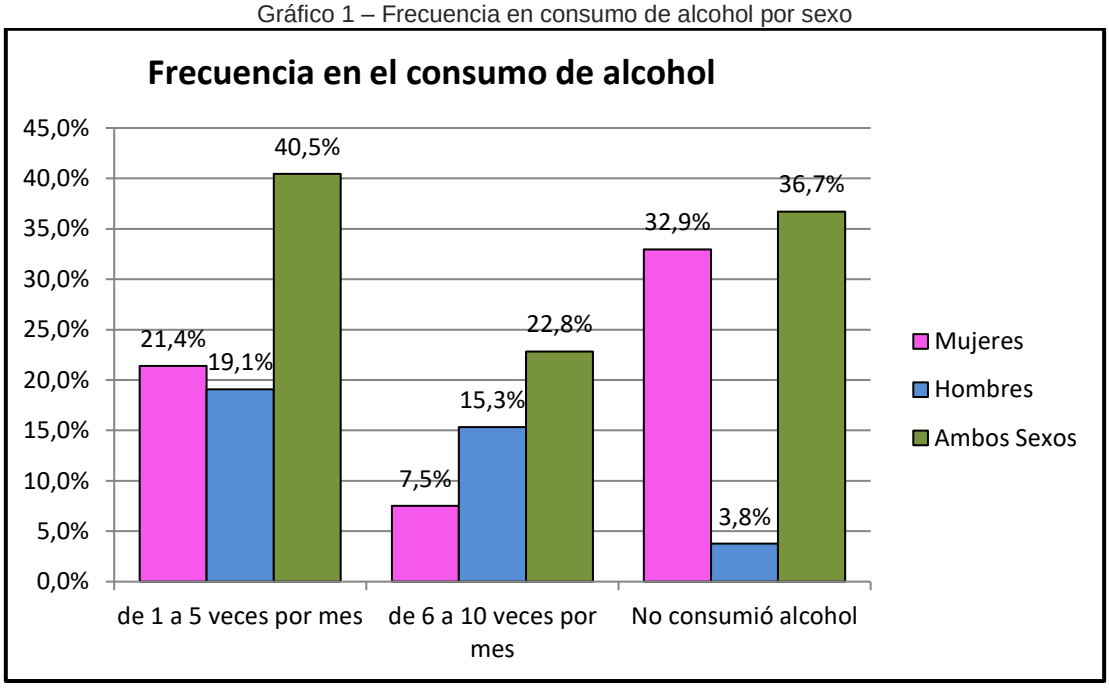
Según la edad la menor prevalencia se observó en el grupo etario de 75 a 79, registrándose valores similares en los grupos inferiores.

En ambos sexos la prevalencia de fumadores a diario fue 20,8% y el 46% en fumadores ocasionales. De los no fumadores el 22,3% corresponde a los que dejaron de fumar y el 54% nunca fumaron siendo casi cuatro veces más en mujeres que hombres. La media de la edad de inicio de fumar fue de 28 años, siendo menor en el grupo de hombres (26 años). La media de la edad en la cual se dejó de fumar fue de 55 años para ambos sexos.

Consumo de Alcohol

En ambos sexos el 39.6% reportó ser abstemio de toda la vida, siendo el 35.85% en mujeres y el 3.8%en hombres

Gráfico 1 – Frecuencia en consumo de alcohol por sexo



Fuente: Encuesta STEPS 2021

El 60.4% consumió alcohol alguna vez en su vida y el 53.8% bebió alcohol en los últimos 12 meses

La frecuencia de consumo en el último año fue: 40,5% de 1 a 5 veces por mes, y un 22.8% de 6 a 10 veces por mes con mayor prevalencia en hombres (15.3%) siendo un 11.3% correspondiente al rango de edad de 60 a 69 años.

De los bebedores actuales el 17,6 % reportó que consumió alcohol no registrado

Dieta. Consumo de frutas y vegetales

En cuanto al consumo de frutas se observa que un 57% de la población consume al menos una porción de frutas de 5-7 días por semana y el 21% consume al menos una porción 1 o 2 veces por semana. El 35% del grupo etario de 60 a 69 años es el mayor consumidor de frutas. La media de números de días por semana que se consumen frutas, en mujeres es de 5.21, en hombres es 4.23 y en ambos sexos 2.33 días.

De la población encuestada el 15% reportó un consumo de verduras de 1-2 días por semana, el 13% de 3-4 días y el 72 % de 5-7 días.

El 61,8% consume aceite vegetal en la preparación de sus alimentos y solo el 41,6% de la población consume alimentos fuera de su casa

Actividad Física

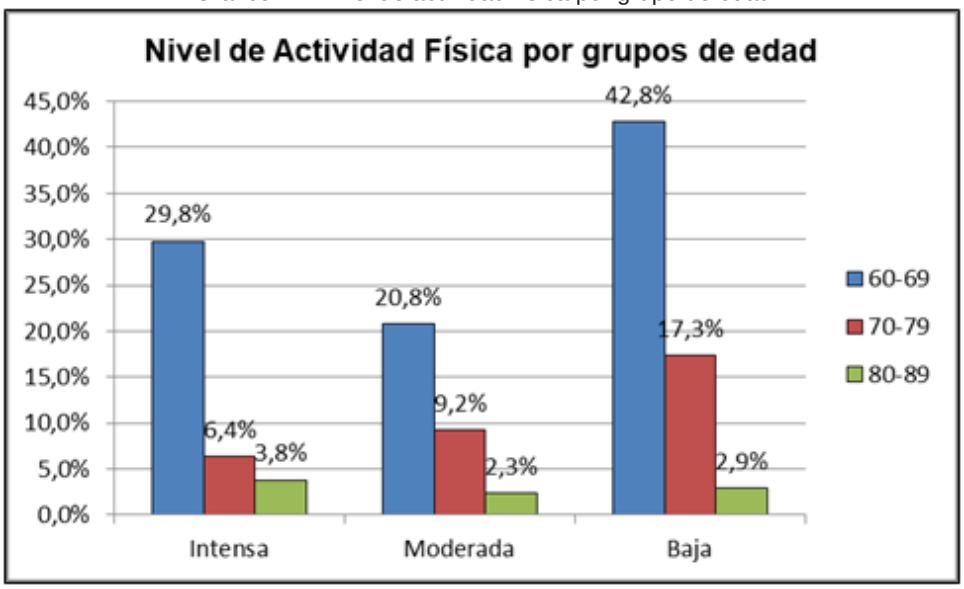
De la población que trabaja el 39,9 % realiza una actividad física intensa en el trabajo, observándose el mayor porcentaje (29,8 %) en el grupo etario de 60 a 69 años para ambos sexos

El 22,3 % realiza una actividad física moderada y el 63% una actividad baja. Con medias de 1,24 horas por día para actividad, intensa 2,7 horas para actividad moderada y 3,3 hs para actividad baja

En cuanto a la actividad física realizada en el tiempo libre el 49,7% realiza alguna actividad física observándose mayor prevalencia (26%) en el grupo etario de 60 a 69 años con una frecuencia de 1 a 3 días por semana para ambos sexos

El valor de sedentarismo es de 50,3% en la población que no trabaja

Gráfico 2 – Nivel de actividad física por grupo de edad



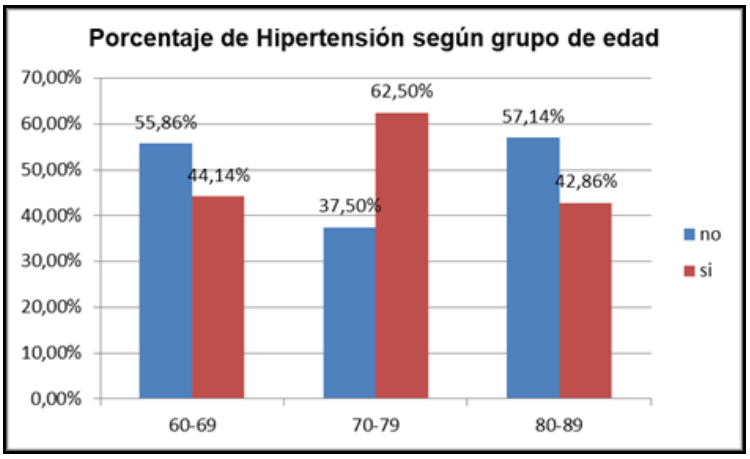
Tensión arterial

De los datos obtenidos se observa una media de Presión Diastólica de 140,59 mmHg en mujeres y de 146.05 mmHg en hombres. Con una moda más alta en mujeres (160mmHg) que hombres (150mmHg). Estos datos arrojan una prevalencia en presión elevada (> a 140mmHg)

En cuanto a la media de presión sistólica entre los grupos, en las mujeres fue de 83.20 mmHg y en los hombres de 88,74 mmHg,

El 49,3% de la población adulta tiene presión alta, con mayor prevalencia en mujeres 48.6% que hombres 38.2%.

Gráfico 4 – Porcentaje de Hipertensión según grupo de edad



De los hipertensos se observó una mayor prevalencia en la población etaria de 70-79 años (62.50%). El 25% de los encuestados confirmó tener antecedentes de hipertensión en la familia

Diabetes

El 21,97% de la población encuestada reportó ser diabética, con 24,30% en mujeres y 18,18% en hombres. El 23% de la población reportó tener antecedentes de diabetes en la familia

Perfil lipídico

Al 44,51% le han dicho que tiene colesterol alto con mayor prevalencia en mujeres (30,64%) del cual solo el 10,98% recibió un tratamiento oral en las últimas 2 semanas

Historia familiar

Se indagó a la población sobre si recibieron recomendaciones o sugerencias en el estilo de vida a fin de prevenir o disminuir algún factor de riesgo cardiovascular, para lo que se encontró que el 27% hace alguna dieta especial y el 25% recibió consejos para aumentar la actividad física realizada. Con respecto a los antecedentes familiares se observó que el 25% tiene familiares con hipertensión siendo este el valor de mayor prevalencia seguida por la diabetes (23%)

Medidas antropométricas

La prevalencia de sobrepeso se definió con un IMC superior a 25 e inferior a 30, prevalencia de obesidad con IMC mayor a 30, la suma de ambos se reporta como población con exceso de peso.

A partir de los datos obtenidos de mediciones de peso y altura se calculó el IMC obteniendo una media de 28,25 en mujeres y 29,30 en hombres; siendo la moda más alta en hombres (30,47) que mujeres (26,64)

La media del perímetro de cadera fue 106,98 cm para mujeres y 105,94 cm para hombres, siendo la moda mayor en mujeres que hombres

La media de perímetro de cintura 100,55 cm para mujeres y 104,83 cm para hombres, con igual moda para ambos sexos. Con estos datos se calculó la índice cintura/cadera siendo la media 0,99 en hombres y en mujeres de 0,94; con índice superior a 0,94 y mujeres con 0,82 tendrán un riesgo elevado de sufrir enfermedades cardiovasculares o metabólicas.

Medidas bioquímicas

Glucemia

Con respecto a la medida de glucemia se observó una media de 106,21 mg/dl en mujeres y 116,20 mg/dl en hombres sin discriminar población diabética de no diabética. Aun así, se observa que la moda para esta variable ingresa dentro de los valores normales (70 a 110 mg/dl).

Se observó una media de 115,11 mg/dl en el grupo etario de 60 a 69 años. La moda se conservó dentro de los valores esperados.

Analizando el porcentaje de frecuencia de las glucemias se clasificó la población en hipoglucémica (2,31%), normo glucémica (72,83%) e hiperglucemia (24,86%), reportándose el 21,7% de la población con diabetes.

Colesterol

El colesterol en ayunas arrojo una media de 196,51 mg/dl con una moda de valor igual a 200mg/dl. Observándose en el grupo etario de 60 a 69 años una mayor media (201 mg/dl) cayendo sobre el límite superior de colesterol normal, se encontró que el grupo etario de 70 a 79 es el único que contiene una moda(160mg/dl) dentro del valor considerado normal (hasta 200 mg/dl).

Al estudiar el componente HDL de colesterol se observa una media de 47,34 mg/dl.

Al 44,51% le han dicho que tiene colesterol alto con mayor prevalencia en mujeres (30,64%) del cual solo el 10,98% recibió un tratamiento oral en las últimas 2 semanas

Creatinina

Se observa una media de 0,94 con una moda de 1,14. Así toda la población presenta valores normales de creatinina en ayunas.

Triglicéridos

Los triglicéridos en ayunas tienen una media de 154,55 mg/dL con una moda de 105 mg/dL. Según la edad para el grupo de 60 a 69 años se obtuvo una media de 164,04 mg/dL con una moda de 100mg/dL. Aunque hay diferencias entre las medias de los grupos de 70 a 79 años y 80 a 89 años, las modas caen por debajo del valor normal esperado (150 mg/ dL). En cuanto al sexo se observa que los valores caen sobre el límite superior esperado para ambos sexos.

Colonoscopia

Para el caso de rastreo de cáncer de colon se vio que solo un 5,20% de la población se realizó el examen. De los cuales el 55,56% fueron mujeres y 44,44% hombres

Sangre Oculta en Materia Fecal

Solo el 16,76% se realizó el test de sangre oculta en materia fecal (SOMF) y se observa mayor prevalencia en mujeres 62,07% que hombres 37,93%

Mamografía y Papanicolaou

El 70% de las mujeres se realizó una mamografía en los últimos años, el 60,3% se realizó un PAP en los últimos 5 años

Estudio de próstata

El 83,33% de la población masculina nunca se realizó el examen de tacto rectal.

CONCLUSIONES

El riesgo de padecer alguna Enfermedad No Transmisible está directamente relacionado con los distintos Factores de Riesgo. La encuesta STEPS 2021 genera una línea de base para el análisis de la prevalencia de factores de riesgo en adultos mayores a 60 en la ciudad El Carmen, esta información es relevante para fortalecer la vigilancia, políticas y acciones de prevención y control de las ENT. Se observó que la población mayor de 60 años, a pesar de que la mayoría es jubilada y tiene un ingreso económico mínimo, vive en condiciones edilicias y cotidianas normales para subsistir. Los que participaron en este estudio el 61.85% fueron mujeres y el 38.15% hombres, de los cuales se observó que el sexo masculino rehusaba contestar fielmente la encuesta y fue difícil captar información para una mejor distribución de datos.

Control de tabaco
Los resultados de STEPS muestran que hay una prevalencia mayor en hombres que alguna vez fumaron, y en la actualidad el hábito ocasional es el más frecuente en ambos sexos, en nuestras visitas no se evidenció padecimientos de EPOC ni asistencia de O2 tampoco se encontró fumadores de cigarrillos electrónicos, habanos, pipas o el uso de chicles antitabaco. La media de la edad en la cual se dejó de fumar fue de 55 años para ambos sexos. Según la OMS fumar cualquier tipo de tabaco reduce la capacidad pulmonar, lo que conlleva un mayor riesgo de sufrir afecciones pulmonares graves y puede aumentar la gravedad de las enfermedades respiratorias. En la actualidad, y con respecto a la pandemia por COVID-19 que resulta ser un patógeno sumamente infeccioso que ataca principalmente a los pulmones. El tabaquismo deteriora la función pulmonar, lo que hace más susceptible a la persona frente a infecciones por virus, bacterias y otras afecciones respiratorias. Los datos de investigación disponibles hasta la fecha parecen indicar que los fumadores tienen un mayor riesgo de desarrollar síntomas graves y de fallecer a causa de COVID-19.

Consumo de alcohol
Los resultados de la encuesta registraron una frecuencia de consumo de 1 a 5 veces por mes, y solo 2 de cada 10 lo realizaron con una frecuencia mayor a esta, también se vio una mayor prevalencia en hombres correspondiente al rango de edad de 60 a 69 años.

De los bebedores actuales el 17,6 % reportó que consumió alcohol en exceso; aunque se notó reticencia al responder esta consigna. Resulta altamente preocupante la tendencia creciente de consumo de alcohol en el país, tanto en adultos mayores como en jóvenes.

La OMS recomienda tres medidas para bajar el consumo nocivo de alcohol; reducir la capacidad de compra por menor disponibilidad, por ejemplo, reducir horarios y expendio en lugares públicos; reducir la exposición a publicidad e incrementar impuestos y precios.

Alimentación
En cuanto al consumo de frutas y verduras se observa que la mayoría de los encuestados consumen al menos una porción de frutas y/o verduras diariamente.

Se observó que no alcanzan a cumplir con las metas de la OMS sobre el consumo de frutas y verduras (5 porciones diarias equivalente a 450g), así también, que en general la población no consume alimentos fuera de su casa o alimentos procesados, por ser una población arraigada a sus tradiciones y costumbres caseras ellos consumen lo que producen y elaboran, además de encontrarse limitados económicamente por ser uno de los grupos etarios de más bajos recursos. A pesar de esto los adultos mayores tratan de consumir frutas y/o verduras de 5 a 7 días de la semana.

Actividad física
Los resultados de la presente encuesta mostraron que los adultos mayores de la ciudad del Carmen se desplazan caminando o en bicicleta por ser una zona medianamente urbanizada sin largas distancias para recorrer al movilizarse de un lugar a otro, realizan caminatas manteniendo una actividad física moderada.

Otra variable a considerar sobre el aumento del sedentarismo y la actividad física leve es la pandemia por SARS-Cov-2, debido a que muchos adultos mayores respetaron el aislamiento preventivo.

La inactividad física contribuye a la epidemia creciente de obesidad y expresa la necesidad de profundizar las políticas públicas para promover la actividad física en toda la población con propuestas deportivas y recreación en espacios comunitarios, acorde a los adultos mayores.

Sobrepeso y obesidad
La población adulta de la ciudad de El Carmen presenta en ambos sexos un IMC mayor al recomendado, lo que equivale a que la población presenta sobrepeso y obesidad, es decir un IMC mayor o igual a 25 kg/m^2.

La prevalencia de obesidad en adultos es mayor en hombres que en mujeres, al presentar una media de IMC de 29.30 kg/m^2. Esto puede estar relacionado con niveles de actividad física insuficiente.

Hipertensión arterial
En la ciudad de El Carmen, el 49,13% de los adultos mayores es hipertensa, definido como algún profesional le dijo que tiene la presión alta. La prevalencia es mayor en mujeres en un 61.18% que en hombres.

En la medición de presión arterial se observó que la población presentaba, en general, la presión arterial elevada sin conocimiento de padecer HTA, Hipertensión Arterial, siendo una afección silenciosa y que requiere de una búsqueda activa y rastreo sistemático.

Existe un grupo de personas con HTA bajo tratamiento que presentaron niveles normales de presión arterial en el momento de la medición.

No obstante, aún se requiere reforzar las intervenciones de captación de pacientes y mejorar la adherencia al tratamiento, ya que algunas personas que conocen su diagnóstico no siguen el tratamiento y el grupo que toma medicación, no llega a controlar su presión arterial.

Diabetes
El 21,97% de la población refirió que algún profesional le dijo que tiene diabetes, siendo mayor el porcentaje en mujeres (24.30%) que en hombres (18.18%).

A través de las mediciones bioquímicas de la encuesta de glucemia, el 24,86% de los adultos mayores presenta hiperglucemia y el 72,83% presenta una glucemia normal en ayunas, esta cifra no excluye al grupo de personas con diabetes bajo tratamiento que presentaron niveles normales de glucemia al momento de la medición.

Existe un grupo de adultos mayores que desconocen ser propensos a tener diabetes o que ya la padecen. Se observó al realizar las encuestas que muchos pacientes no se realizaban los controles médicos correspondientes. Es decir que, para reducir la prevalencia de diabetes, es fundamental implementar las políticas de prevención y control de la obesidad.

Colesterol y Triglicéridos
Se registró una media de 201,00 mg/dl de colesterol total en sangre, siendo la mayor media que presenta el grupo etario de 60 a 69 años. Se observa la mayor prevalencia de hipercolesterolemia en mujeres que en hombres, incluyendo a quienes toman medicamentos para el colesterol elevado.

A través de las mediciones bioquímicas de la encuesta, se observó que las mediciones de triglicéridos se comportan de igual forma.

La prevalencia de dislipemia (o alteración del perfil de lípidos) de la población requiere combinar múltiples variables. Es necesario aclarar que no puede concluirse que un valor aislado de colesterol total elevado significa que la persona tiene dislipemia, dado que la encuesta se trata de un estudio epidemiológico y no de una evaluación clínica.

Prácticas preventivas de cánceres prevalentes
En el grupo de mujeres de este estudio el 72.89% se ha realizado alguna vez una prueba de PAP, siendo que el 14.01% nunca se realizó y el 13.08% no recuerda haberse realizado el examen. Estas cifras muestran que la población se realiza los exámenes preventivos de cáncer de cuello de útero, pero así también se observa la falta de interés e información debido a que muchas mujeres se realizaron el PAP hace más de 5 años.

El tamizaje de cáncer de mama en mujeres muestra que alrededor del 30 % de que nunca o no recuerda haberse realizado una mamografía, así también como en el caso del tamizaje de cáncer de próstata, colon y recto, en ambos sexos es insuficiente, dada la limitada implementación de políticas y del modelo de atención de salud.

BIBLIOGRAFÍA

¿QUÉ ES LA ENCUESTA NACIONAL DE FACTORES DE RIESGO? - Ministerio de Salud de la Nación - Año 2005.

ASIS Año 2002 de la Provincia de Jujuy, realizado por la Dirección Provincial de Sanidad, presentado en la 1º Reunión Zonal de los Valles en junio del año 2003, en la Ciudad de Monterrico, Jujuy, Argentina.

ASIS Año 2004 del Área Programática V, realizado por el servicio de Epidemiología del Hospital Nuestra Señora del Carmen, presentado en la 3º Reunión Zonal de los Valles en Marzo del Año 2005, en la Ciudad de Perico, Jujuy, Argentina.

ASIS de Médicos Comunitarios por Puesto de Salud del Área Programática – Año 2018

Berríos Carrasola X. LA PREVENCIÓN DE LAS ENFERMEDADES CRÓNICAS NO TRANSMISIBLES DEL ADULTO. Boletín Esc. de Medicina, Universidad Católica de Chile 1994; 23: 53-60. http://escuela.med.puc.cl/paginas /publicaciones/ Boletin/html/Salud_Publica /1_13.html. Acceso 01/02/2019

Castellanos P. Sobre el concepto de salud-enfermedad. Descripción y explicación de la situación de salud. Boletín Epidemiológico. Organización Panamericana de la Salud. Vol. 10, Nº 4,1990. ISSN 0255-6669.

DETERMINANTES DE LA SALUD http://atencionprimaria.wordpress.com/ 2007/11/09/ - distintas-visiones/ determinantes de la salud

4° ENCUESTA NACIONAL DE FACTORES DE RIESGO – INFORME DEFINITIVO – Ministerio de Salud de la Nación - Año 2019 -

INDIVIDUOS ENFERMOS Y POBLACIONES ENFERMAS. Boletín Epidemiológico. Organización Panamericana de la Salud. Vol. 6, Nº 3, 1985. ISSN 0255-6669

Montenegro R. ATENCIÓN PRIMARIA Y MEDICINA FAMILIAR: SALUD Y MEDIO AMBIENTE. Acceso Julio 2019 en http://www.lulu.com/items/volume_36/560000/ 560065/ 1/print/560065.pdf

Publicación OMS:" PREVENCIÓN DE LAS ENFERMEDADES CRÓNICAS: UNA INVERSIÓN VITAL" .http://www.who.int/chp/chronic_disease_report/part1/ es/ índex. html, acceso 01/02/2019.

Silberman P. EL PROCESO SALUD - ENFERMEDAD, SUS DETERMINANTES. Acceso Febrero 2020 en http://www.fcm.uncu.edu.ar/medicina/posgrado/ inssjp/lectura /Modulo%202.3%20La%20salud%20y%20sus%20determinantes.

Las Condiciones de Salud de las personas mayores- Boletín Nº2 – 2017 UCA.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis de los cambios en los modelos dinámicos de procesos del comportamiento social durante la pandemia COVID19

Julián Pucheta[a,b], Martín Herrera[b], Carlos Salas[b], Daniel Patiño[c], Cristian Rodríguez Rivero[d]

[a] Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba.

[b] Facultad de Tecnologías y Ciencias Aplicadas, Universidad Nacional de Catamarca, Catamarca.

[c] Instituto de Automática - Universidad Nacional de San Juan, San Juan.

[d] Wiskunde en Informatica - Universiteit van Amsterdam, Amsterdam.

Datos de contacto: jpucheta@unc.edu.ar. Tel +54 351 5353800

RESUMEN

Se presenta un análisis de cambios de modelos de procesos dinámicos descriptos por variables que representan al comportamiento social desde el punto de vista de la movilidad de las personas y del de índices de la economía en el marco de la pandemia Covid19. Aquí se emplea como proxy del comportamiento social a la movilidad descrita por Google y Apple para relacionarlo con la evolución temporal de los infectados diarios de Covid19. Además, se emplea como proxy del proceso socioeconómico a índices relacionados con la economía global, donde se analizan dos de evolución ascendente (MSFT Microsoft y Nasdaq, Inc.) y otro con evolución estanca (WTI precio del galón de petróleo) y se relacionan con los contagios diarios de Covid19. En este último caso es complejo detectar una región territorial de influencia dada la cantidad de orígenes de influencias que tienen los índices seleccionados, pero sí se puede estudiar el impacto del primer pico en China y la posterior evolución en el mundo, en especial en nuestro país y en Holanda. Se concluye que el modelo subyacente ha cambiado en etapas distintas, según los meses del año y que luego de mediados de 2021 está en camino de alcanzarse un equilibrio inestable, con el agregado de las nuevas posibilidades que da el proceso de vacunación y las normas de convivencia social. Se concluye que deben analizarse qué decisiones conviene tomar a nivel social de política pública y qué decisiones personales por cada individuo.

Palabras Clave: Modelos dinámicos, Análisis estadístico, identificación de sistemas, procesos estocásticos.

INTRODUCCIÓN

Se presenta un análisis de procesos dinámicos descriptos por variables que representan al comportamiento social desde el punto de vista de la movilidad de las personas y del de índices de la economía. Aquí se emplea como proxy del comportamiento social a la movilidad descrita por Google [1] y Apple [2] para relacionarlo con la evolución temporal de los infectados diarios de Covid19 [3]. Además, se emplea como proxy del proceso socio-económico a índices relacionados con la economía global, donde se analiza uno de evolución ascendente (MSFT Microsoft [4] y Nasdaq, Inc. [5]) y otro con evolución estanca (WTI precio del galón de petróleo [6]) y se relaciona con los contagios diarios de Covid19 [3]. En este último caso es complejo detectar una región territorial a nivel país o continente de influencia dada la diversidad de componentes o actores que influyen en los índices seleccionados, pero sí se puede estudiar el impacto del primer pico en China y la posterior evolución en el mundo, en especial en Argentina y en Holanda. Puede observarse que la evolución ha sido en etapas distintas, según los meses del año y que luego de mediados de 2021 está en camino de alcanzarse un equilibrio inestable, con el agregado de las nuevas posibilidades que da el proceso de vacunación y las normas de convivencia social.

En la primera etapa de evolución de la pandemia, que incluye el intervalo temporal desde puede el 1º de enero del 2020 hasta el 30 de abril del 2020, donde gigantes tecnológicas se involucraron en la divulgación de información de movilidad de las personas, como Apple a partir del 17 de enero de 2020 y Google a partir del 15 de febrero de 2020. La protagonista pasa a ser la movilidad porque los Gobiernos implementaron reglas de movilidad estrictas para necesidades de movilizarse no esenciales. En ese período la movilidad disminuye a un mínimo. Mientras tanto, las empresas de laboratorios empezaron a coordinar esfuerzos y a principios de Mayo 2020 se hizo el primer llamado público a voluntarios que quieran probar la vacuna por parte de Pfizer [7]. Aquí termina la segunda etapa de evolución de la pandemia ya que, al aparecer, dada la expectativa por la vacuna, las variables socioeconómicas comenzaron a tener un comportamiento no correlacionado con los contagios.

En septiembre de 2020 empieza una segunda ola de contagios en Europa, que tuvo menos impacto que la primera etapa en las variables socio-económicas estudiadas. No obstante, se evidencia el cambio de la movilidad de las personas para el caso de Holanda que disminuyó respecto de la primera etapa.

En julio de 2021 se tiene una nueva ola de contagios, con nuevas variantes de Covid19, y el desarrollo de las variables socio económicas se ubicaron en caminos de normalizarse. Se observa que no hay impacto negativo en la economía y que la movilidad de las personas en Argentina evolucionó en el sentido de la normalidad. No obstante, en Holanda en la movilidad en la categoría de Parques se nota una merma respecto del período anterior, y en la de Transporte público. Analizando los contagios y la movilidad en Argentina, se puede pensar que la movilidad social está muy cerca de la normalidad, aunque los contagios no sean totalmente suprimidos.

Necesidad de modelos de procesos dinámicos

Un modelo es una representación simplificada de la realidad, y su utilidad es la posibilidad de tomar decisiones para que esa realidad evolucione de una manera deseada. La realidad se enfoca mediante procesos para acotar el problema, y se asume que tiene variables que modifican su evolución, y que ésta a su vez puede ser descripta por una variable variante en el tiempo. Aquí se propone un modelado para establecer la evolución de la pandemia a lo largo del tiempo y del espacio mediante relaciones matemáticas. Cada modelo tendrá en cuenta variables que son relevantes para la representación de la evolución de la pandemia, con el fin de establecer mecanismos que puedan mitigar el daño que ésta provoca en la sociedad. Las epidemias han sido objeto de estudio de la ciencia al menos desde 1760,

cuando apareció la propuesta de Bernoulli [8]. Luego, en 1911 se aplicó dicha propuesta para la prevención de la malaria [9]. A partir de 1927 se empezó a tomar el tópico desde la matemática por parte de Kermack [10] que formalizó resultados.

Modelo básico susceptibles infectados recuperados

Un modelo pionero es el de Kermack [10] orientado a la pandemia Covid 19 [11]. El modelo propone transiciones de un grupo susceptible a un grupo infectado, y de éste a un grupo recuperado,

(1) $$S_n \rightarrow I_n \rightarrow R_n$$

y propone que la suma de las funciones temporales sea una constante,

(2) $$S_n + I_n + R_n = N$$

donde N es el total de la población, y es constante, S_n representa a los susceptibles, I_n a los infectados, R_n a los recuperados, asumiendo que para n=0 se tiene que $S_0>0$, $I_0>0$ y $R_0=0$.

El ritmo de cambio de los susceptibles dada la transición a infectados, depende del tamaño de los susceptibles y del tamaño de los infectados [12], siendo los recuperados una fracción de los infectados. Así, para el caso de la evolución temporal diaria, es decir, cada 24 horas, se tiene,

(3) $$\begin{cases} S_{n+1} = S_n - \beta S_n I_n \\ I_{n+1} = (1 + \beta S_n - \gamma) I_n \\ R_{n+1} = R_n + \gamma I_t \end{cases}$$

con las condiciones iniciales para t=0 se tiene que $S_0>0$, $I_0>0$ y $R_0=0$. Además, ® es la velocidad de infección, que se define como el cociente

(4) $$\beta = \frac{\kappa p}{N}$$

siendo ⁄ el promedio de contactos por persona por unidad de tiempo y *p* es la probabilidad de contagio. A su vez, © es la velocidad de remoción y se define como

(5) $$\gamma = \frac{1}{D}$$

siendo D la duración media de la enfermedad.

Como acciones posibles para mitigar la pandemia, se puede observar que en general que In crece si ®S-©>0. Entonces, definiendo

(6) $$R_0 = \frac{\beta}{\gamma} S = S \frac{\kappa}{N} pD$$

se busca que $R_0 \leq 1$. Para ello, se pueden realizar diversas acciones, como por ejemplo reducir D con vacunas antivirales, y reducir la transmisibilidad p con medidas de higiene, barbijos, e incluso reducir κ con medidas de aislamiento entre los individuos.

En las Fig. 1 y Fig. 2 se muestran evoluciones obtenidas mediante este modelo. Nótese que para los parámetros establecidos el resultado a los 200 días de tiempo es diferente para cada caso. Para el primero, la totalidad de la población susceptible se termina infectando, y para el segundo se infecta hasta el 70% del total. Aquí se evidencia que modificando el distanciamiento β o agregando vacunas D se puede evitar la multiplicación de los infectados.

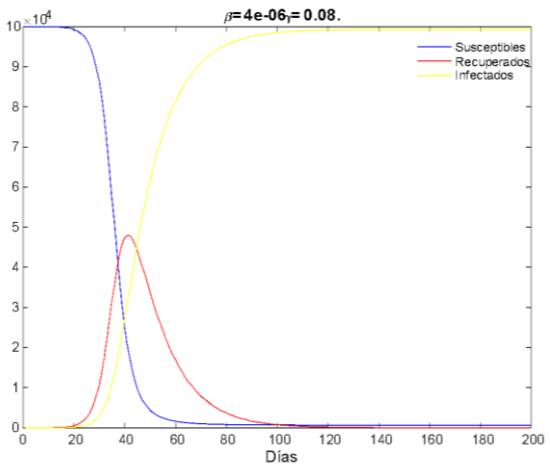


Fig. 1. Modelo SIR de pandemias. En 200 días todos los Susceptibles se infectan.

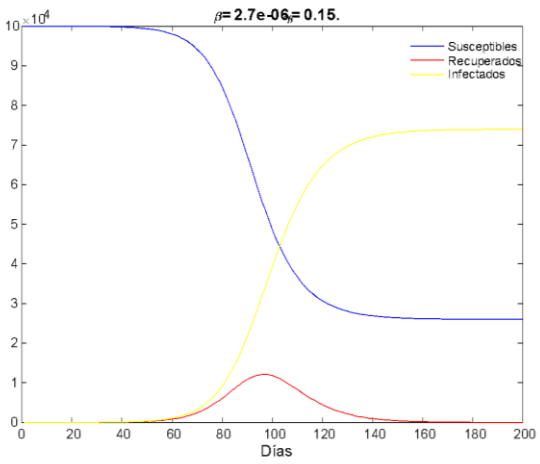


Fig. 2. Modelo SIR donde se infectan algunos de los susceptibles luego de 200 días de evolución.

A este modelo se le han agregado diferentes variables, como por ejemplo sujetos Expuestos, Fallecidos, testeos de sujetos asintomáticos y pre sintomáticos como en [13] donde se muestra su uso para un control automático de la pandemia. Incluso hay modelos que incorporan inteligencia artificial para realizar las proyecciones de los infectados y el paso a la nueva normalidad [14], y también sobre la situación global automatizada con un día de latencia en la actualización del tipo mapa de color [15], y de evolución temporal [16].

METODOLOGÍA

Se emplea la técnica de modelado para procesos dinámicos, y se comparan los diferentes resultados a partir de su desempeño y de su evidencia según la función de intercorrelación [17] [18]. Así, la conducta de las personas debería modificarse a partir de que se conoció la pandemia, con intención de mitigar los contagios que están relacionados con la cantidad de fallecidos. Por lo tanto, se puede establecer un patrón de conducta a partir de las indicaciones de movilidad, que debería cambiar respecto del paso de los días en el sentido de minimizar los contagios para evidenciar una conducta social responsable y virtuosa. Para ello se mide la diferencia entre el proceso real y el modelo, que debería ser siempre negativo, si se define como

$$e_{n+h} = \frac{\left(\hat{y}_{n+h} - y_{n+h}\right)}{y_{n+h}} 100 \tag{7}$$

donde se está comparando la salida real medida del proceso respecto de la salida entregada por el modelo dinámico a intervalos *h*, ajustada hasta un determinado tiempo, como máximo *n*. Si la conducta social fuese acorde al COVID19, entonces el valor siempre debería ser positivo y cuanto más grande sea, mayor es la mejora en la conducta social. Si se inspeccionan los datos, se evidencia que no hay rasgos de estacionalidad, pero sí de estacionalidad, tendencia y dinámica de las mediciones respecto de las entradas medidas, para lo cual se van a implementar técnicas para esta clase de procesos [17].

Propuesta de modelo dinámico

El esquema de modelado propuesto es simple, bien conocido en la literatura [17] siempre que las series temporales que representan a la entrada y salida del modelo tengan una intercorrelación diferente de la del ruido blanco [17], para realizar el cálculo como se muestra en la Fig. 3. Aquí se propone el uso del ajuste por el gradiente descendente, donde el error de ajuste está dado por

$$e_n = y_n - f(\theta_n, x_n, y_{n-1}) = y_n - \hat{y}_n \tag{8}$$

donde y_n es el valor actual de la salida del proceso, f(·) indica la salida del modelo que se está ajustando con parámetros θ_n a definir en cantidad y en valores, y x_n es el valor de la entrada actual. Se asume que se dispone de un total de datos muestreados con intervalos unitarios desde 0 hasta n.

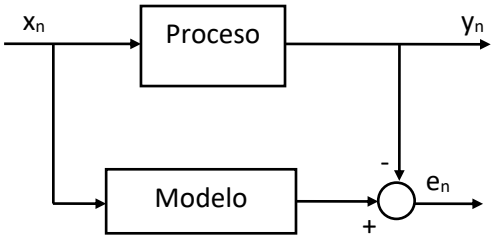


Fig. 3. Representación de la señal de error entre un proceso real y su modelo. El modelo tiene parámetros que son ajustados mediante la secuencia de error.

Aquí f() se implementa como un sistema dinámico lineal, definiendo al vector de los parámetros θ como

$$\theta^T = \begin{bmatrix} b_1 & b_2 & \cdots & b_{o_x} & a_2 & a_3 & \cdots & a_{o_y} \end{bmatrix} \tag{9}$$

se define que la función f() es

$$\hat{y}_n(\theta, x_n, y_{n-1}) = b_1 x_n + b_2 x_{n-1} + \ldots + b_{o_x} x_{n-o_x+1} - a_2 y_{n-1} - \ldots - a_{o_y} y_{n-o_y+1} \tag{10}$$

donde los parámetros serán ajustados en el sentido de minimizar el error (8) en donde interviene la medición actual y_n, que no aparece en (10).

Para el ajuste de dichos parámetros, se define el funcional a minimizar respecto de los parámetros como

$$J_n = \frac{1}{2}\sum_{n=n_i}^{n_f} e_n^2 = \frac{1}{2}\sum_{n=n_i}^{n_f} \{y_n - f(\theta, x_n, y_{n-1})\}^2 \tag{11}$$

donde están definidos dos tiempos específicos de ajuste que son n_i y n_f como tiempo inicial y tiempo final para el ajuste del modelo. Este ajuste se realiza una vez por cada iteración, durante la que se hará variar cada componente de $\backslash$. Los parámetros serán ajustados durante una etapa de tiempo, pero luego quedarán fijos para decidir la conducta social definido en (7).

Para ajustar los parámetros del modelo, que son las componentes a_i y b_i del vector θ, se procede empleando el método del gradiente descendente [17]. Para ello, se empieza minimizando a (11) respecto de dichas componentes hallando sus derivadas parciales, por lo tanto

$$\frac{\partial J_n}{\partial b_i} = e_n x_{n-i+1}, i = 1,2...o_x. \tag{12}$$

$$\frac{\partial J_n}{\partial a_i} = -e_n y_{n-i+1}, i = 2,3....o_y. \tag{13}$$

Con estos incrementos se definen las cantidades del algoritmo de gradiente descendente para el instante *n*, como

$$b_i := b_i + \gamma e_n x_{n-i+1}, i = 1,2...o_x. \tag{14}$$

$$a_i := a_i - \gamma e_n y_{n-i+1}, i = 2,3...o_y. \tag{15}$$

con el valor de γ como ganancia de ajuste o paso de ajuste. Con las variables {*x,y*} correctamente condicionadas, éste valor puede estar en el orden de 10^{-3} a 10^{-6} para los casos aquí estudiados. Por lo tanto, dados un conjunto de datos de mediciones hasta un instante *n*, $\{x_n,y_n\}$ se requiere elegir la relación de correspondencia entre la entrada *x* y la salida *y* de tal manera que el modelo muestre máxima evidencia. El algoritmo expresado en (14) (15) depende de las condiciones iniciales de θ, ya que es incremental, por lo que se proponen valores de orden reducido en dimensión y luego se procede a aumentar los valores de o_x, o_y si el modelo no muestra la suficiente evidencia. El criterio para medir la evidencia del modelo con el vector θ actual consiste en estudiar el comportamiento del error definido como

$$e_{cm} = \frac{1}{n_f - n_i}\sum_{n=n_i}^{n_f} e_n^2 \tag{16}$$

denominado error cuadrático medio del ajuste, y cambia luego de cada iteración del algoritmo. Por lo tanto, cuando la pendiente de cambio indica que e_{cm} ha aumentado de una iteración a otra, es porque el ajuste ya ha mostrado una evidencia de modelo desmejorada y corresponde detener las iteraciones. A partir de esa instancia, se puede realizar el cálculo con el modelo obtenido y la entrada x_n, con valores de *n* superiores al empleado en el cálculo del modelo, para comparar con los datos reales como indica (7). El criterio de evidencia de modelo aquí empleado es el de la inter correlación entrada salida [18]. Así, el modelo obtenido que presente la función de intercorrelación más parecida a la que presentan los datos originales, será el modelo elegido.

Evidencia de modelo

Para determinar cuál es el modelo que mejor describe la evolución del proceso, se emplea la relación de correspondencia entre dos series temporales conocida como intercorrelación [17] [18]. Se supone disponible una determinada cantidad de datos observados, por lo que existirán errores debido a que la cantidad es finita. Recordando que para el cálculo de la función de intercorrelación, se parte de la expresión

$$\phi_{xy}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x(t)y(t+\tau)dt \tag{17}$$

y para intervalos discretos unitarios se tiene

$$\phi_{xy}(j) = \frac{1}{N-j}\sum_{i=0}^{N-j} x(i)y(i+j) \tag{18}$$

con j=0, 1, …, M, y además i=0, 1, ….. N. donde M es el número de puntos de la función intercorrelación, N es el número de puntos medidos de la señal. Cabe aclarar que normalmente M es al menos la mitad de N, ya que la función intercorrelación está definida para N mucho mayor que M para asegurar una buena representación [18]. Nótese que para poder obtener un modelo que permita predecir el comportamiento del proceso subyacente, la función de intercorrelación entre las variables involucradas debe presentar una forma que el modelo debe respetar y seguir. De aquí la importancia de poder establecer una selección de variables que presenten una firma de intercorrelación posible de modelar.

Implementación en contagios diarios

Se emplearon los algoritmos propuestos para modelar la evolución de los contagios diarios, y el modelo ha ido cambiando a medida que pasaron las etapas. En la primera etapa, se obtuvieron evoluciones como se muestra en las Fig. 4, Fig. 5 y Fig. 6. Este tipo de modelado fue útil para determinar el cambio de conducta social y de allí tomar decisiones sobre las restricciones de movilidad en la sociedad. En la Fig. 4 se detalla la evolución proyectada y la

evolución real del número de contagios diarios. La discrepancia entre el número generado por el modelo y el medido es 14553-5352 o sea 9200 casos menos. Allí se consideró una movilidad de Apple con restricciones constantes al 20 de julio de 2020. Esta discrepancia es un indicador de que la sociedad estaba atenuando el efecto del Covid19 en un sentido virtuoso. Además, analizando las funciones de intercorrelación mostradas en las Fig. 5 y Fig. 6 se destaca que hay una discrepancia pero que está asociada a una parte pequeña de las series. A partir de este momento, puede generarse un nuevo modelo ya que la evidencia de modelo empieza a tener una discrepancia que irá en aumento.

Para el caso que se requiera mejorar el desempeño del modelo, sólo es necesario replegar el tiempo de la ventana de cómputo. El resultado para el nuevo proceso con su nuevo modelo subyacente, esto es la nueva conducta social respecto de la movilidad, se muestra en la Fig. 7 y en la Fig. 8 la evidencia de cada modelo. Aquí los resultados muestran un mejor ajuste, cuestión que se debe a que se disponen más datos que en el caso anterior, y el cambio de conducta presenta una tendencia.

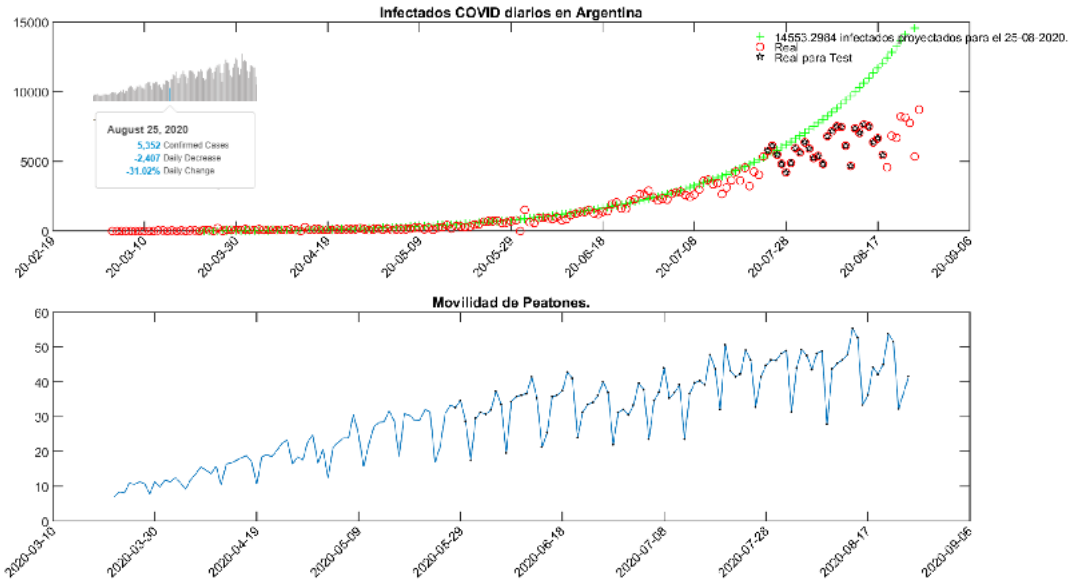


Fig. 4. Modelado del proceso de infectados en la primera etapa de la pandemia, para la república Argentina.

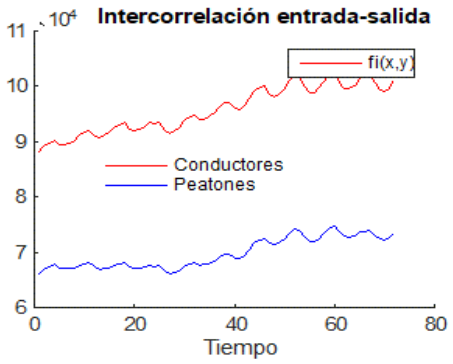


Fig. 5. Firma de la función intercorrelación definida en (18) entre la entrada medida y la salida medida.

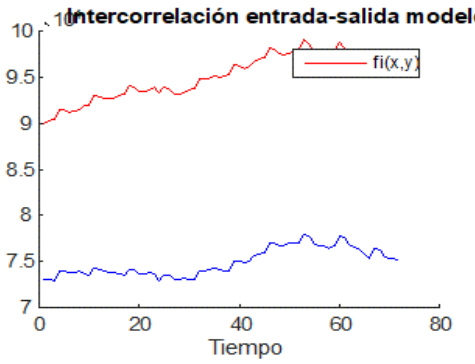


Fig. 6. Función Intercorrelación entre la salida del modelo y la entrada medida.

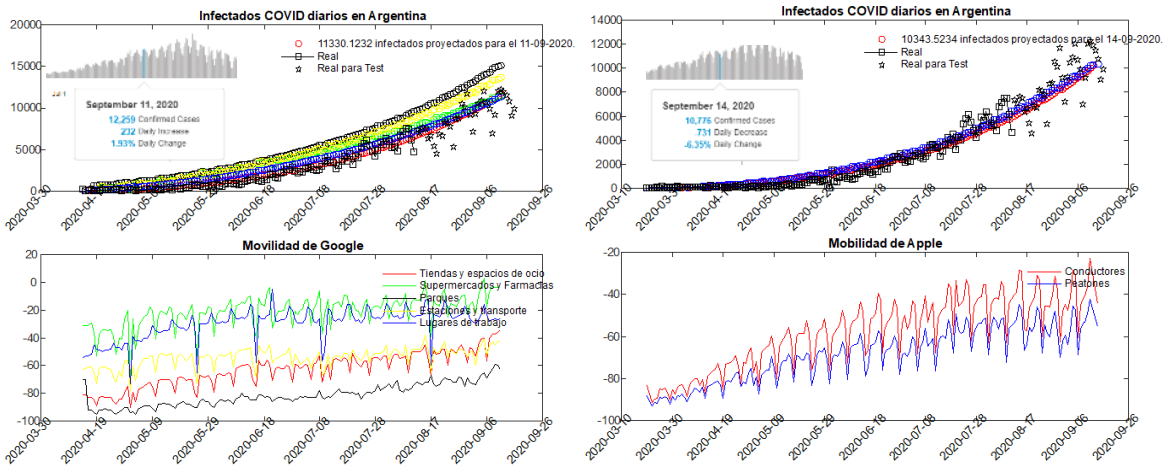


Fig. 7. Evolución de los valores dinámicos del cálculo según la movilidad de Google (derecha) y Apple (izquierda). Los datos empleados en el cálculo se indican con cuadrados y finalizan el 11 de agosto de 2020. Desde el 12 de agosto en adelante se hace empleo del modelo ajustado para predecir. Está superpuesto el valor dado por la OMS para ese día en cada caso.

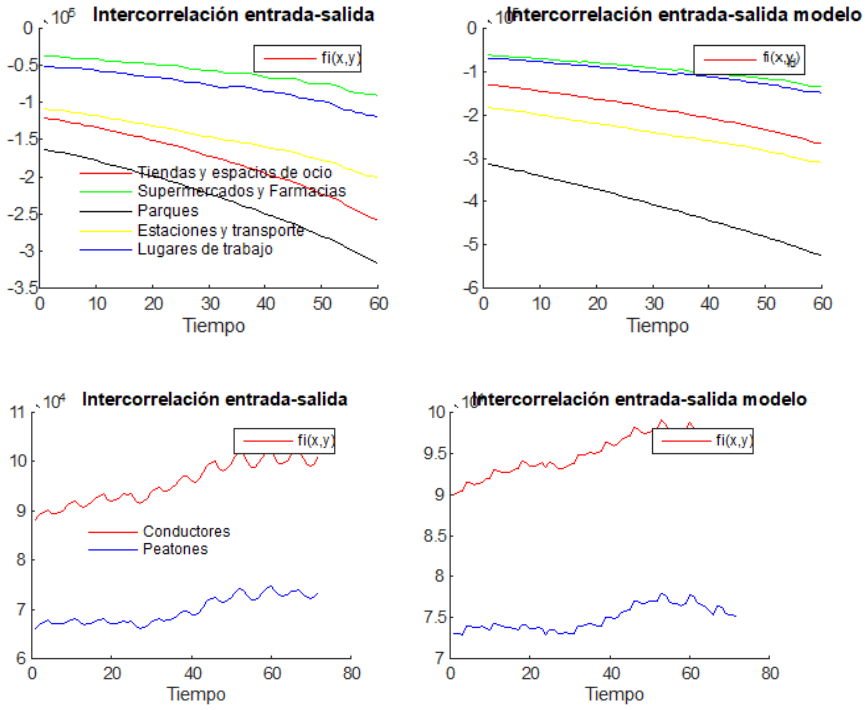


Fig. 8. Funciones de intercorrelación entre las variables dato y las identificadas, considerando como entradas a Google (arriba) y a Apple (abajo).

En virtud de las discrepancias mostradas en las figuras Fig. 4 y Fig. 7, se concluye que los contagios pueden modelarse a partir de la movilidad, pero el modelo va cambiando en el sentido de minimizar los contagios por lo que el algoritmo debe replegar su ventana de cálculo. Por lo tanto, es conveniente implementar esta clase de modelado para inferir el cambio en la correspondencia de infectados respecto de la movilidad asumiendo que las personas van adoptando conductas sociales que minimizan los contagios. Nótese que el proceso de vacunación comenzará en un período posterior y tendrá un impacto que modifica la correspondencia.

RESULTADOS OBTENIDOS

La evolución de los contagios es importante para poder determinar la evolución de las variables socio-económicas, teniendo como proxi de estos índices bursátiles como el precio del galón de petróleo [6], el precio de las acciones de Microsoft [4], y las acciones de Nasdaq Inc. [5] que incluye el desempeño de las empresas principales de dispositivos electrónicos. El análisis exige una evidencia de existencia de una correspondencia dinámica, y para ello se emplea el mismo método mostrado para el caso de la movilidad de personas. En la Fig. 9 se puede observar que hay una fuerte dependencia de los índices hasta abril de 2020 y está marcado el día 20 de abril de 2020. Allí termina una etapa de comportamiento social donde se evidencia un pánico y tendencia en liquidar los papeles por parte de los tenedores, llegan a abrir el galón de petróleo con valores negativos y un récord de volumen negociado quintuplicando la cantidad normal. En esa etapa se comenzó una búsqueda exhaustiva de la vacuna [19].

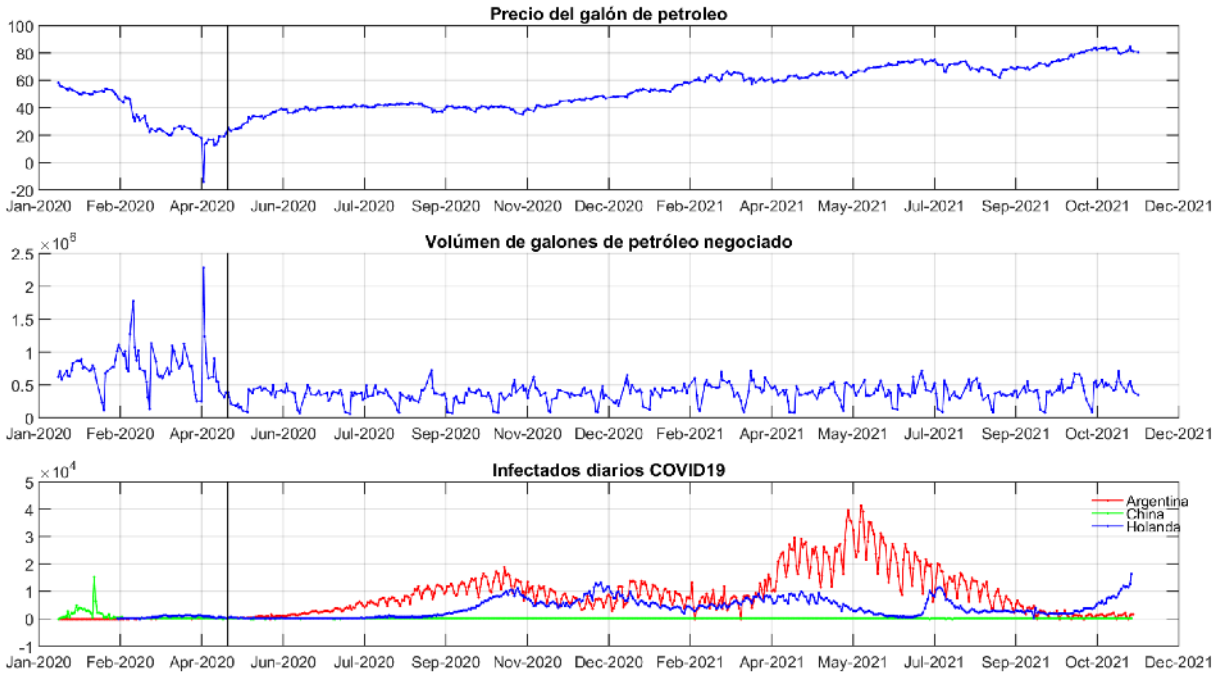


Fig. 9. Evolución de las variables socioeconómicas (Precio y volumen negociado de galón de petróleo) según los infectados diarios de Covid19. Se indica el 20 de abril de 2020.

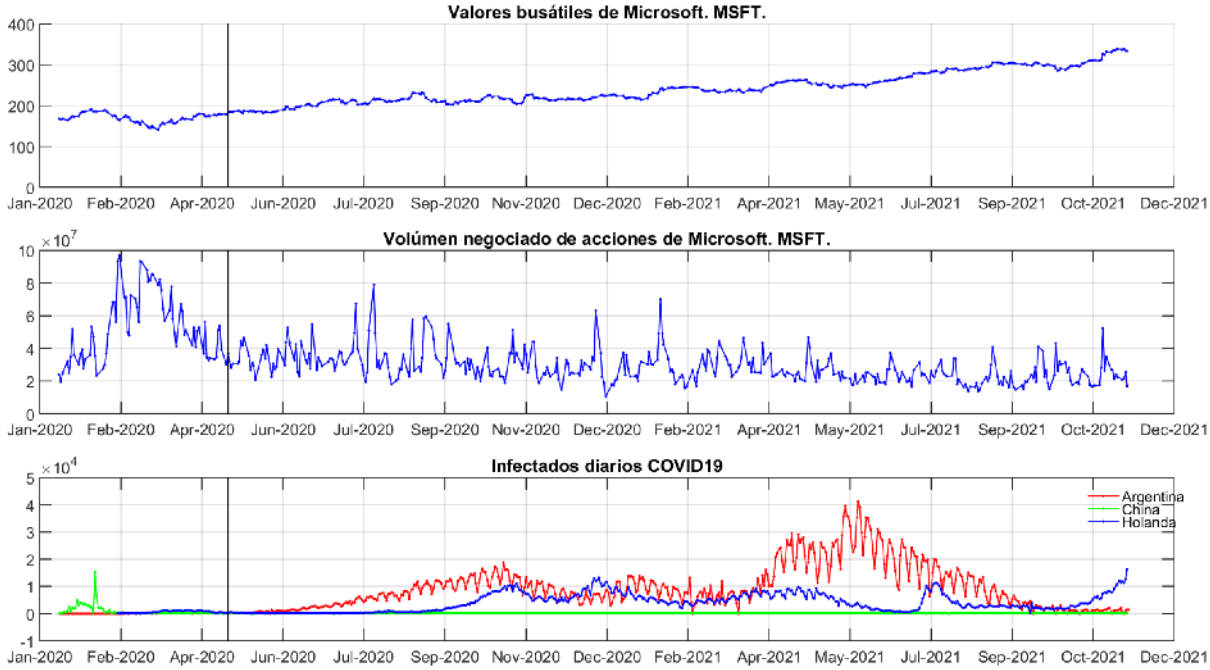


Fig. 10. Evolución de las variables socio económicas (Precio de cada acción y volumen negociado de Microsoft) según los infectados diarios de Covid19. Se indica el 20 de abril de 2020.

En el caso de buscar alguna actividad económica que se ha beneficiado con el impacto de la pandemia, se puede pensar en la evolución de las acciones de Microsoft y su volumen negociado respecto de los contagios diarios. En la Fig. 10 está la misma marca temporal de máxima inestabilidad que evidenció el precio del galón de petróleo en la Fig. 9. La variable describe un volumen negociado muy grande en los días previos pero la tendencia del valor es alcista. Aún con las nuevas olas de infectados de Covid19 la tendencia continuó en alza. Esta evolución indica que la actividad socioeconómica ha logrado convivir con la pandemia e incluso ha mejorado su estabilidad en cuanto al volumen negociado y la tendencia siempre alcista del valor.

Otra actividad económica que se ha beneficiado con el impacto de la pandemia es la de servicios financieros, como por ejemplo Nasdaq Inc. [5]. En la Fig. 11 se muestra la evolución temporal de los valores diarios por acción, que no es mucho mayor que la de MSFT, pero nótese que el volumen negociado comenzó a incrementar su nivel antes de la marca del 20 de abril de 2020 y luego permaneció alta.

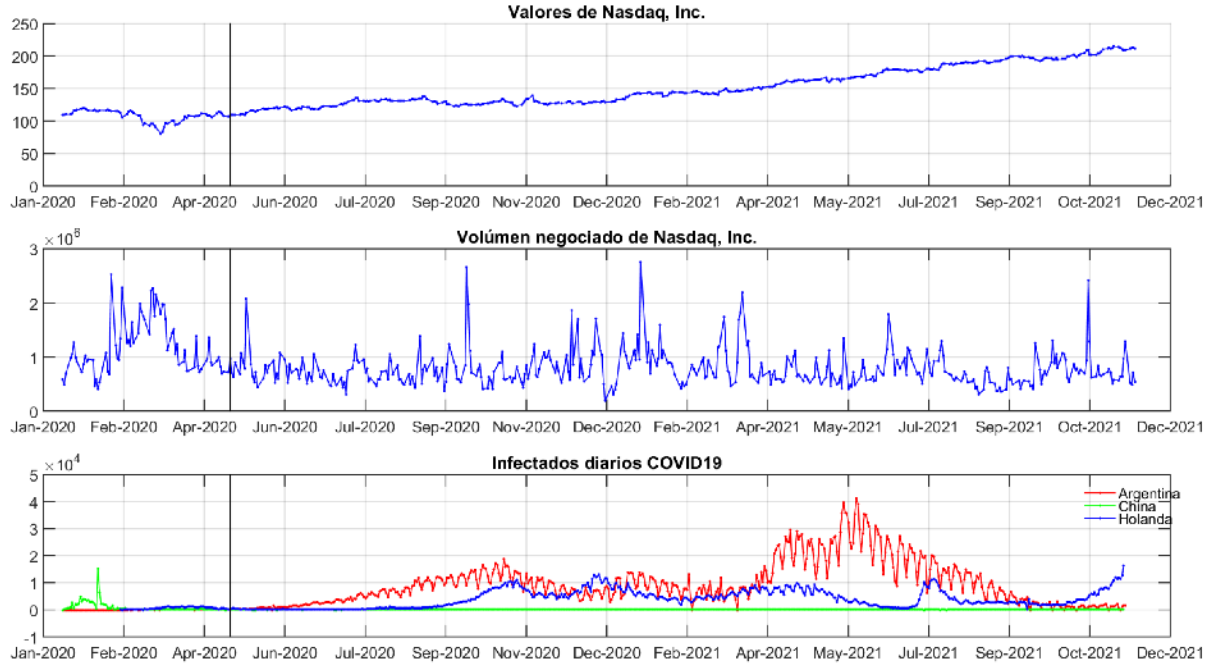


Fig. 11. Evolución de las variables socio económicas (Precio por acción y volumen negociado de Nasdaq, Inc.) según los infectados diarios de Covid19. Se indica el 20 de abril de 2020.

En los casos estudiados se observa que a partir del 20 de abril de 2020 la inestabilidad permanece un tiempo determinado, Esto es porque en mayo de 2020 apareció la convocatoria de los principales laboratorios para probar la vacuna para Covid19 en humanos [7]. Eso generó una expectativa en las personas y la pandemia fue tratada con otra óptica, según evidencian las variables mostradas en las Fig. 9, Fig. 10 y Fig. 11 a partir de junio de 2020. Este razonamiento tiene evidencia estadística si se analiza la Fig. 12.

En la Fig. 12 se detalla el cálculo de la función intercorrelación expresada en (18) entre las series temporales de los infectados diarios de Covid19 en China y las variables socioeconómicas WTI, MSFT y Nasdaq Inc. Están separadas temporalmente en las cuatro etapas mencionadas, con algunos días de solapamiento. En la primera gráfica se puede observar que la evolución de la función (18) es muy distinta cuando se observa a la WTI (color negro) respecto de las otras dos. La serie WTI indica una gran inestabilidad (pánico) respecto del comportamiento de las personas responsables de establecer el precio según la oferta y demanda, evidenciando una oferta pocas veces vista y un volumen negociado 5 veces superior al normal en el día del quiebre que fue el 20 de abril de 2020. En la segunda etapa, el comportamiento de las funciones (18) está detallado en la gráfica de la derecha arriba de la Fig. 12. Para este caso, se tiene que las tres series tuvieron una tendencia a la baja, con un comportamiento errático, pero en aumento del número de contagiados de Covid19, por lo que el modelo dinámico es de fase no mínima, es decir, inestable.

En la tercera etapa, mostrada en la gráfica de abajo a la izquierda de la Fig. 12, se destaca que el comportamiento del proceso ha sido similar y convergente a una independencia entre las variables, aunque hay una firma que evidencia un modelo dinámico estable.

En la última etapa, se destaca que las intercorrelaciones tienen una firma similar y el modelo dinámico subyacente es estable, por lo que puede concluirse que los efectos de los contagios en China no afectan a las variables socio-económicas de una forma sostenida y nociva como sí lo fue en la primera etapa.

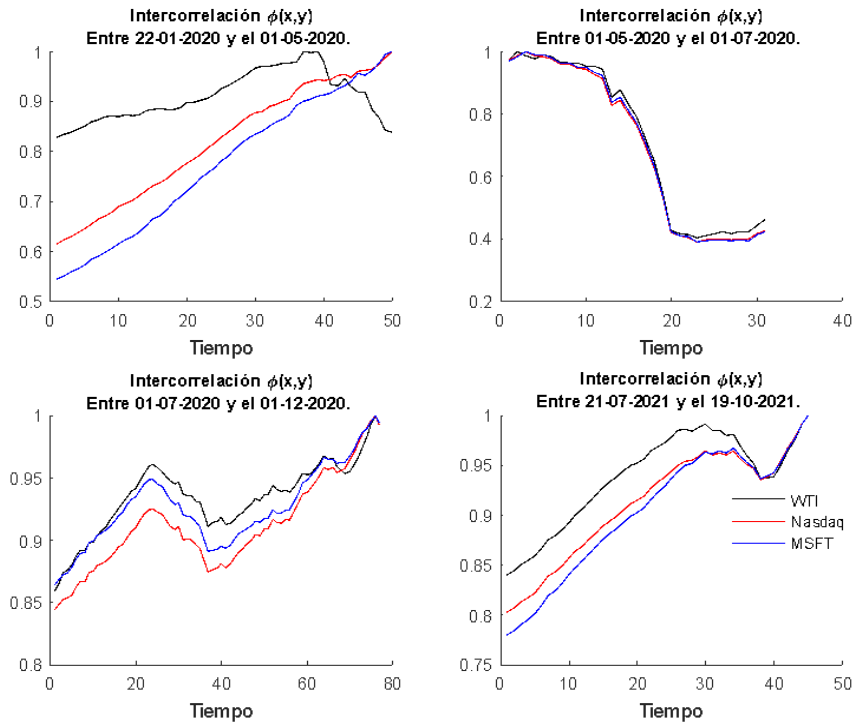


Fig. 12. Intercorrelación entre las variables socio-económicas y los infectados diarios de Covid19 en China. Cada una es calculada mediante la expresión (18), y normalizada para que su rango sea unitario.

Etapa de vacunación masiva mundial

El proceso de vacunación en el mundo comenzó a fines de 2020 en voluntarios [7], y se masificó a principio de 2021 [20]. No obstante, el impacto en el número de contagios diarios fue diferente según el país. En la Fig. 13 se muestra la evolución en Argentina, donde se destaca el aumento de contagios diarios que existió cuando comenzó el proceso de vacunación. Ese efecto también aparece en Holanda, como se ve en la Fig. 14, donde un brote de contagios aparece luego de finalizar de vacunar a su población.

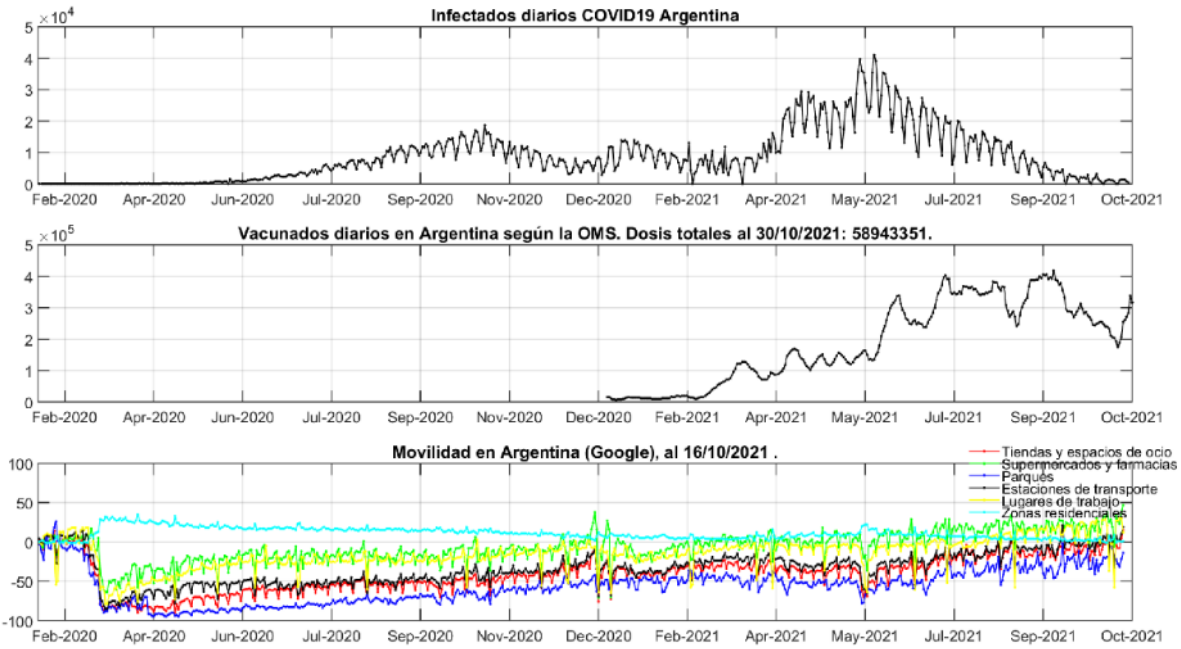


Fig. 13. Evolución de los infectados diarios, vacunados y movilidad de las personas según Google en Argentina.

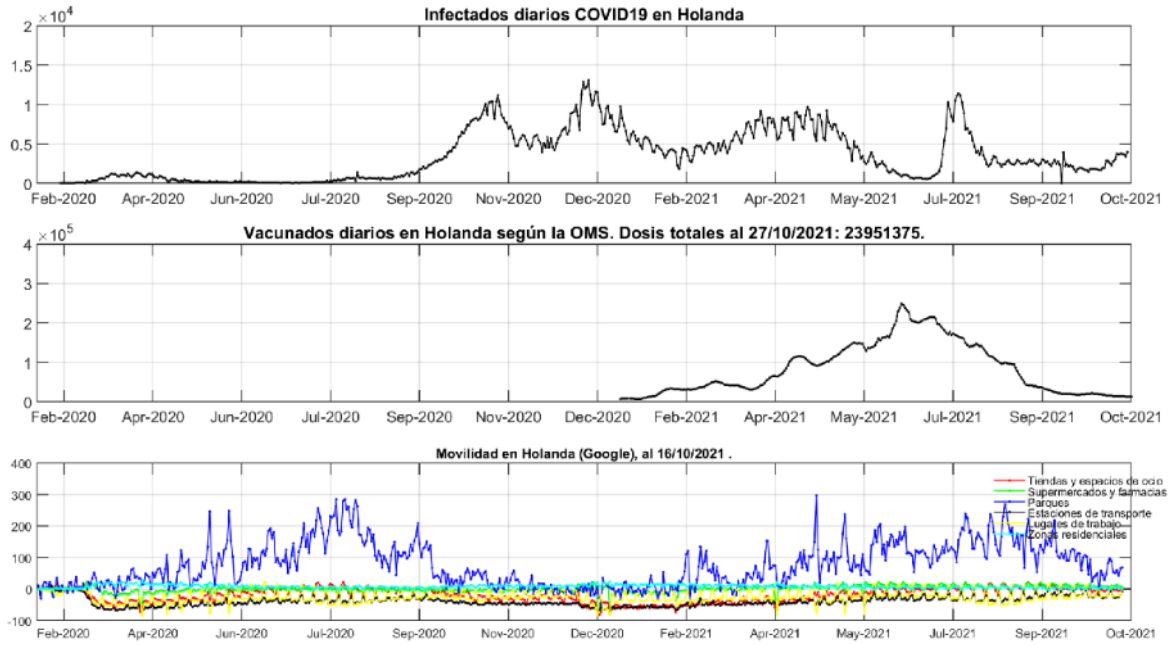


Fig. 14. Evolución de los infectados diarios, vacunados y movilidad de las personas según Google en Holanda.

Variables socio-económicas antes y después de la pandemia

Para finalizar el análisis, en la Fig. 15 se puede observar la evolución en el período previo al primer brote de Covid19 de las variables bursátiles mencionadas que tienen una tendencia similar a la que tenían antes de que comience la pandemia. Incluso la variable que ha sido afectada en el sentido desfavorable, que es el precio del galón de petróleo, ha llegado a su precio nominal medio anual, y lo ideal es que quede rondando por allí dependiendo fundamentalmente de las estaciones climáticas en Europa. No obstante, observando la dinámica que tiene el volumen negociado, se evidencia la inestabilidad del desde el 1 de marzo hasta el 20 de abril de 2020, hecho que no tiene antecedente en el periodo analizado.

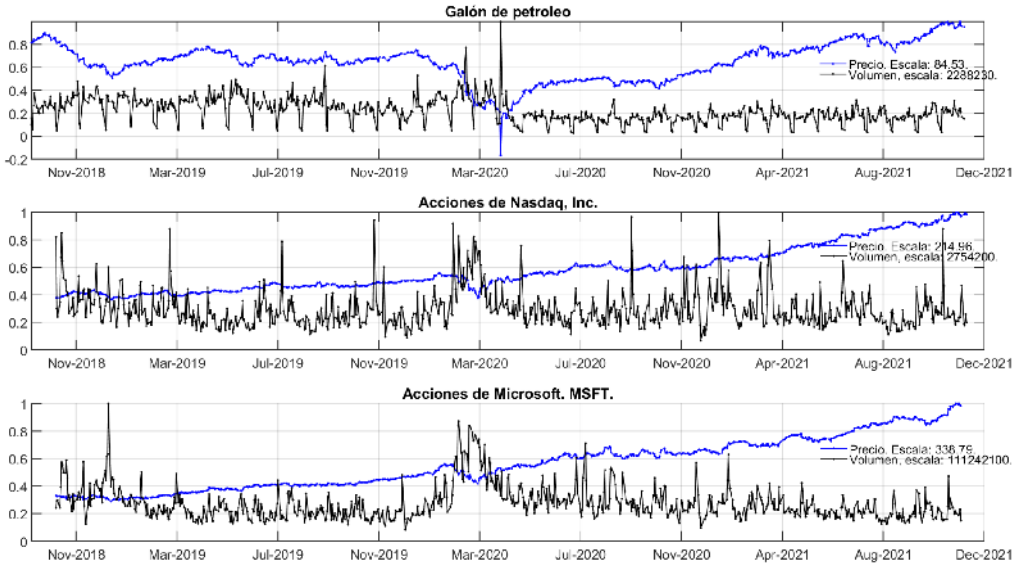


Fig. 15. Variables socio económicas de empresas que han sido impactadas por la pandemia en abril de 2020.

CONCLUSIONES

En conclusión, a partir de la evidencia que los indicadores económicos mostrados tienen tendencia a la recuperación, y en algunos casos acercándose a la normalidad, se puede decir que los efectos de la pandemia han sido minimizados. El cambio en la conducta social, y en los mecanismos de producción de tecnología para combatirla ha ayudado a que la sociedad tenga expectativa positiva ya que el virus está en camino de la contención. Con respecto a la contribución del análisis al soporte en la toma de decisiones, se tienen dos aspectos a considerar. Un aspecto de política pública en la administración de los recursos para la sociedad, donde debe enfocarse en la inversión en la tecnología de la información y comunicaciones, además del fortalecimiento de las estructuras de salud. El otro aspecto es el individual, donde cada persona no puede dejar de usar las barreras que son el barbijo, higiene personal y distanciamiento o evitar las concentraciones de personas en ambientes no ventilados.

BIBLIOGRAFÍA

[1] https://www.google.com/covid19/mobility/

[2] https://covid19.apple.com/mobility

[3] https://covid19.who.int/

[4] https://finance.yahoo.com/quote/MSFT

[5] https://finance.yahoo.com/quote/NDAQ/

[6] https://finance.yahoo.com/quote/CL%3DF

[7] Primeras pruebas en voluntarios. 5 de Mayo de 2020. https://www.pfizer.com/news/press-release/press-release-detail/pfizer_and_biontech_dose_first_participants_in_the_u_s_as_part_of_global_covid_19_mrna_vaccine_development_program

[8] Bernoulli, D. (1760). Esai d' une nouvelle analyse de la mortalité causée par la petite vérole, et des advantages de l'inoculation pour la prévenir. Mém Math. Phys.Acad. Roy. Sci., 1-45.

[9] Ross R. The prevention of malaria. London: John Murray; 1911. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2302055/pdf/brmedj07224-0010.pdf

[10] Kermack, W. O. y McKendrick, A.G. (1927). "Contributions to the Mathematical Theory of Epidemics". Proc. Roy. Soc. A. vol. 115, pp. 700-721.

[11] Amster, Pablo. https://www.youtube.com/watch?v=7XKPydEnDjA&feature=youtu.be

[12] Hamer, W.H. (1906) The Milroy Lectures on Epidemic Disease in England—The Evidence of Variability and Persistence of Type. The Lancet. Volume 167, Issue 4305, 3 March 1906, Pages 569-574. https://doi.org/10.1016/S0140-6736(01)80187-2.

[13] H. D. Patiño, S. Tosetti, J. Pucheta and C. R. Rivero, "Control of COVID-19 Outbreak for Preventing Collapse of Healthcare Capacity based on Social Distancing, Confinement and Testing-Quarantining", 2020 IEEE Congreso Bienal de Argentina (ARGENCON), 2020, pp. 1-6, doi: 10.1109/ARGENCON49523.2020.9505448. (2020).

[14] https://covid19-projections.com/path-to-herd-immunity/

[15] https://globalepidemics.org/key-metrics-for-covid-suppression/

[16] J. Pucheta, C. Salas, M. Herrera, H. D. Patiño and C. R. Rivero, "Análisis y Modelado de Procesos Dinámicos para Medir el Cambio de Conducta Social en el Marco del COVID-19", 2020 IEEE Congreso Bienal de Argentina (ARGENCON), 2020, pp. 1-6, doi: 10.1109/ARGENCON49523.2020.9505520. (2020).

[17] Ljung, Lennart. System Identification: Theory for the User (2nd Edition). Prentice-Hall. 1999.

[18] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. Discrete-Time Signal Processing. Upper Saddle River, NJ: Prentice-Hall, 1999.

[19] Principales laboratorios colaboran en la búsqueda de la vacuna para el Covid19.

https://cincodias.elpais.com/cincodias/2020/03/23/companias/1584965257_070426.html

[20] Mathieu, E., Ritchie, H., Ortiz-Ospina, E. *et al*. A global database of COVID-19 vaccinations. Nat Hum Behav (2021)

[21] https://finance.yahoo.com/

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Aplicación de un Análisis de Procrustes Generalizado para evaluar el estado de plantaciones de dos años de tres variedades de eucaliptos en dos microambientes de La Esperanza, provincia de Jujuy

Juan Manuel Solís[1,2], Santiago De Tellería[2], Agustín Montenegro[3], Julián Quispe[2]

[1]Facultad de Ciencias Agrarias, Universidad Nacional de Jujuy, San Salvador de Jujuy.
[2]Tecnicatura Universitaria Forestal, Universidad Nacional de Jujuy, sede San Pedro de Jujuy
[3]Universidad Nacional de la Plata

Datos de contacto: juanmasolis@fca.unju.edu.ar. Tel. 3884 389632

RESUMEN

El Análisis de Procrustes Generalizado (GPA por sus siglas en inglés) constituye un método de análisis multivariado muy conveniente para el análisis de variabilidad multiambiental, ya que permite el ajuste de matrices bidimensionales a configuraciones parciales que luego conforman una matriz de consenso que ajuste mejor a la estructura intrínseca de la variabilidad de los fenómenos observados.

En este trabajo se realizó un GPA sobre plantaciones de dos años de tres genotipos de eucalipto en dos microambientes o lotes, a fin de determinar la contribución relativa que sobre la variabilidad total observada tuvieron el genotipo y el ambiente, como una herramienta para la toma de decisiones.

De esta forma, se pudo identificar cuáles fueron los genotipos más “estables” *entre* ambientes, cuáles los más “variables”, y aquellos que aportaron mayor variación *dentro* de cada ambiente.

Palabras Clave: Análisis de Procrustes Generalizado, La Esperanza, Eucalipto, FactoMineR

INTRODUCCION

El Análisis de Procrustes Generalizado (GPA, por sus siglas en inglés) fue introducido por Gower (1975) como una técnica estadística multivariada para analizar matrices de datos tridimensionales. Este método permite abordar el análisis de datos provenientes de un conjunto de objetos o individuos en diferentes condiciones experimentales (ambientales o temporales), con un arreglo en "matrices o tablas multivías", en las cuales cada dato es originado por tres modos o vías: individuos x variables x condiciones (Del Médico, 2015), aunque en los últimos años se han desarrollado métodos para analizar más de tres modos.

Commandeur (2016), describe el principio de GPA de la siguiente manera:

Supongamos que tenemos dos configuraciones X_1 y X_2, cada una de las cuales conteniendo las coordenadas de los mismos *p* estímulos en *dos* dimensiones.

Además, supongamos que los puntos asociados a los estímulos en X_1 y X_2 son dibujados en una hoja transparente. El problema a resolver por un GPA es: ¿cómo mover estas dos hojas, superpuestas una con otra, de forma tal que la suma cuadrática de las distancias entre cada par de puntos sea la menor posible?

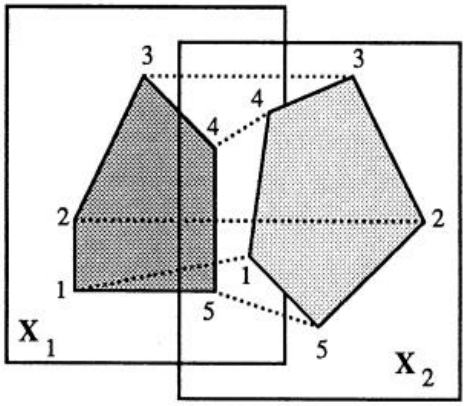


Figura 1. Configuraciones hipotéticas en dos dimensiones.

Para minimizar la suma cuadrática de las distancias entre cada par de puntos, las "hojas transparentes" pueden ser manipuladas de varias formas. Sin embargo, hay una restricción importante: únicamente consideramos transformaciones que mantengan las distancias relativas entre puntos **dentro** de cada configuración.

Las transformaciones que cumplen con la restricción impuesta pueden ser:

1. *Traslación.*
2. *Reflexión.*
3. *Escalamiento.*
4. *Rotación*.

Ninguna de las transformaciones anteriores modifica las relaciones internas entre los puntos de cada configuración. La estimación de la función de mínima distancia cuadrática se base en el siguiente teorema:

"Dados *n* puntos en un espacio *m-dimensional* la suma del cuadrado de las distancias entre los *n* puntos entre sí, equivale a *n* veces la suma del cuadrado de las distancias entre los *n* puntos y su *centroide*."

El teorema anterior puede ser expresado algebraicamente de la siguiente de la siguiente manera: sea x_j el vector columna que contiene las coordenadas del punto *j* (j= 1,…,n), en un espacio m-dimensional, y z un vector columna de coordenadas del centroide de los *n* puntos, es decir, $\mathbf{z} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x_j}$, entonces

$$\sum_{j<k} (\mathbf{x_j} - \mathbf{x_k})'(\mathbf{x_j} - \mathbf{x_k}) = \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{x_j}\,'\mathbf{x_j} - \mathbf{n^2 z'z}$$

Obtenidas las transformaciones de las configuraciones originales, se obtiene la **configuración de consenso** como el promedio de las anteriores.

La transformación Procrustes (escalamiento, rotación y traslación), en términos matriciales, pueden ser expresados del siguiente modo:

$$\mathbf{Y_k} = \rho_k \mathbf{C_k H_k} + \mathbf{T_k}$$

Donde $\mathbf{Y_k}$ representa la transformación de procrustes, ρ_k el factor de escala, $\mathbf{C_k}$ es una configuración resultante de un Análisis de Componentes Principales (ACP) aplicado sobre una table $\mathbf{X_k}$ de datos conformadas por *n* filas (individuos) y *p* columnas (variables), $\mathbf{H_k}$ la matriz ortogonal de rotación de dimensión *pxp* y $\mathbf{T_k}$ la matriz traslación de dimensión *nxp* (Del Médico, 2015).

Los tres últimos elementos son encontrados minimizando la Suma de Cuadrados Residuales (SCR):

$$SCR = \sum_{j=1}^{n}\sum_{k=1}^{K} \Delta^2\,(\mathbf{P_i^{(k)}}, \mathbf{G_i})$$

donde $\Delta^2(P_i^{(k)}, G_i)$ es la distancia euclídea entre el punto $P_i^{(k)}$ y el centroide de las *k* puntos análogos de $P_i^{(k)}$, llamado G_i (Del Médico, 2015).

A partir de 2014, el Ingenio La Esperanza (provincia de Jujuy) decidió activar el área forestal como actividad complementaria a la caña de azúcar con una ampliación en la producción de eucaliptos. A partir de esta iniciativa, se probaron distintas variedades de eucaliptus y técnicas de plantación, incorporando material genético de buenos resultados en otras regiones del país.

Los sitios plantados fueron zonas abandonadas por la caña de azúcar, potreros ganaderos abandonados, o tierras habilitadas no aptas para la producción de la caña de azúcar, con lo cual se hace necesario contar con herramientas que permitan analizar y caracterizar diferentes genotipos como respuesta a variaciones ambientales a micro – escala, a fin de identificar aquellos con mayor aptitud o capacidad de adaptación.

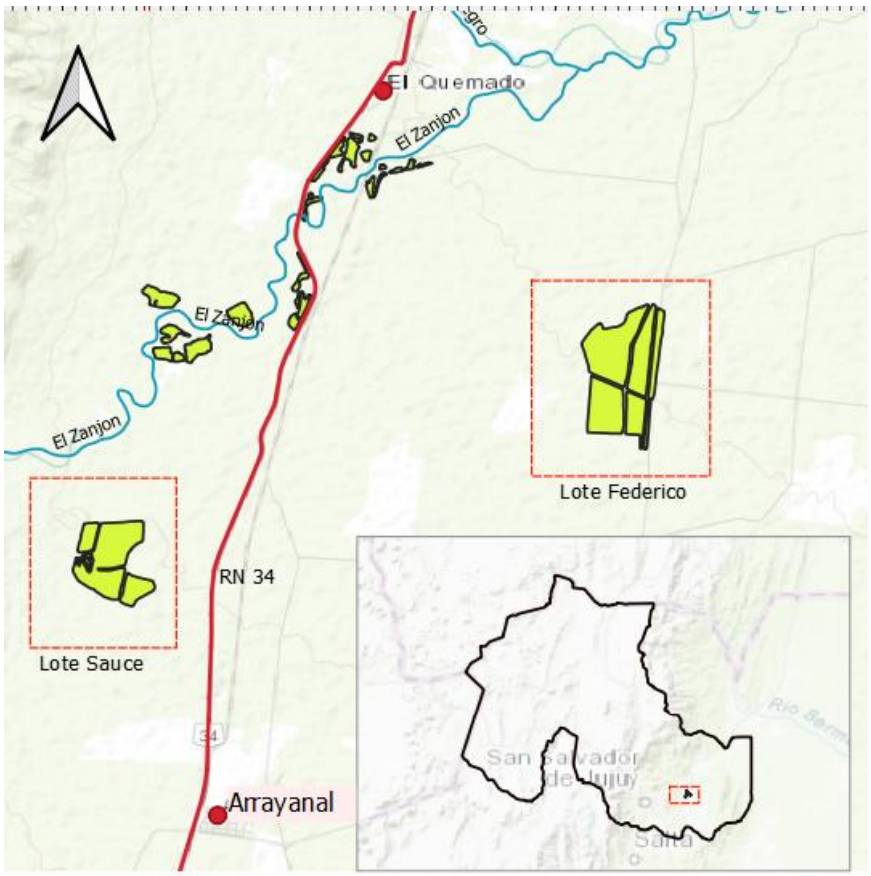
En este trabajo se realizó el análisis GPA sobre diferentes variables medidas en tres variedades y/o híbridos de eucaliptos en plantaciones de dos años realizadas en dos microambientes o lotes de plantación pertenecientes al Ingenio La Esperanza.

METODOLOGÍA

Los datos analizados provienen de un trabajo de monitoreo de plantaciones forestales de eucaliptus realizados en agosto de 2017 en dos microambientes o lotes de plantación pertenecientes al Ingenio la Esperanza (San Pedro de Jujuy), lote *Federico* (86 ha) y Lote *Sauce* (45 ha), con las variedades *Eucaliptus camaldulensis*, *Eucaliptus saligna* y el híbrido *Eucaliptus grandis x camaldulensis*, todos plantados en 2015 con un marco de plantación 3 x 3 (Figura 2).

La muestra incluyó 3.686 plantas. Para el muestreo se realizaron transectas diagonales delimitadas por caminos o cortafuegos.

Figura 2. Ubicación de los lotes Sauce y Federico



Las variables registradas en dicho monitoreo sobre cada planta fueron:

1. *Diámetro a la altura del pecho* (DAP), cuyo valor fue utilizado para estimar el *área basimétrica.* Se utilizó un calibre con vernier que se dispuso de forma perpendicular al eje del tallo a una altura de 1,3 m.
2. *Altura total de la planta*.
3. *Estado sanitario:*
   a. Sano (s)
   b. Muerto (m)
   c. Falla (f): cuando no se encontró el árbol
   d. Decrépito (de): con síntomas de marchitez

Además de las variables detalladas para la caracterización del estado sanitario, se reconocieron otras variables que no se incluyeron en el análisis.

A partir de las mediciones anteriores, se construyeron las siguientes variables para cada lote y variedad/híbrido:

1. Área basimétrica por ha en m^2/ha, obtenida sobre individuos no muertos (AB).
2. Altura promedio en metros de árboles no muertos (H).
3. Porcentaje de individuos sanos (S).
4. Porcentaje de individuos muertos (M).
5. Porcentaje de fallas (F).

6. Porcentaje de individuos con síntomas de marchitez (DE).

La variabilidad asociada a cada variedad de eucalipto y/o ambiente (lote) fue analizada por medio de un análisis de Procrustes Generalizado empleando la función *GPA* de la librería *FactoMineR* del programa estadístico R (versión 3.6.1), la cual considera casos perdidos.
Los argumentos de esta función son los siguientes:

- "DF": es un dataframe con *n* filas y *p* columnas (variables cuantitativas).
- "Tolerance": umbral de tolerancia por debajo del cual el algoritmo se detiene (cuando la diferencia entre la función de pérdida de GPA en el paso n y n + 1 es menor que la tolerancia).
- "nbiteration": número de iteraciones (por defecto 200).
- "scale": si los datos son escalados (por defecto TRUE).
- "group": vector con el número de variables en cada grupo.
- "name.group": vector con el nombre de cada grupo.
- "graph": si realiza el gráfico de la matriz de consenso (por defecto TRUE).
- "axes": selección de dimensiones (equivalente a componentes principales).

En este caso se trabajaron con tres filas (una por cada variedad/híbrido) y doce columnas (seis variables agrupadas en dos microambientes o lotes, Tabla 1).
Dado que las variables presentaron diferentes escalas de medición, se trabajó con las mismas de forma escalada.
Los grupos estuvieron constituidos por cada uno de los microambientes o lotes: *Federico* y *Sauce*.
Para analizar las sumas de cuadrados, se recurrió al atributo "PANOVA" o Análisis de la Varianza del objeto procrustes. Por ejemplo, es posible analizar en qué medida cada variedad aportó variabilidad sobre los datos observados. Además, se realizó una lectura e interpretación del gráfico biplot correspondiente.

DESARROLLO

En la Tabla 1 pueden observarse los valores obtenidos de cada variable, para cada variedad según lote. En líneas generales, *E. grandis x camaldulensis* presentó los mayores valores de área basimétrica por ha y altura media en ambos microambientes.

E. camaldulensis presentó los mayores valores de porcentaje de individuos sanos y los menores de mortalidad y falla.

E. saligna presentó un comportamiento muy variable, con un alto porcentaje de falla en lote Federico y un alto porcentaje de individuos sanos en lote Sauce. Además, presentó un valor relativamente bajo de área basimétrica en lote Federico, pero alto en lote Sauce.

Tabla 1. Tabulación de los valores de las variables analizadas por variedad/híbrido de eucalipto según microambiente o lote.

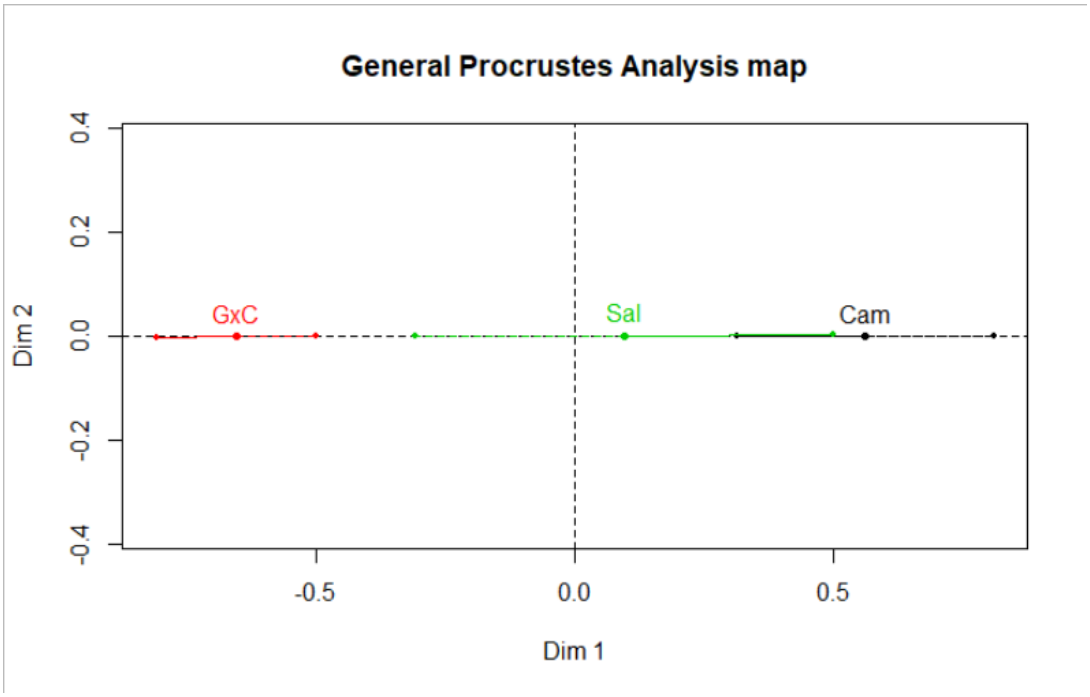
Variedad/ híbrido	Lote Federico						Lote Sauce					
	AB	H	S	M	F	DE	AB	H	S	M	F	DE
	m^2	m	%				m^2	m	%			
E. camaldulensis	1,044535	4,25	53,3	0,55	41,21	4,95	0,157654	2,06	93,06	0	0	0
E. grandis x camaldulensis	2,728826	7,25	28,35	8,66	41,73	0,79	4,881905	7,44	55,83	1,48	22,16	0,44
E. saligna	0,766547	4,37	21,98	4,16	64,22	2,64	4,193512	6,81	65,51	4,3	23,27	2,11

Nota: para el análisis del estado sanitario se consideraron únicamente las variables S, M, F y DE. Es posible que la sumatoria de los valores de estas columnas no resulte 100, en virtud de otras variables que no se tuvieron en cuenta para el análisis.

Una vez creado el objeto “procrustes” en R, se construyó un biplot de las dos dimensiones principales de la matriz de consenso, incluyendo los puntos correspondientes a cada configuración transformada (Figura 3). En este gráfico es posible reconocer cada punto $\mathbf{G_i}$ como el centroide de los correspondientes puntos $\mathbf{P_i^{(k)}}$, habiendo minimizado la suma de cuadrados residuales (SCR).

El biplot muestra que la dimensión 1 parece explicar prácticamente toda la variación observada en el grupo de variables analizadas. De hecho, la suma de cuadrados explicada por esta dimensión representa un 99,99% de toda la variación observada. En consecuencia, es suficiente tomar como referencia el eje asociado a la dimensión 1 para realizar el análisis de la variabilidad por variedad y/o lote.

Figura 3. Biplot de matriz consenso y configuraciones transformadas



El híbrido *E. grandis x camaldulensis* parece mostrar un comportamiento multivariado diferente (y opuesto) a las variedades *E. saligna* y *E. camaldulensis*. Lo anterior puede ser

observado a través de la posición relativa y la distancia entre el centroide de cada configuración de consenso, y el par coordenado (0,0) del eje dimensional 1.
Asimismo, la variabilidad observada para *E. grandis x camaldulensis* parece ser la menor, de acuerdo a la longitud de los vectores que ligan los puntos de cada una de las configuraciones con el centroide determinado por la matriz de consenso dentro de cada grupo. En términos del modelo, hubo una buena concordancia entre las configuraciones parciales de cada matriz de datos.
En tanto que la variabilidad observada en *E. saligna* fue la mayor.
A continuación, se muestra la tabla del Análisis de la Varianza de Procrustes (PANAVA) correspondiente al ajuste de la matriz de consenso:

Tabla 2. PANAVA, indicando la suma de cuadrados del ajuste de la matriz de consenso por cada variedad/híbrido de eucalipto.

Variedad/híbrido	Suma cuadrados modelo	Suma cuadrados residuales	Suma cuadrados totales
E. camaldulensis	**31,43**	6,17	37,60
E. grandis x camaldulensis	**42,86**	2,40	45,26
E. saligna	0,89	**16,26**	17,14
Suma	75,17	24,83	100,00

La variedad con mayor suma de cuadrados del modelo es la que mayor variabilidad aportó al consenso, marcando su carácter diferencial con el resto.
Las variedades con mayor suma de cuadrados residual son las que mayores diferencias presentan *entre* ambientes, y las que tienen menor valor de suma de cuadrados residual son las de comportamiento más homogéneo.
En la Tabla 2 se observa que el "peso" de la suma de cuadrados del modelo asociado a *E. grandis x camaldulensis* fue el mayor, indicando que fue el híbrido que mayor aportó a la variabilidad observada *dentro* de cada microambiente, seguido por *E. camaldulensis*.
En contrapartida, *E. saligna* realizó la mayor contribución relativa a la suma de cuadrados residuales, lo que puede interpretarse como que fue la variedad menos estable o más variable *entre* microambientes.

CONCLUSIONES

El Análisis de Procrustes Generalizados (GPA) permitió sintetizar y describir de forma completa la variabilidad observada de dos variedades y un híbrido de eucalipto en dos microambientes de la Esperanza (provincia de Jujuy), por medio de ajustes de configuraciones parciales y la generación de una matriz de consenso.

Constituye una alternativa muy recomendable para el análisis multivariado de la variabilidad por ambientes, ya que permite comparar los aportes ponderados que cada genotipo realiza *dentro* y *entre* ambientes.

Se pudo observar que el híbrido *E. grandis x camaldulensis* de dos años fue el genotipo más estable *entre* ambientes, además de presentar los mayores valores de AB por ha y altura promedio.

E. camaldulensis también presentó una relativa estabilidad entre ambientes. Además, fue el que presentó los menores valores de mortalidad y tasas de falla.

E. saligna fue el genotipo que presentó mayor variabilidad *entre* ambientes, es decir, fue el menos estable.

Se espera que este trabajo pueda contribuir al reconocimiento del método de GPA como una técnica óptima para el análisis multivariado de componentes de variabilidad en estudios multiambientales en el campo forestal.

BIBLIOGRAFÍA

Bruno, C. y Balzarini, M. 2010. Ordenaciones de material genético a partir de información multidimensional. Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo

Commandeur, Jacques. 1991. MATCHING CONFIGURATIONS. Department of Psychometrics and Research Methodology. DSWO Press, Leiden University. Países Bajos.

Del Médico, A. y Vitelleschi, M. 2015. ANÁLISIS PROCRUSTES GENERALIZADO. UNA APLICACIÓN EN EL ÁREA AGRÍCOLA. Vigésimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística.

Gower, J. y Dijksterhuis, G. 2005. “PROCRUSTES PROBLEMS”. Oxford: Oxford University Press.

Gower, J. .2004. The geometry of biplot scaling. Biometrika, 91 705-714.

Grice, J. y Assad, K. 2009. GENERALIZED PROCRUSTES ANALYSIS: A TOOL FOR EXPLORING AGGREGATES AND PERSONS. Applied Multivariate Research, Volume 13, No. 1.

Lê, S.; Josse, J. And Housson, F. 2008. FactorMineR: an R package for multivariate analysis. Journal of Statistical Software.

Stegmann, L. y Delgado Gómez, D. 2002. A Brief Introduction to Statistical Shape Analysis. Informatics and Mathematical Modelling, Technical University of Denmark.

Torcida, S. y Perez, S. 2018. ANÁLISIS DE PROCRUSTES Y EL ESTUDIO DE LA VARIACIÓN MORFOLÓGICA. REVISTA ARGENTINA DE ANTROPOLOGÍA BIOLÓGICA. Volumen 14, Número 1, Páginas 131-141.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Método de exposición cuasi-inducida: asignación de responsabilidad

Almudena Sanjurjo de No, Blanca Arenas Ramírez, José Mira McWilliams, Francisco Aparicio Izquierdo

Institución: Instituto Universitario de Investigación del Automóvil Francisco Aparicio Izquierdo (INSIA-UPM), Escuela Técnica Superior de Ingenieros Industriales (ETSII-UPM), Universidad Politécnica de Madrid (UPM), Madrid (España)

Datos de contacto: almudena.sanjurjo.no@gmail.com

RESUMEN

Determinar las tasas de accidentalidad de diferentes colectivos de conductores es un objetivo importante en seguridad vial. Para ello, es necesario contar con los niveles de exposición de los conductores de manera desagregada.

Sin embargo, esta información no es conocida al nivel requerido, por lo que se recurre al método de exposición cuasi-inducida, que permite determinarla de forma relativa a partir de los datos de las bases de accidentes de tráfico. La hipótesis principal del método es que los conductores no responsables constituyen una muestra aleatoria de la población de conductores. Por tanto, determinar correctamente la responsabilidad del conductor es clave.

Hasta el momento, la asignación de responsabilidad se hace, fundamentalmente, en base a las infracciones del conductor y de velocidad, dejando de lado otros registros que podrían incrementar la probabilidad de que un conductor sea responsable.

En este trabajo se aplica la metodología Self-Organizing Maps (SOM) para explorar el conjunto de infracciones y condiciones desfavorables de los conductores, que pueden influir en su responsabilidad.

Los resultados obtenidos indican que la clasificación de conductores mejora en comparación con la clasificación tradicional: más conductores clasificados, mayor número de variables consideradas y mayor calidad esperada en el proceso de asignación de responsabilidad.

Palabras Clave: análisis multivariante, asignación de responsabilidad, método de exposición cuasi-inducida, seguridad vial, self-organizing maps (SOM)

INTRODUCCIÓN

Determinar las tasas de accidentalidad de diferentes colectivos de conductores es muy importante en seguridad vial con el objetivo de establecer medidas preventivas que traten de evitar el accidente o minimicen sus impactos, tal y como han señalado Stamatiadis y Deacon (1997); Redondo et al. (2000); Hing et al. (2003); Jiang y Lyles (2007, 2010); Lenguerrand et al. (2008); Chandraratna y Stamatiadis (2009); Lardelli et al. (2011); Jiang et al. (2012, 2014); Haque et al. (2013); Huggins (2013); Martínez et al. (2013); Pulido et al. (2016).

Las tasas de accidentalidad de un colectivo de conductores se definen como el cociente entre el número de accidentes de este colectivo y la exposición del mismo (Chandraratna y Stamatiadis, 2009; Gómez y Aparicio, 2010; Huggins, 2013). Por tanto, para determinarlas es necesario contar con alguna medida de los niveles de exposición del colectivo en cuestión, lo que supone un importante desafío entre los investigadores de seguridad vial (Redondo, 2000; Chandraratna y Stamatiadis, 2009; Gómez y Aparicio, 2010; Lardelli et al., 2011; Haque et al., 2013; Pulido et al., 2016). Por ello, se recurre al método de exposición cuasi-inducida, que permite la estimación relativa de la exposición de un grupo de conductores a partir de la información contenida en la base de datos de accidentes (Stamatiadis y Deacon, 1997; Lardelli et al., 2005; Redondo et al., 2000; Lenguerrand et al., 2008; Chandraratna y Stamatiadis, 2009; Cooper et al., 2010; Lardelli et al., 2011; Huggins, 2013; Jiang et al., 2012, 2014; Pulido et al., 2016).

La hipótesis principal de este método es que los conductores no responsables en accidentes entre dos turismos constituyen una muestra aleatoria de la población general de conductores (Stamatiadis y Deacon, 1997; Hing et al., 2003; Lardelli et al., 2005, 2011; Yan et al., 2005; Yan y Radwan, 2006; Jiang y Lyles, 2007, 2010, 2011; Lenguerrand et al., 2008; Chandraratna y Stamatiadis, 2009; Gómez y Aparicio, 2010; Cooper et al., 2010; Mohaymany et al., 2010; Jiang et al., 2012, 2014; Haque et al., 2013; Martínez Ruíz et al., 2013). Por tanto, la correcta asignación de responsabilidad a partir de la información de la base de datos de accidentes constituye una importante tarea para la estimación posterior de la exposición relativa.

El registro de accidentes de tráfico con víctimas de España no contiene información específica sobre la responsabilidad de cada conductor implicado en un accidente, aunque sí registra las infracciones cometidas por los mismos, así como el estado de dichos conductores (Martínez et al., 2013; Pulido et al., 2016). Sin embargo, no existe un consenso claro acerca de cuáles son las variables que se deberían utilizar para realizar tal asignación de responsabilidad, dado que algunos autores como DeYoung et al. (1997), Jiang y Lyles (2007, 2010), Jiang et al. (2012, 2014) pusieron de manifiesto que asignar la responsabilidad en base a comportamientos no relacionados con la conducción podrían sesgar los resultados. Por ello, fundamentalmente en los últimos años, los investigadores se han inclinado más por asignar la responsabilidad del conductor en base a acciones de conducción peligrosas (Jiang y Lyles, 2010, 2011; Jiang et al., 2012, 2014), principalmente englobadas en las variables infracción del conductor e infracción de velocidad.

En esta investigación se aplica la metodología de clúster Self-Organizing Maps (SOM) para llevar a cabo el análisis de un conjunto más amplio de variables que podrían influir sobre la responsabilidad de los conductores, pero no suelen ser tenidas en cuenta por no ser determinantes totales de la responsabilidad. La investigación se divide en dos objetivos fundamentales: (a) Establecer cuáles podrían ser las variables más relevantes en la asignación de responsabilidad y (b) Realizar una propuesta de asignación de responsabilidad que será comparada con la asignación tradicionalmente realizada.

MATERIAL Y METODOLOGÍA

La base de datos utilizada para llevar a cabo esta investigación fue proporcionada por la Dirección General de Tráfico (DGT) y contiene información de los conductores implicados en accidentes de tráfico ocurridos en España entre 2004 y 2013.

Inicialmente la base de datos fue sometida a un proceso de filtrado para únicamente mantener los accidentes en vía interurbana de tipo frontal, frontolateral, lateral y de alcance, entre dos turismos. Además, se realizó un importante trabajo de depuración de la base de datos filtrada, dado que se detectaron errores que podrían sesgar los resultados de la posterior investigación. Por tanto, la base de datos de partida contaba con un total de 836.598 conductores y tras el filtrado y la depuración de la misma esta quedó finalmente reducida a 145.904 conductores.

La base de datos cuenta con la información relacionada con el conductor (edad, género, infracciones, etc.), con el accidente (día, lugar, etc.) y con el vehículo (tipo, defectos, etc.), lo que constituye un total de 113 variables. A partir de las mismas, se seleccionaron todas aquellas que se consideró que podrían afectar, en mayor o menor medida, a la asignación de responsabilidad del conductor. Por tanto, las variables seleccionadas para la construcción posterior del Self-Organizing Maps (SOM) son: infracción del conductor, infracción de velocidad, infracción administrativa, consumo de alcohol y/o drogas, enfermedad súbita, sueño, cansancio y/o preocupación (englobada bajo la variable sueño), defecto físico previo del conductor y estado del vehículo.

Por otro lado, para poder trabajar con estas variables aplicando la metodología SOM, ha sido necesario llevar a cabo un proceso de transformación de los valores de las mismas de forma que éstas pasasen de ser categóricas a numéricas. Por convenio, se adoptó el valor de 0 para indicar que dicha infracción o condición no estaba presente y el valor de 2 para indicar justo lo contrario. Adicionalmente, se ha utilizado el valor intermedio de 0,25 para indicar que se ignora si está presente o no dicha infracción. Para tomar este valor, se ha partido de dos hipótesis: (a) el mismo tendría que estar comprendido entre 0 ó 2, dado que su categoría es intermedia a las otras dos y (b) si este valor está indicado como “Se ignora” es más probable que sea porque no estuviese presente, por lo que debe asignarse un valor más cercano al 0. Para validar esta cuestión se realizó un análisis de sensibilidad de los mapas SOM para distintos valores (comprendidos entre 0,25 y 1) para los casos en los que el valor de una o más variables era desconocido. Los resultados obtenidos, que no son mostrados aquí por no ser objeto de esta publicación, demostraron que no era significativa la elección de este valor de entre todos los valores probados. Por tanto, basándonos en las hipótesis planteadas, se escogió el valor de 0,25 para los casos en los que se desconocía el valor de la variable. Más información puede encontrarse en Sanjurjo-de-No et al. (2021).

Como se ha indicado anteriormente, la metodología empleada para llevar a cabo esta investigación es el método clúster conocido como Self-Organizing Maps (SOM).

Self-Organizing map es una técnica de aprendizaje no supervisado que fue desarrollada por Kohonen (1990) y forma parte de las técnicas de Machine Learning. El propósito de SOM es representar datos que, originalmente están en un espacio multidimensional, en un espacio de dimensión más reducida (típicamente 2 ó 3 dimensiones), pero manteniendo la estructura topológica original de los datos. Así, conductores que por sus características estuviesen cercanos en el espacio original, deberían seguir estando cercanos en el espacio proyectado. Como señalan algunos autores como Liu, P. (2009), Kohonen, T. (2013), Kohonen, T. (1998) o Lagus, K. (2002), esto supone una importante ventaja del SOM con respecto a otros métodos de clúster, dado que permite la visualización de las nubes de puntos generadas por la proyección de los datos.

En la presente investigación, la metodología SOM permite realizar un análisis conjunto de las 8 variables anteriormente seleccionadas. De esta forma, se realiza un análisis conjunto de los conductores en el espacio original de 8 dimensiones el cual es proyectado sobre el espacio de 2 dimensiones, pero manteniendo la topología original de los datos de los conductores. Esto tiene por objetivo ayudar a identificar patrones de comportamiento entre los conductores en función de las infracciones que estos pueden o no haber cometido, lo que permitirá arrojar luz sobre la responsabilidad de los mismos. Además, al apoyar la asignación de responsabilidad con esta metodología se está proporcionando más información acerca de cómo son los datos de manera multivariante, dado que se tiene en cuenta la distancia conjunta de todas las variables. Esto hace que el proceso de asignación de responsabilidad sea menos radical, esperando con esto que la asignación sea más precisa.

DESARROLLO

El mapa SOM de infracciones sobre el que se reparten los 145.904 conductores de la base de datos es el que se muestra en la Figura 1. Este mapa está dividido en 25 clústers o nodos (dimensión 5x5), donde cada uno alberga el número de conductores que se indica en negro en la parte superior de cada clúster. En rojo, se indica el número del clúster en cuestión. Así, por ejemplo, el clúster 15 contiene un total de 30.729 conductores, mientras que en el clúster 23 no hay ningún conductor. Un análisis más detallado puede encontrarse en Sanjurjo-de-No et al. (2021).

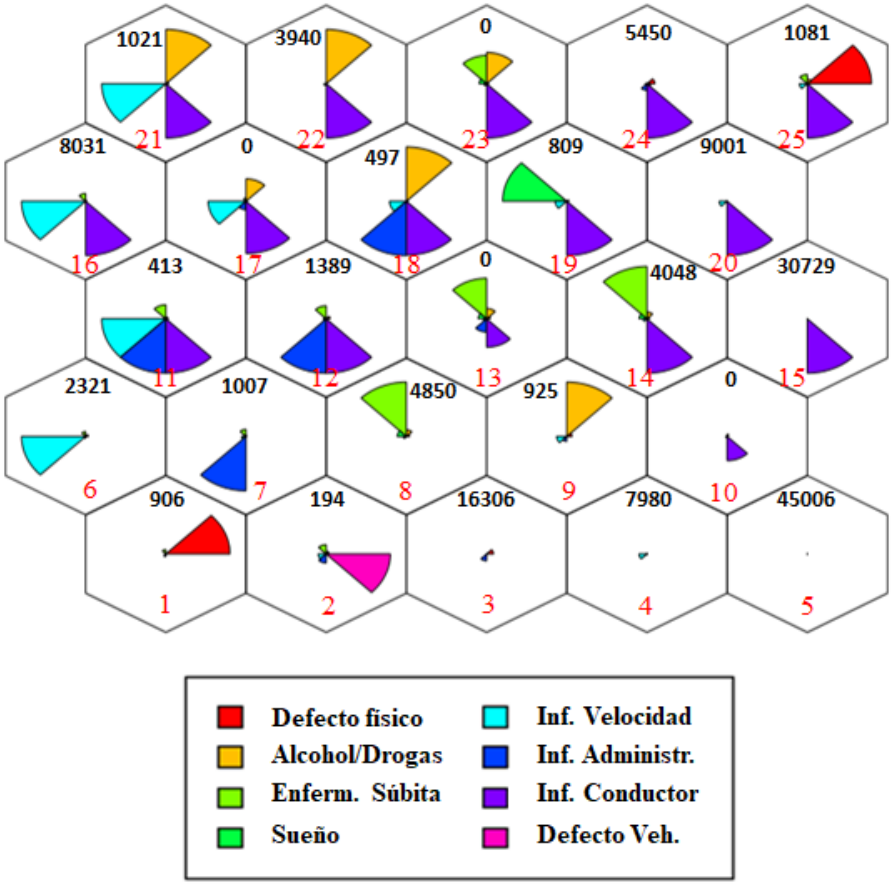


Figura 1. Mapa SOM de infracciones

En cada clúster es posible visualizar las características de los conductores en el espacio proyectado de 2 dimensiones, cuando originalmente estaban en el espacio de 8 dimensiones (una dimensión por cada variable introducida en el SOM). Cada uno de los vectores de colores que se muestran en cada clúster indica el valor promedio de la variable que representan para todos los conductores que han sido asignados a ese clúster. En la Tabla 1 se indica

numéricamente dicho valor promedio para cada una de las variables en cada uno de los clústers en los que hay algún conductor asignado.

	Defecto físico	Alcohol/Drogas	Enfermedad súbita	Sueño/Cansancio	Inf. Velocidad	Infr. Administrativa	Inf. Conductor	Defecto del vehículo
Cluster 1	2	0.02	0.02	0.02	0	0.09	0	0
Cluster 2	0.10	0.10	0.05	0.06	0.25	0.33	0	2
Cluster 3	0.22	0	0	0	0	0.25	0	0
Cluster 4	0	0	0	0	0.25	0.03	0	0
Cluster 5	0	0	0	0	0	0	0	0
Cluster 6	0.09	0.02	0.02	0.02	2	0.07	0	0
Cluster 7	0.02	0.03	0.03	0.07	0	2	0	0
Cluster 8	0.10	0.24	0.27	0.29	0	0.03	0	0
Cluster 9	0.17	2	0	0	0.32	0.18	0	0
Cluster 11	0.07	0.07	0.07	0.08	2	2	2	0.07
Cluster 12	0.13	0.05	0.06	0.08	0	2	2	0.05
Cluster 14	0.09	0.25	0.26	0.25	0	0.04	2	0
Cluster 15	0	0	0	0	0	0	2	0
Cluster 16	0.02	0.04	0.04	0.04	2	0.03	2	0.03
Cluster 18	0.05	2	0	0	0.52	2	2	0.04
Cluster 19	0.07	0	0	2	0.36	0.02	2	0.02
Cluster 20	0	0	0	0	0.25	0.02	2	0
Cluster 21	0.09	2	0	0	2	0.03	2	0.02
Cluster 22	0.03	2	0	0	0	0.04	2	0.03
Cluster 24	0.25	0	0	0	0	0.24	2	0.12
Cluster 25	2	0.17	0.04	0.02	0.25	0.02	2	0

Tabla 1. Valor promedio de cada variable en cada clúster

Variables importantes en la asignación de responsabilidad

En este apartado se realiza un análisis de las variables introducidas en el SOM con el objetivo de determinar cuáles son aquellas que mejor identifican a los grupos o clústers de conductores del SOM. Se debe llegar a una solución de compromiso entre usar un número excesivo de variables para realizar la asignación de responsabilidad y tener en cuenta el suficiente número de variables para la identificación de los grupos de conductores responsables y no responsables de forma que se pierda la menor cantidad posible de información.

En la Tabla 1 y Figura 1 puede observarse que la variable que mejor divide a los conductores en dos grupos es la infracción del conductor, dado que, entre todos los clústers con conductores asignados, esta variable siempre adopta valor 0 ó 2 y no un valor intermedio entre estos. Por tanto, se trata de la variable que mejor define la responsabilidad de los conductores. Además, se observa que la infracción de velocidad es una variable también relevante en la agrupación de los conductores en diferentes clústers y, al igual que la infracción del conductor, mide comportamientos de conducción peligrosos. Por tanto, esta variable también debería ser considerada en el proceso de asignación de responsabilidad.

La variable defecto físico previo aparece sólo en los clústers 1 y 25 del mapa, lo que hace pensar que no es significativa en la responsabilidad de los conductores. Los defectos físicos de los conductores son principalmente defectos de visión y audición y estos defectos están relacionados con la edad de los conductores (Sanjurjo-de-No et al., 2020) y hay investigaciones, como la realizada por Sagar et al. (2020), que indican que los conductores mayores tienen más probabilidad de ser responsables en un accidente que los conductores pertenecientes a otras franjas de edad. Esta variable debería, por tanto, ser sometida a análisis futuros que evalúen más profundamente su grado de influencia sobre la responsabilidad del conductor.

Por otro lado, se observa que las variables denominadas “Estado del vehículo” y “Sueño” aparecen de manera aislada en muy pocos casos o aparecen combinadas con otras variables, como la infracción del conductor, que ya por sí sola permitía clasificar a los conductores. Por lo que, se considera que estas variables podrían no ser tenidas en cuenta en el proceso de asignación de responsabilidad.

En relación con las infracciones administrativas, se considera que, presumiblemente, la única infracción de este tipo que podría ser relevante sobre la responsabilidad es no haber pasado la Inspección Técnica Reglamentaria del Vehículo (ITV) cuando esta va unida a un mal estado de dicho vehículo. Sin embargo, no se ha identificado ningún clúster en los que ambas variables aparezcan simultáneamente. Por tanto, se consideran poco relevantes a la hora de realizar la asignación de responsabilidad.

Finalmente, en relación a la variable "Enfermedad súbita", es importante señalar que en la base de datos de accidentes no se especifica el tipo de enfermedad súbita sufrida por el conductor que la padece, lo que facilitaría su tratamiento. Esta variable no aparece por sí sola en muchos de los clústers, por lo que podríamos no considerarla en el modelo. Sin embargo, teniendo en cuenta que las enfermedades súbitas se definen como aquellas que aparecen sin ser esperadas y normalmente hacen perder las facultades normales de la persona que las padece (desmayo, ataque de ansiedad, infarto, etc), podríamos considerar que puede afectar a la probabilidad de que un conductor que la padezca pueda ser responsable de un accidente de tráfico. Por esta razón, análisis adicionales también deberían llevarse a cabo en relación a esta variable para determinar el grado de influencia de la misma sobre la responsabilidad de los conductores.

En la Tabla 2 se muestran finalmente las variables consideradas más importantes para realizar la asignación de responsabilidad.

Variables más influyentes en la asignación de responsabilidad	Variables menos influyentes en la asignación de responsabilidad	Relevancia desconocida en el proceso de asignación de responsabilidad
Infracción del conductor	Infracción administrativa	Defecto físico previo
Infracción de velocidad	Sueño	Enfermedad súbita
Consumo de alcohol/drogas	Estado del vehículo	

Tabla 2. Variables más relevantes para la asignación de responsabilidad con mapas SOM

Propuesta de asignación de responsabilidad en base al SOM

En este apartado se realiza el análisis del mapa SOM anteriormente creado desde el punto de vista de la responsabilidad de los conductores. Para ello, se tratará de identificar el perfil de cada uno de los conductores en función del clúster al que estos pertenezcan.

Como puede observarse en la Figura 2, es posible establecer una frontera en el mapa, por criterio experto, en función de la combinación de variables que aparece en cada uno de los clústers. Esta frontera divide el mapa en dos regiones: potencialmente responsables y potencialmente no responsables.

En la región superior (Potencialmente responsables) se localizan todos aquellos conductores que han cometido una o múltiples infracciones o presentan condiciones desfavorables para la conducción. Se considera que estos conductores son responsables, dado que el perfil de infracciones presentes en esta zona del mapa.

Por otro lado, en la región inferior del mapa (Potencialmente no responsables) se encuentran los conductores que con seguridad no han sido responsables (en el clúster 5) y aquellos conductores que habría que analizar en profundidad junto con el otro conductor del accidente, con el objetivo de determinar si es posible la clasificación de los mismos.

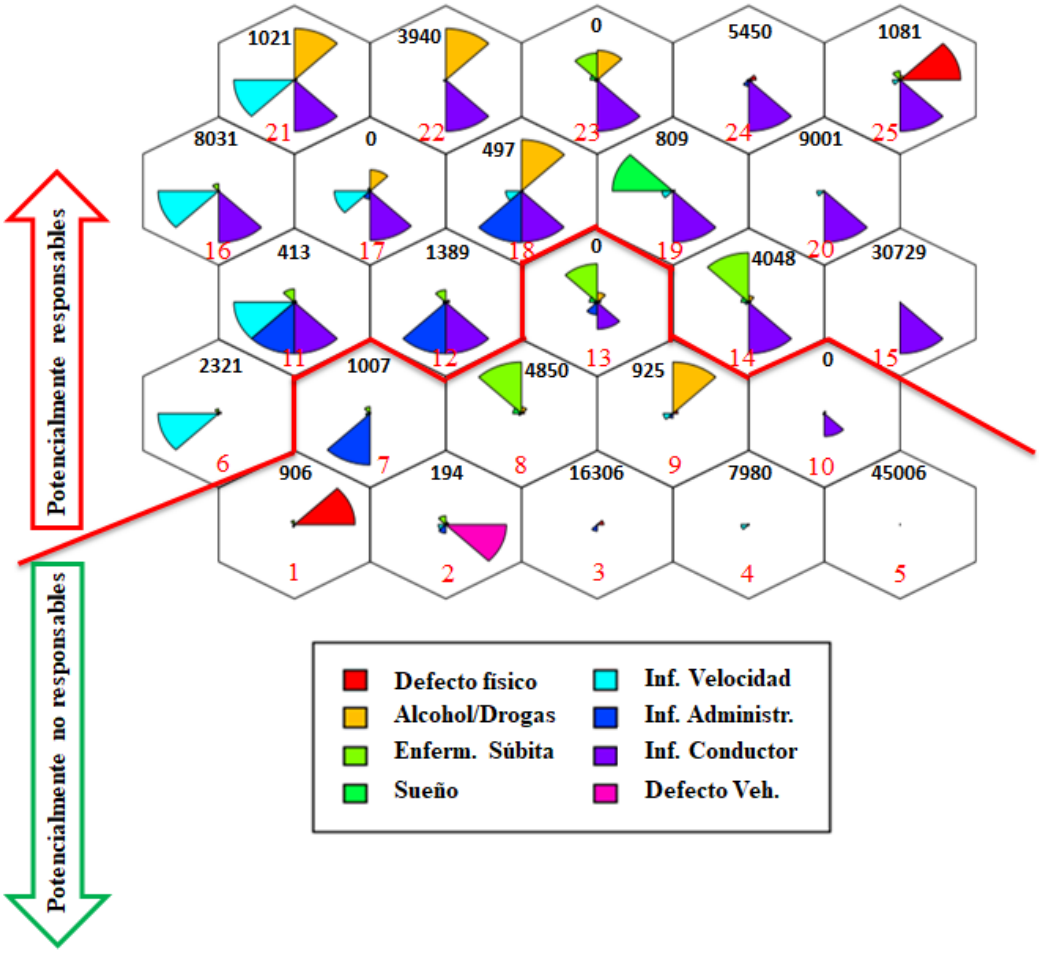


Figura 2. Mapa SOM de infracciones con la frontera de separación de regiones

El análisis conjunto de las parejas de conductores implicados en un mismo accidente dio lugar a la identificación de varias casuísticas posibles: (I) Responsable / No Responsable: Si un conductor está situado en la parte superior del mapa y el otro conductor con el que tuvo el accidente está situado en la parte inferior del mapa (establecido por la frontera), entonces el primer conductor podría considerarse responsable del accidente, mientras que el segundo sería considerado como no responsable; (II) Responsable / Responsable: Si ambos conductores están situados en la parte superior del mapa de acuerdo con la frontera establecida, entonces ambos podrían considerarse responsables y, por lo tanto, no se tendrían en cuenta en el análisis; (III) No responsable / No responsable: Si ambos conductores implicados en un mismo accidente están en el clúster 5, entonces los dos podrían ser considerados no responsables y, por lo tanto, podrían no ser tenidos en cuenta en el posterior análisis; Y (IV) Casos de Análisis: Si ambos conductores están en la parte inferior del mapa de acuerdo con la frontera establecida y, al menos uno de ellos en un clúster diferente al 5, entonces habría que realizar análisis adicionales para determinar si es posible saber cuál de ellos podría haber sido el responsable del accidente.

En la Tabla 3, se muestra el reparto de conductores en cada una de estas categorías y, puede observarse como, a partir de la interpretación del mapa SOM, podrían llegarse a clasificar un total de 83,76% de los conductores en Responsable / No responsable.

	Número de conductores	%
Responsable / No responsable	122.208	83,76%
Responsable / Responsable	7.626	5,23%
No responsable / No responsable	2.014	1,38%
Casos de Análisis	14.056	9,63%
TOTAL	145.904	100%

Tabla 3. Clasificación de los conductores con mapas SOM

Los resultados obtenidos usando los mapas SOM para realizar la asignación de responsabilidad fueron comparados con los resultados obtenidos cuando dicha asignación se hace teniendo en cuenta únicamente la infracción del conductor y la de velocidad, que son las variables más comúnmente utilizadas para realizar esta asignación. En este caso, se considera que un conductor es responsable si ha cometido infracción del conductor y/o de velocidad y no será responsable en caso contrario.

Los resultados son los que se muestran en la Tabla 4, donde se observa que se logra clasificar al 72,47% de los conductores.

Por tanto, realizando la asignación de responsabilidad utilizando la metodología SOM logramos rescatar a un mayor número de conductores para el análisis posterior. Además, tiene en cuenta una mayor cantidad de variables a la hora de realizar dicha asignación. Por lo que se espera que la calidad de la misma sea mayor.

	Número de conductores	%
Responsable / No responsable	105.736	72,47%
Responsable / Responsable	7.706	5,28%
No responsable / No responsable	13.098	8,98%
Casos de Análisis	19.364	13,27%
TOTAL	145.904	100%

Tabla 4. Clasificación de los conductores en función sólo de las infracciones del conductor y de velocidad (método tradicional)

CONCLUSIONES

En esta investigación se propone el uso de la metodología Self-Organizing Maps (SOM) de cluster como herramienta alternativa de ayuda para la asignación de responsabilidades. Con el SOM es posible identificar diferentes patrones en los conductores en relación a las infracciones que estos han cometido. Esto permite, en primer lugar, identificar las variables más y menos relevantes para llevar a cabo la asignación de responsabilidad y, en segundo lugar, nos servirán como herramienta de ayuda para llevar a cabo el proceso de asignación de responsabilidad.

Así, se ha observado que las variables más relevantes a la hora de asignar la responsabilidad son la infracción del conductor, la infracción de velocidad y también el consumo de alcohol y/o drogas.

Finalmente, la distribución de los conductores en el mapa SOM permite ayudar en la identificación de la responsabilidad de los conductores, clasificando a un 11,29% más de los conductores que con la asignación tradicional que, fundamentalmente, tiene en cuenta sólo las infracciones del conductor y las de velocidad. Por tanto, además de conseguir una mayor clasificación de los conductores aplicando la metodología SOM, se está considerando, de

manera multivariante, un mayor número de variables, por lo que se espera que la calidad de la clasificación sea mayor.

BIBLIOGRAFIA

1. Chandraratna, S. and Stamatiadis, N. (2009). Quasi-Induced Exposure Method: Evaluation of not-at-fault assumption. Accident Analysis and Prevention 41, pp. 308-313.

2. Cooper, P.J., Meckle, W., Andersen, L. (2010). The efficiency of using non-culpable crash-claim involvements from insurance data as a means of estimating travel exposure for road user sub-groups. Journal of Safety Research 41, pp. 129-136.

3. DeYoung, D., Peck, R., Helander, C. (1997). Estimating the exposure and fatal crash rates of suspended/revoked and unlicensed drivers in California. Accident Analysis and Prevention 29 (1), pp. 17-23.

4. Gómez, A., Aparicio, F. (2010). Quasi-Induced Exposure: The choice of exposure metrics. Accident Analysis and Prevention 42, pp. 582-588.

5. Haque, M., Washington, S., Watson, B. (2013). A Methodology for estimating exposure-controlled crash risk using Traffic Police Crash Data. Social and Behavioral Sciences 104, pp. 972-981.

6. Hing, J., Stamatiadis, N., Aultman-Hall, L. (2003). Evaluating the impact of passengers on the safety of older drivers. Journal of Safety Research 34, pp. 343-351.

7. Huggins, R. (2013). Using speeding detections and numbers of fatalities to estimate relative risk of a fatality for motorcyclists and car drivers. Accident Analysis and Prevention 59, pp. 296-300.

8. Jiang, X., Lyles, R. (2007). Difficulties with quasi-induced exposure when speed varies systematically by vehicle type. Accident Analysis and Prevention 39, pp. 649-656.

9. Jiang, X., Lyles, R. (2010). A review of the validity of the underlying assumptions of quasi-induced exposure. Accident Analysis and Prevention 42, pp. 1352-1358.

10. Jiang, X., Lyles, R. (2011). Exposure-based assessment of the effectiveness of Michigan's graduated driver licensing nighttime driving restriction. Safety Science 49, pp. 484-490.

11. Jiang, X., Qiu, Y., Lyles, R., Zhang, H. (2012). Issues with using police citations to assign responsibility in quasi-induced exposure. Safety Science 50, pp. 1133-1140.

12. Jiang, X., Lyles, R., Guo, R. (2014). A comprehensive review on the quasi-induced exposure technique. Accident Analysis and Prevention 65, pp. 36-46.

13. Kohonen, T. (1998). The self-organizing map. Neurocomputing 21, pp. 1-6.

14. Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks 37, pp. 52-65.

15. Lagus, K. (2002). Text retrieval using self-organized document maps. Neural Processing Letters 15, pp. 21-29.

16. Lardelli, P., Jiménez, J.J., Luna, J.D., García, M., Moreno, O., Bueno, A. (2005). Comparison between two Quasi-Induced Exposure Methods for studying risk factors for Road Crashes. American Journal of Epidemiology 163 (2), pp. 188-195.

17. Lardelli, P., Luna, J.D., Jiménez, E., Pulido, J., Barrio, G., García, M., Jiménez, J.J. (2011). Comparison of two methods to assess the effect of age and sex on the risk of car crashes. Accident Analysis and Prevention 43, pp. 1555-1561.

18. Lenguerrand, E., Martin, J.L., Moskal, A., Gadegbeku, B., Laumon, B., the SAM group (2008). Limits of the quasi-induced exposure method when compared with the standard case-control design. Application to the estimation of risks associated with driving under the influence of cannabis or alcohol. Accident Analysis and Prevention 40, pp. 861-868.

19. Liu, P. (2009). A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes. Neurocomputing 72, pp. 2902-2908.

20. Martínez, V., Lardelli, P., Jiménez, E., Amezcua, C., Jiménez, J.J., Luna, J.D. (2013). Risk factors for causing road crashes involving cyclists: An application of a quasi-induced exposure method. Accident Analysis and Prevention 51, pp. 228-237.

21. Mohaymany, A.S., Kashani, A.T., Ranjbari, A. (2010). Identifying Driver Characteristics influencing Overtaking Crashes. Traffic Injury Prevention 11, pp. 411-416.

22. Pulido, J., Barrio, G., Hoyos, J., Jiménez, E., Martín, M.M., HouwinG, S., Lardelli, P. (2016). The role of exposure on differences in driver death rates by gender and age: Results of a quasi-induced method on crash data in Spain. Accident Analysis and Prevention 94, pp. 162-167.

23. Redondo, J.L., Luna, J.D., Jiménez, J.J., García, M., Lardelli, P., Gálvez, R. (2000). Application of the Induced Exposure Method to compare risks of Traffic Crashes among different types of Drivers under different Environmental Conditions. American Journal of Epidemiology 153 (9), pp. 882-891.

24. Sanjurjo-de-No, A., Arenas-Ramírez, B., Mira, J., Aparicio-Izquierdo, F. (2020). Driver Pattern Identification in Road Crashes in Spain. IEEE Access 8, pp. 182014-182025.

25. Sanjurjo-de-No, A., Arenas-Ramírez, B., Mira, J., Aparicio-Izquierdo, F. (2021). Driver Liability Assessment in Vehicle Collisions in Spain. International Journal of Environmental Research and Public Health 18(4):1475.

26. Stamatiadis, N., Deacon, J. (1997). Quasi-Induced Exposure: Methodology and Insight. Accident Analysis and Prevention 29 (1), pp. 37-52.

27. Yan, X., Radwan, E., Abdel-Aty, M. (2005). Characteristics of Rear-End Accidents at Signalized Intersections using Multiple Logistic Regression Model. Accident Analysis and Prevention 37, pp. 983-995.

28. Yan, X., Radwan, E. (2006). Analyses of Rear-End Crashes based on Classification Tree Models. Traffic Injury Prevention 7, pp. 276-282.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

VARIABLES DEL INGRESO A LA LIC. MÉDICO CIRUJANO DE LA UAN DURANTE LA CONTINGENCIA POR COVID-19.

De Jesús Espinoza Nadia Grisell

Ibáñez Andrade José Israel

Rodríguez Altamirano David.

Universidad Autónoma de Nayarit, Tepic Nayarit, México.

Centro de Estudios Tecnológicos Industrial y de Servicios Núm.100

(Josefa Ortiz de Domínguez) Tepic Nayarit, México.

israel.andrade@uan.edu.mx. 311-110-94-58

RESUMEN
Las variables de ingreso a la educación superior en México han sido poco estudiadas desde una perspectiva econométrica. La carrera en Medicina Humana al ser una profesión muy demandada y con pocos espacios en las universidades públicas en México, es un ejemplo de que, en el proceso de ingreso, hay variables que pueden incidir en los resultados. Esta investigación tuvo por objeto identificar variables en el ingreso a la carrera de medicina de la Universidad Autónoma de Nayarit en el contexto de la contingencia sanitaria por virus SARS-Cov-2. Se encontró a partir de la aplicación de un modelo de regresión logística multinomial que las variables que influyeron mayormente fueron el promedio de bachillerato y la escolaridad máxima de los padres, teniendo mejores resultados en el proceso de selección aquellos estudiantes con promedios de bachillerato altos y con un grado de estudio mínimo de licenciatura de los padres. Se concluyó que aplicar un modelo de regresión logística multinomial es adecuado, pero se deben incluir otras variables de estudio para una mayor comprensión del fenómeno; y, que, a partir del estudio, se pueden tomar decisiones para fortalecer el ingreso desde la universidad, a través de programas vinculados con el bachillerato.

Palabras Clave: Variables de ingreso al nivel superior, Médico cirujano, contingencia sanitaria.

INTRODUCCIÓN

La contingencia sanitaria por el virus SARS- Cov2 en el mundo dejó ver claramente las brechas sociales existentes dentro del sistema educativo en el mundo. México no fue la excepción y los diferentes niveles educativos se sometieron a cambios necesarios.

El proceso de ingreso al nivel superior sufrió cambios importantes. De manera interna, dentro de la universidad se adaptaron los criterios de selección; y, de manera externa, los estudiantes traían experiencias de aprendizajes diversas y condiciones sociales que permearon en su resultado de ingreso.

Esta investigación es un análisis de variables seleccionadas relacionadas con el ingreso a la carrera de Médico Cirujano de la Universidad Autónoma de Nayarit basado en un ejercicio econométrico utilizando una regresión logística multinomial. Se presenta en sus resultados las variables significativas en el objeto de estudio.

METODOLOGÍA

Para el desarrollo de este trabajo se tomó la base de datos de aceptados al Programa Académico (PA) de Médico Cirujano en el proceso de ingreso de agosto de 2020. Esta base de datos es realizada por la Dirección de Seguimiento Académico de Estudiantes, correspondiente a la Secretaría Académica quien es la que lleva el proceso de ingreso de estudiantes a nivel superior de la Universidad Autónoma de Nayarit.

Esta base de datos es construida a partir del cuestionario que cada estudiante llena a su ingreso para poder acceder a su registro CENEVAL que conlleva la aplicación del examen. Dicho cuestionario es muy extenso y a partir de este se realizan informes para caracterizar a los aspirantes que es parte de una actividad prioritaria dentro de la universidad.

Es importante mencionar que el total de estudiantes registrados al PA de Medicina en dicha base de datos es de 1349. Este total se sometió a una segunda fase en donde lograron ser aceptados 157, para finalmente terminar en un total de 164 inscritos (tomando en cuenta el programa de Apoyo Adicional).

La base de datos contempla las variables género, promedio de bachillerato, nivel máximo de estudios de los padres, régimen de sostenimiento del bachillerato, apoyo en la guía de estudio y estatus. La base de datos se analizó con el programa Statistical Package for the Social Sciences (SPSS) y el programa EViews.

Además de la estadística descriptiva para generar un panorama general sobre la población estudiada, también se utilizó una regresión logística multinomial debido a que las variables utilizadas son politómicas.

Como todo fenómeno educativo, la complejidad de la realidad a partir de su representación en los sujetos puede carecer de objetividad de los datos; por ello, fue que se decidió utilizar este modelo para tratar de analizar el fenómeno del ingreso a la carrera de Médico Cirujano de la UAN ante la alta demanda y el bajo número de espacios.

DESARROLLO

Contexto del fenómeno estudiado

El ingreso a la Educación Superior en México es un fenómeno social que ha adquirido mayor importancia en los últimos años. Como ocurre en muchos países del mundo, la demanda a la

educación superior rebasa los espacios disponibles dentro de ella. De acuerdo con la Organización para la Cooperación y el Desarrollo Económico (2019), “actualmente los programas de salud y bienestar también son relativamente comunes (10.1% frente a 13% del promedio de la OCDE)”; sin embargo, la diferencia entre la oferta de espacios y la demanda de estos, está aún más marcada.

En este contexto, las presiones sociales a favor de incrementar la cobertura en educación superior han producido un gran incremento en el número de estudiantes. En el año 2015, se reportaban 126.296 jóvenes estudiantes de medicina; de los cuales concluyeron los estudios 14.781 y finalmente se titularon 13.081.

En el momento actual, existen aproximadamente 165 escuelas o facultades de medicina en el país y continúan abriéndose nuevas escuelas privadas las cuales vienen a restar espacios clínicos a las escuelas públicas. “Se estima que actualmente, en el sistema de educación en México, hay 10,7 graduados de medicina por cada 100,000 habitantes, no muy lejano de la media de 12,1 graduados en la OCDE” (León, R., Lara, V., Abreu, L., 2018).

Sin embargo, aunque aparentemente las cifras son altas, la realidad es que, en las universidades públicas, los espacios al programa académico de Medicina siempre son limitados para la demanda que se tiene. Y es que, la autonomía de cada uno de los estados en México permite la toma de decisiones sobre la forma de organizar la educación de la medicina en su territorio. Por lo cual, cada estado puede definir cuántas escuelas de medicina pueden funcionar en su territorio (Pierdant, G. y Grimaldo, J., 2013).

En el caso del Estado de Nayarit, actualmente la Universidad Autónoma de Nayarit es quien oferta la carrera de Médico Cirujano. Sin embargo, el Instituto Universitario de Ciencias Médicas y Humanísticas de Nayarit (institución educativa privada), ha sido una alternativa para quienes no logran ingresar a la UAN, ofreciendo la carrera de Medicina General Interna. Por su parte, la Universidad Autónoma de Nayarit, se apega a la normatividad de la Secretaría de Salud de Nayarit, en cuanto al límite de espacios – cama, que aseguren a los egresados del programa un espacio para su residencia.

El hecho es que, a pesar de los esfuerzos por incrementar la calidad educativa de las universidades, esto no deriva en más espacios educativos. Como menciona Guzmán, C. y Serrano, O. (2011), se ha agudizado el problema de la desigualdad de oportunidades para ingresar al nivel superior en el mundo; aunado a esta situación debemos tener en cuenta la contingencia sanitaria por el virus SARS- Cov2, que contribuyó a agudizar aún más las brechas sociales correspondientes al ingreso a la educación superior.

Por lo anterior, como afirma Ramírez, L. (2019), las universidades públicas no logran cubrir la demanda de la población en edad escolar. Como menciona Alcántara, A. y Villa, L. (2014), inherente a la expansión de la matrícula de la educación superior en Latinoamérica, es desigual por las brechas sociales de la población. Por ello, los programas de Medicina en México no son para todos ya que socialmente es sinónimo de estatus que se traduce en otras variables socioeconómicas que los distingue.

Es por ello que este trabajo presenta un análisis sobre algunas variables que pueden ayudar a explicar quienes ingresan o no a este tipo de programas educativos bajo el contexto de la contingencia sanitaria por el virus SARS- Cov2.

Es necesario precisar que, desafortunadamente, existe un estado del arte muy limitado sobre factores que inciden en el ingreso al nivel superior. Generalmente es mayormente estudiado la permanencia, que el ingreso. Sin embargo, ambas comparten elementos que puede ayudar a explicar el fenómeno. Variables como el género, la situación económica, el capital cultural familiar, el promedio de bachillerato, entre otros, son las más utilizadas para explicar fenómenos en el área educativa.

El presente trabajo se enfoca en determinar relaciones de variables que son significantes en el proceso de ingreso a la educación superior, específicamente al programa de Médico Cirujano de la Universidad Autónoma de Nayarit en el proceso de ingreso de agosto de 2020.

El ingreso al programa de Médico Cirujano en la Universidad Autónoma de Nayarit

La Universidad Autónoma de Nayarit alberga al mayor número de estudiantes de nuevo ingreso al año en el Estado de Nayarit. En este sentido, el programa de Médico Cirujano es el que más demanda tiene en cada periodo de ingreso.

En cuanto a los aceptados, es de notar que el número de aceptados en promedio es de 166 aspirantes incluyendo el programa de Apoyo Adicional efectuado por instancias internas como un beneficio para docentes y personal administrativo de la institución.

Además, el número de aspirantes al programa cada año oscila entre los 1300 y 1500. Sin embargo, es importante mencionar que, a pesar de los esfuerzos por abrir espacios para matrícula, se ve una diferencia significativa entre la demanda y la oferta. Por esta razón, surge el interés por estudiar el fenómeno desde una visión econométrica.

La aplicación de un modelo econométrico puede aportar una visión diferente del fenómeno estudiado y dar elementos para la toma de decisiones. Lo importante es poder identificar a partir de variables ya dadas cuál es la que tiene una mayor significancia en el ingreso a este programa.

El objetivo de este trabajo fue identificar variables con una significancia en el ingreso al programa de Médico Cirujano de la Universidad Autónoma de Nayarit considerando la pandemia como una causa probable del movimiento de las variables de ingreso a la educación superior.

Resultados del análisis

Como primer paso realizamos la tabla de distribución de frecuencias para observar las características generales de nuestra base. El total de observaciones fueron de 1349 en donde un 61% fueron mujeres y el resto hombres. Cabe mencionar que en el resultado final de los 193 con el estatus “inscritos”, 112 fueron mujeres; es decir, un 58% del total. Y es que, como menciona Cortez, A. *et al.* (2004), el acceso a las escuelas de medicina en la cultura occidental no representa problema para la mujer, ya que aproximadamente 50% de los alumnos que ingresan a las escuelas de medicina son mujeres, semejante a lo que sucede en otras carreras consideradas previamente exclusivas para el sexo masculino.

Así también, el nivel máximo de estudio de los padres del aspirante fue de educación obligatoria que implica la Educación Básica y Media Superior con un 57.1%. De ahí, con un 42.3% se encuentran los padres con nivel licenciatura; y, con un porcentaje debajo del 1% se encuentran dos extremos de la trayectoria escolar: padres analfabetos y padres con posgrado.

Esta variable nos da una imagen muy interesante sobre el capital cultural que posiblemente el aspirante tenga. Si bien, como menciona Robledo, P., García, J. y Díez, C. (2009), la colaboración de la familia en tareas educativas produce efectos positivos, no sólo para el alumno, mejorando su rendimiento, potenciando en él el desarrollo de actitudes positivas hacia el colegio, la adquisición de hábitos regulares de estudio o la mejora de su autoestima, sino también para los padres, al contribuir al aumento de su conocimiento sobre el desarrollo y la educación de los hijos, al incremento del número de interacciones de calidad con el centro educativo a la consecución de un desarrollo más ajustado de su autoestima parental. Es por ello, que se podría esperar que esta variable pueda aproximar un mejor entendimiento sobre el fenómeno.

Por otra parte, en cuanto al uso de la guía CENEVAL proporcionada dentro del proceso de admisión, se encontró que el 58.9% admitió utilizarla. Esto es importante destacarlo porque esta herramienta tiene la finalidad de acercar al estudiante a la estructura y forma en como debe ser contestado el examen. Y, aunque el contenido de la guía no es precisamente el examen original, no deja de ser una herramienta que podría posibilitar un mejor entendimiento y aprovechamiento de los tiempos para resolver el examen.

Sobre la variable promedio de bachillerato se observa que los promedios con la categoría "baja" marcados con un "0" equivalen a un 13.6%. Por el contrario, el grueso del porcentaje de esta variable se encuentra en los promedios de entre 8 y 10 que es un promedio "alto" marcado como "1". En este contexto se esperaría que fueran quienes pudieran estar entre los aceptados. Es importante destacar que el promedio de bachillerato es el segundo criterio que se tomó en cuenta para el ingreso al programa académico de Médico Cirujano, por lo que es una variable fuerte de la que se plantea una alta significancia para el ingreso.

Finalmente, se observa como de 1349 aspirantes, el total en estatus 1 (inscritos) fue solo del 14.3%. claramente se observa un problema de cobertura educativa que más adelante se abordará en el análisis.

De acuerdo con el análisis descriptivo de los datos, el efecto de las variables género, promedio de bachillerato, escolaridad de los padres, régimen del bachillerato y apoyo en la guía CENEVAL fueron elegidas para este análisis por los supuestos que teóricamente se tienen sobre ellas en la influencia en el área académica y en un proceso de ingreso al nivel superior.

De acuerdo con los datos, la única variable que presenta una significancia alta con el estatus (inscritos/aceptados), fue la de promedio de bachillerato con una correlación de Pearson de *.076*. Los términos relación o asociación son equivalentes y se usan para designar aquella área de la estadística en la que se evalúa la covariación entre al menos dos variables (Hernández, J. et al. 2018). Partiendo de este supuesto, podemos inferir que el promedio de bachillerato tiene una relación o asociación con la probabilidad de ser aceptado en el nivel superior.

Lo anterior si lo comparamos con las situaciones problemáticas que trajo la contingencia sanitaria, se puede suponer que los resultados del bachillerato se vieron afectados por las condiciones en las cuales se tuvo que adaptar el sistema educativo mexicano a un modelo virtual. De ahí que, los estudiantes que no lograron adaptarse al cambio o sus condiciones eran limitadas, posiblemente influye en un resultado académico.

Ya en su conjunto, se observa que la variable dependiente estatus y la variable independiente promedio de bachillerato, son las que muestran visualmente un mayor comportamiento.
Ahora bien, para ampliar el análisis, se decidió jugar con las variables a fin de poder observar que comportamiento tenían de manera individual mediante el sistema EViews.

Los resultados arrojaron lo siguiente:

1. Se observa que el género define una probabilidad del 7% de ser aceptado.
2. Se observa que el promedio de bachillerato define una probabilidad del 41% de ser aceptado.
3. Se observa que la escolaridad máxima de los padres define una probabilidad del 25% de ser aceptado.
4. Se observa que estudiar la guía CENEVAL define una probabilidad del 11% de ser aceptado.

Hasta este momento el modelo Probit se había utilizado como herramienta para un mejor entendimiento de las variables y poder visualizar cuál de ellas resultaba más benéfico para el modelo; sin embargo, para realizar el análisis final se decidió utilizar una Regresión Logística Multinomial con el total de variables debido a que se hicieron pruebas quitando algunas variables independientes pero el modelo siguió dando resultados bajos.

En este caso, la variable dependiente fue la variable “Estatus”, que representa el ser aceptado o rechazado en el programa de Médico Cirujano. Mientras que las covariables o variables independientes, fueron el género, la escolaridad máxima de los padres, el apoyo en la guía CENEVAL y el promedio de bachillerato.

En una radiografía general, es evidente que existe una falta de cobertura para la alta demanda que la carrera de Médico Cirujano tiene en la Universidad Autónoma de Nayarit; y, aunque las políticas de la Secretaría de Salud son muy claras en la determinación de espacios en hospitales para estudiantes del área de la salud, lo cierto es que hay un 85.7 % de aspirantes que se quedaron fuera en el 2020.

De acuerdo con el resultado, es necesario contemplar otras variables del fenómeno que ayuden a encontrar posibles relaciones significativas con referencia al ingreso a la educación superior. Otros autores sugieren los factores socioeconómicos como una posible explicación.

En cuanto a las pruebas de razón de verosimilitud, se encontró que el promedio de bachillerato es el que alcanza un nivel más alto de significancia con un .003. Y aunque de acuerdo con la regla de significancia, estudiar la guía de CENEVAL se aproxima a un resultado esperado.

Hasta este punto, se constata que el modelo propuesto a partir de las variables seleccionadas para la explicación de posibles relaciones o significancias entre ellas tiene un nivel de explicación bajo. Sin embargo, esto ayudó a considerar que la variable que más incide en la población estudiada es el promedio de bachillerato. Como se mencionó en el análisis descriptivo por cada variable, el promedio de bachillerato tuvo un 86.4% de nivel “alto”.

Ahora bien, en cuanto a pronósticos, se muestran datos muy interesantes. Por ejemplo, el modelo explica casos particulares que en teoría supone que una combinación de variables que conllevan un resultado negativo sería:

1. Promedio bajo.
2. No estudiar la guía CENEVAL.
3. Padres analfabetos.

Sin embargo, como ya se demostró en el presente análisis las relaciones de significancia esperadas no fueron las esperadas. La explicación se reduce a que al trabajar con personas hay un mundo de combinaciones y relaciones entre muchas variables y factores que inciden en un resultado. Para el caso del ingreso a la educación superior, las posibilidades son muy variadas ya que todo giraría de acuerdo con el contexto y a la relación del sujeto con él.

CONCLUSIONES

La demanda al programa de Médico Cirujano en la Universidad Autónoma de Nayarit año con año sigue creciendo, no así los espacios para matrícula. En este sentido existe un problema de cobertura debido a que las políticas institucionales condicionan estos espacios y, por ende, muchos aspirantes se quedan en el camino; más aún, en una situación de contingencia como la vivida mundialmente por el virus SARS Cov- 2.

Existe la necesidad entonces de revisar estas políticas y repensar el papel del Médico dentro de una sociedad que más allá de reflejar un estado saludable, muestra una sociedad enferma que necesita de un mayor número de médicos para su atención considerando las nuevas condiciones en salud que la sociedad necesita a partir de la nueva normalidad.

Aplicar un modelo de regresión logística multinomial es adecuado para el fenómeno estudiado; sin embargo, es necesario incluir otras variables para obtener un resultado

definitivo que ayude a comprender el ingreso a este programa, debido a que los fenómenos sociales presentan una complejidad que va más allá de probabilidades. En el caso del ingreso a la Educación Superior, muchas variables se hacen presentes en el análisis. En este trabajo solo se tomó las variables de género, el promedio de bachillerato, si se estudió la guía CENEVAL y el nivel de estudios máximo de sus padres. En general si se encontraron relaciones en todas, pero únicamente el promedio de bachillerato resultó la variable con una mayor significancia con relación al estatus (ingreso) al programa académico de Médico Cirujano. Y, aunque no fue significativo el género, un dato interesante es que más de la mitad fueron mujeres con el 58% del total de inscritos.

Este estudio permitió visualizar la importancia que tuvo el promedio para el ingreso a la carrera de Médico cirujano, el cual se vio afectado ya que los estudiantes cambiaron drásticamente la modalidad de estudio del trabajo presencial a uno a distancia, afectando su desempeño escolar y finalmente sus calificaciones.

Además, la pandemia orilló a la institución a tomar decisiones con respecto al proceso de ingreso adaptándolo a las nuevas condiciones y que no afectará al proceso o el ingreso de los estudiantes, quienes como ya se detalló anteriormente tienen pocas probabilidades de ingresar por los números reducidos de espacios.

Los pronósticos del modelo fueron muy variados, pero se destaca que tener un promedio bajo, no estudiar la guía CENEVAL y tener padres sin ningún nivel de estudio, si marca una diferencia en el resultado del ingreso, el cual como se explicó pudo agudizarse ante la contingencia sanitaria.

Finalmente, el estado del arte sobre factores o variables que condicionan el ingreso a la educación superior es muy limitado. Generalmente se enfocan los estudios sobre las trayectorias escolares y el impacto de las variables en el rendimiento académico. Por lo tanto, se deja este estudio como una línea guía para próximos análisis.

BIBLIOGRAFÍA

Alcántara, A. y Villa, L. (2014). Desigualdad social y educación superior. Revista Universidades, núm. 59, enero-marzo, pp. 4-8.

Cortez, A., Fuentes, C., López. M., Velázquez, C., Farías, O., Olivares, J. y González, A. (2004). Medicina académica y género. La mujer en especialidades quirúrgicas. Gaceta Médica México. Vol.141 no.4 México jul./ago.

Guzmán, C. y Serrano, O, (2011). Las puertas del ingreso a la educación superior: el caso del concurso de selección a la licenciatura de la UNAM. Rev. Educación superior. Vol.40 no.157 México ene./mar. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-27602011000100002#nota

Hernández, J., Peñaloza, E., Rodríguez, J., Chacón, J., Tolosa, C., Arenas, M., Carrillo, S.y Bermúdez, V. (2018). Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. Revista AVFT. Vol. 37, Núm. 5.

León, R., Lara, V., Abreu, L. (2018). Educación médica en México. Fundación Educación Médica. Vol. 21 (3): 119-128.

Pierdant, G. y Grimaldo, J. (2013). La discrepancia entre la apertura de nuevas escuelas de medicina en México y la planeación de recursos humanos en salud. Revista Investigación en Educación Médica.

Ramírez, L. (2019). La desigualdad en educación superior en México a través del estudio de las trayectorias escolares. Revista CoPaLa. Año 4, Número 7, enero-junio 2019. Pp. 175-187

Robledo, P., García, J. y Díez, C. (2009). La edad y el nivel cultural-educativo de los padres como factores relacionados con la implicación en la educación de los hijos. INFAD Revista de Psicología International Journal of Developmental and Educational Psychology, Nº2, 2009. ISSN: 0214-9877. pp:485-492

Organización para la Cooperación y el Desarrollo Económico (2019). Educación Superior en México: resultados y relevancia para el mercado laboral. Publicado en Paris. Disponible en: https://doi.org/10.1787/9789264309432

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis descriptivo inteligente de las tendencias al Éxito Académico en estudiantes universitarios de la carrera Ingeniería Industrial UNJu, empleando técnicas de Minería de datos

Coro, Octavio Daniel; Farfán, José H.

Institución: Facultad de Ingeniería, Universidad Nacional de Jujuy. San Salvador de Jujuy.

Datos de contacto: odcoro@fi.unju.edu.ar, jhfarfan@fi.unju.edu.ar

RESUMEN

El trabajo realiza un análisis inteligente descriptivo, utilizando técnicas de Minería de Datos. Se seleccionaron algunas variables que evidencian las pautas asociadas al Éxito Académico de las cursadas de los alumnos de la carrera de Ingeniería Industrial de la Facultad de Ingeniería - UNJu. Para realizar dicho análisis y obtener algunos resultados de las cursadas antes mencionadas. Operando para ello, con algunas tablas de la Base de Datos académica generada por el sistema SIU Guaraní desde el año 2000 al año 2018. La metodología utilizada para el presente trabajo es la KDD (Descubrimiento de Conocimiento en Base de Datos), que consta de cinco etapas, que, básicamente un proceso para el procesamiento y análisis de los datos, que nos permitieron obtener los resultados mencionados.

Palabras Clave: Éxito académico, Análisis Descriptivo, Análisis Predictivo, Minería de datos, Base de datos académica.

INTRODUCCIÓN

La permanencia y la graduación de los estudiantes suele ser una problemática relevante compartida por la mayoría de los sistemas universitarios e instituciones que lo componen. De acuerdo al nivel alcanzado en sus indicadores se constituye en condición de existencia o desaparición de aquellas universidades que dependen en forma casi exclusiva de su matrícula, como es el caso de las universidades privadas, o de condicionamiento presupuestario en el caso de las universidades estatales (Lattuada, 2017).
El rendimiento académico es un fenómeno complejo que suele abordarse en las universidades desde distintos indicadores. Con frecuencia se consideran las calificaciones promedio obtenidas en distintos períodos de la estancia en la universidad, tasas de aprobación de materias y créditos, tasas de retención-deserción, tasas de graduación, períodos de graduación respecto de planes de estudio (Sánchez & Chinchilla Brenes, 2005). Estos indicadores son relativamente sencillos de calcular y medir, dado que se basan en recuentos estadísticos, y los informes y reportes que se arman en base a estos números solo reflejan una parte de la historia, la que la estadística descriptiva puede analizar (Fernández, Sánchez, Córdoba, & Largo, 2002).
Ahora bien, si el número de estudiantes que egresan cada año de una universidad fuera el indicador más relevante para estimar la calidad de una universidad, sería deseable no solo identificar las causas que influyen en la deserción de los estudiantes sino también identificar las causas que determinan que a un estudiante le demande mayor o menor cantidad de años finalizar dicha carrera, ya sea una carrera de grado o pregrado. Por ello existen modelos, técnicas y herramientas avanzadas que emplean Inteligencia Artificial para poder identificar patrones y tendencias en los datos analizados (Wang, Rudin, Wagner, & Sevieri, 2015), del mismo modo estos modelos pueden realizar predicciones o pronósticos partiendo del análisis datos históricos, lo que se conoce comúnmente como Aprendizaje Automático o Machine Learning (Witten, Frank, Hall, & Pal, 2016).
El mecanismo que emplea el Machine Learning es muy sencillo, a partir de un conjunto de datos se realiza una división del mismo en tres grupos, uno de entrenamiento, uno de prueba y uno de evaluación. El conjunto de datos de entrenamiento es empleado por el modelo/algoritmo que se desee entrenar, esto es, el modelo aprende la lógica de los datos para realizar predicciones, para clasificar o agrupar datos, el modelo se ajusta a los datos. El conjunto de prueba sirve para testear y ajustar el modelo diseñado. Por último, el conjunto de datos de validación es empleado para obtener indicadores de la eficiencia del modelo planteado. Una vez ajustado y validado el modelo, el mismo es utilizado con datos nuevos y desconocidos para obtener “nuevo conocimiento”.
Una ventaja adicional de este tipo de análisis serviría para producir visualizaciones y representaciones de información con técnicas novedosas que permiten encontrar nuevas relaciones e información a partir de dichas representaciones. Para ello se emplearán técnicas de Visualización de Información (InfoVis) que tienen por objeto presentar la información visualmente, en esencia, para reducir la carga de trabajo cognitivo al sistema perceptivo visual humano (Ware, 2004; Ware, 2008).
La Universidad Nacional de Jujuy posee un Sistema de Gestión Académica llamado SIU Guaraní, a partir del año 2020 se centralizó este sistema, y cada facultad posee una configuración y acceso particular. Cada instancia en cada facultad accede a los datos pertenecientes a su unidad académica, donde se gestiona y almacena la “historia” académica de los alumnos, desde el momento en que ingresan a la facultad hasta que egresan de la misma. Partiendo de esto, se plantea en este trabajo analizar la base de datos de la carrera Ingeniería Industrial, del sistema SIU Guaraní de la Facultad de Ingeniería para identificar las variables y causas que influyen en el éxito académico.

Justificación del proyecto

Anualmente, las autoridades emiten informes que se cargan al sistema SIU Araucano, donde se ve reflejado el número de egresados por carrera y otra información. En paralelo se comparan estos números con otras facultades y a nivel nacional entre todas las universidades. Además es bien sabido, que al menos la Facultad de Ingeniería, realiza un gran esfuerzo para

"retener" a los estudiantes y mejorar el proceso de enseñanza-aprendizaje. Un ejemplo de ello es el sistema de tutorías o el amplio uso de las aulas virtuales, entre otros.
Como se mencionó, los informes estadísticos actuales, si bien son necesarios e importantes, cuentan solo una parte de la historia. Por este motivo, conociendo la realidad desde otra óptica y tomar las decisiones adecuadas nos sirve para mejorar el sistema educativo. Detectando las causas por las cuales el alumno logra un buen nivel académico. Esto se constituye en un reto no solo atractivo desde el punto de vista de conocer la situación actual de una facultad/universidad, sino también desde el punto de vista tecnológico para aplicar técnicas de análisis novedosas a un caso de estudio real. Para que esto logre impactar notoriamente en el día a día de la vida universitaria. Los resultados encontrados seguramente sirvan para tomar decisiones más acertadas al momento de plantear nuevas políticas y esquemas de gestión.

METODOLOGÍA

Se utiliza la metodología KDD (Descubrimiento de Conocimiento en Base de Datos), la cual es básicamente un proceso para el procesamiento y análisis de los datos. El mismo consiste en extraer información en forma de funciones, reglas o gráficos, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos y presentar resultados. Esta metodología KDD consta de cinco etapas que se detallan a continuación.

Etapa de Comprensión del dominio y selección: En primer lugar se trabaja con los datos obtenidos de la Base de Datos del SIU Guaraní de la Facultad de Ingeniería de la UNJu, particularmente se seleccionan solamente para analizar los datos de la carrera de Ingeniería Industrial, con esto se obtiene el conjunto de datos objetivo. Para gestionar estos datos se realizan las gestiones y autorizaciones pertinentes para contar con los mismos y poder analizarlos, con el debido recaudo de no manejar datos sensibles de los alumnos, puesto que no son necesarios para el análisis inteligente que se plantea llevar a cabo.

Etapa de preprocesamiento/limpieza: Luego de analizar la calidad de los datos, se aplican operaciones básicas, para comenzar con la depuración de los datos erróneos en cuanto a fechas incorrectas, códigos de materia, plan, año académicos vacíos. Se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.

Etapa de transformación/reducción: En la misma se realiza el análisis de los atributos para detectar características útiles, para decodificar los datos dependiendo del objetivo del proceso. Se utilizan métodos de reducción de la dimensionalidad; o de transformación para disminuir el número efectivo de variables bajo consideración. También para encontrar representaciones invariantes de los datos. Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos de forma horizontal o vertical. Como por ejemplo en cuanto al atributo fin_vigencia_regularidad, se encontraron fechas correspondientes al año 2090 y 2209, como se observa en el Gráfico 01, seguramente mal cargados por error de tipeo, en total se eliminaron 386 filas.

Etapa de minería de datos: Se plantea el uso de técnicas de Análisis Inteligente Descriptivo asociado a la Minería de datos. En vista de que estas técnicas permiten por un lado extraer datos relevantes y por el otro encontrar relaciones entre los datos que resultan complejas de hallar por métodos habituales (consultas a una base de datos, informes estadísticos, gráficos estáticos). Para ello es necesario desarrollar un modelo de predicción, separar los conjuntos de datos en datos de entrenamiento, prueba y validación para asegurar que las clasificaciones y agrupaciones sean correctas. Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas. Entre las tareas

predictivas están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los clustering y las correlaciones.

Etapa de interpretación/evaluación de datos: Los datos son procesados mediante los operadores y técnicas que ofrece el Software Rapidminer y Excel. Entre las técnicas que se utilizan están las herramientas para el preprocesamiento y preparación de los datos. Se prueban con algunos modelos como Naive Bayes, Generalized linear model, Logistic Regresión, Fast Large Margin y Decision Tree. Una vez que los datos son procesados, los modelos entrenados y validados, se emplean técnicas de inteligencia artificial aplicados con los operadores del software para visualizar las relaciones encontradas. Estas técnicas permiten hallar nuevo conocimiento a partir de representaciones visuales que nada tienen que ver con los gráficos básicos de la Estadística tradicional.

DESARROLLO

Comprensión del dominio y establecimiento de los objetivos

Para la realización del presente trabajo se sacó información de la base de datos del Sistema SIU Guarani de la Facultad de Ingeniería de la UNJu. Con información relevada entre el año 2000 y 2018, con información de la carrera de Ingeniería Industrial, en primera instancia. Se logró confeccionar una tabla depurada en base a otras con la información más relevante para este trabajo. Dicha tabla o archivo cuenta con 53754 registros, y 8 atributos, los cuales se detallan a continuación:

Tabla Industrial:

- Plan: Es el año del plan que tiene la carrera.
- Legajo: Identificación del alumno/a.
- Fecha de Regularidad: Fecha en la cual obtuvo la condición de regular en la cursada de la materia.
- Fin de Vigencia de Regularidad: En el caso de la Facultad de Ingeniería de Jujuy la misma tiene 2 años de duración a partir de la Fecha de Regularidad de la materia o cátedra.
- Año Académico: Año en que se cursó la materia.
- Materia: Nombre o descripción que tiene la cátedra.
- Condición de regularidad y Resultados: Estos dos atributos tienen códigos, cuyas descripciones se detallan en la Tabla auxiliar siguiente:

Tabla auxiliar:

Cond_ Regularidad	Nombre	Descripción	Resultado
1	Libre	Indica si quedó libre en la cursada.	U
2	Abandonó	Abandonó la cursada	U
3	Insuficiente	Reprobó la cursada	R
4	Regular	Regularizó la actividad	A
5	Promocionó	Promocionó la actividad	A

Tabla 00: Tabla Industrial; Atributos seleccionados de la base de datos para realizar el presente análisis y predicción.

- Cond_regularidad: es el atributo que tiene el código de la condición de regularidad que pueden ser del 1 al 5 como se muestra en la tabla anterior.
- Nombre: denota si es el estudiante reviste la condición de libre, abandono, insuficiente, regular o promocionado
- Descripción: detalle de lo que significa cada código.
- Resultado: tiene 3 posibles valores, U: los que quedaron libres o abandonaron, R: los que reprobaron, y A: los que regularizaron o promocionaron.

Objetivos

El presente trabajo consta de los siguientes objetivos:

- Determinar los factores más importantes para la descripción de los resultados obtenidos por los alumnos, principalmente analizando los que regularizaron o promocionaron la materia, según la descripción de la tabla auxiliar.
- Analizar las relaciones y patrones presentes entre los distintos atributos mencionados anteriormente.
- Hallar la distribución de casos de éxito académico de los alumnos de Ing. Industrial de la Facultad de Ingenieria - UNJu
- Determinar un modelo de clasificación de los alumnos priorizando los casos de éxito académico.

En este trabajo de Minería de Datos es relevante establecer un label u objetivo, para este caso de estudio se ha seleccionado el atributo Resultado asociado principalmente a Condición de Regularidad.

Selección de datos

Como la BD tenía inicialmente datos de todas las carreras y planes, para reducir la dimensionalidad del conjunto de datos se seleccionaron solamente alumnos de la carrera Ingeniería Industrial, con el plan 2001 que es el único plan actualmente de esa carrera, y los atributos más relevantes eliminando también datos erróneos, faltantes o redundantes, que no aportan para realizar el presente trabajo. Cabe aclarar que se eliminaron los datos atípicos,

ya que había fechas de fin de vigencia de la regularidad con año 2090 y 2209 en 386 registros, para obtener resultados más representativos.

Los atributos que se eliminaron están marcados en la tabla siguiente:

E15 | 17/08/2007 12:00:00 a. m.

	A	B	C	D	E	F	G	H	I
1	carrera	plan	legajo	fecha_regularidad	fin_vigencia_regu	nd_regularid	resultado	nio_academi	materia
2	1	2001	IND000001	24/09/2004 00:00	31/12/2007 00:00	4	A	2004	ELECTROTECNIA
3	1	2001	IND000001	10/10/2001 00:00	31/03/2090 00:00	4	A	2001	ORGANIZACION DE EMPRESAS
4	1	2001	IND000001	26/12/2001 00:00	31/03/2090 00:00	4	A	2001	ANALISIS MATEMATICO II
5	1	2001	IND000001	09/12/2003 00:00	31/03/2090 00:00	4	A	2003	ESTATICA Y RESISTENCIA DE MATERIALES
6	1	2001	IND000001	04/12/2002 00:00	31/03/2090 00:00	4	A	2002	COSTOS INDUSTRIALES
7	1	2001	IND000001	31/03/2005 00:00	17/08/2007 00:00	1	U	2004	INVESTIGACION OPERATIVA
8	1	2001	IND000001	27/12/2004 00:00	31/12/2007 00:00	4	A	2004	MECANICA Y MECANISMOS
9	1	2001	IND000001	16/12/2005 00:00	16/12/2007 00:00	1	U	2005	INSTALACIONES Y CONTROL
10	1	2001	IND000001	29/09/2004 00:00	31/12/2007 00:00	4	A	2004	FORMULACION Y EVALUACION DE PROYECTOS
11	1	2001	IND000001	12/04/2006 00:00	12/04/2008 00:00	2	U	2005	INVESTIGACION OPERATIVA
12	1	2001	IND000001	13/12/2004 00:00	31/12/2007 00:00	4	A	2004	TERMODINAMICA Y MAQUINAS TERMICAS
13	1	2001	IND000001	17/12/2003 00:00	31/03/2090 00:00	4	A	2003	ECONOMIA Y DIRECCION DE EMPRESAS
14	1	2001	IND000001	17/02/2005 00:00	17/08/2007 00:00	4	A	2004	INGENIERIA LEGAL
15	1	2001	IND000001	08/07/2005 00:00	17/08/2007 00:00	2	U	2005	MECANICA DE LOS FLUIDOS
16	1	2001	IND000001	24/09/2004 00:00	31/12/2007 00:00	4	A	2004	ELECTROTECNIA
17	1	2001	IND000001	10/10/2001 00:00	31/03/2090 00:00	4	A	2001	ORGANIZACION DE EMPRESAS
18	1	2001	IND000001	26/12/2001 00:00	31/03/2090 00:00	4	A	2001	ANALISIS MATEMATICO II
19	1	2001	IND000001	13/12/2004 00:00	31/12/2007 00:00	4	A	2004	TERMODINAMICA Y MAQUINAS TERMICAS
20	1	2001	IND000001	04/12/2002 00:00	31/03/2090 00:00	4	A	2002	COSTOS INDUSTRIALES
21	1	2001	IND000001	31/03/2005 00:00	17/08/2007 00:00	1	U	2004	INVESTIGACION OPERATIVA
22	1	2001	IND000001	27/12/2004 00:00	31/12/2007 00:00	4	A	2004	MECANICA Y MECANISMOS
23	1	2001	IND000001	16/12/2005 00:00	16/12/2007 00:00	1	U	2005	INSTALACIONES Y CONTROL
24	1	2001	IND000001	17/12/2003 00:00	31/03/2090 00:00	4	A	2003	ECONOMIA Y DIRECCION DE EMPRESAS
25	1	2001	IND000001	09/12/2003 00:00	31/03/2090 00:00	4	A	2003	ESTATICA Y RESISTENCIA DE MATERIALES

Industrial

Tabla 01: Tabla Industrial; Atributos seleccionados de la base de datos para realizar el presente análisis y predicción.

Análisis y Exploración de Datos

En el siguiente gráfico se muestra como están distribuidos los atributos condición de regularidad y resultado, que son los de mayor importancia al momento de definir el éxito académico.

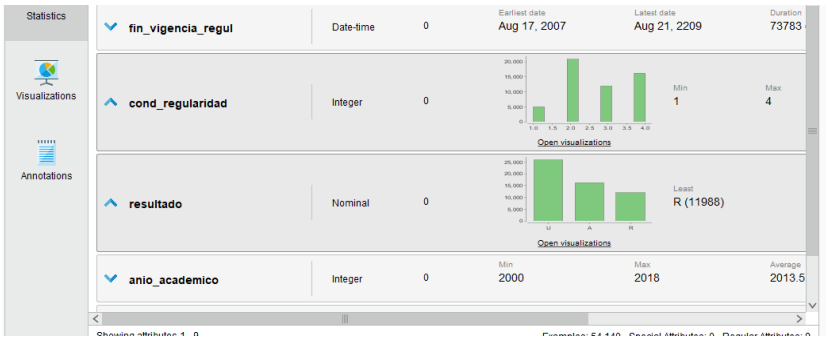


Gráfico 01: Muestra en gráfico de barras la distribución de los atributos: condición de regularidad y resultado.

En el Gráfico 02 de boxplot, el atributo cond_regularidad es de tipo numérico, se observa cómo se distribuyen los datos, particularmente se destaca que la mediana toma el valor 3 que son los que desaprobaron la materia, como primer cuartil se obtiene 2 que son los alumnos que abandonaron la cursada, mientras que el mínimo valor es 1, lo cual señala los que quedaron libres en la cursada. También se aprecia que no hay casos de alumnos que hayan promocionado con cond_regularidad = 5, por ese motivo coincide el tercer cuartil con el máximo y con el bigote superior. Aclarando además que los códigos que se consideran éxito académico son 4 y el 5, los cuales corresponden a los que obtuvieron la regularidad o la promoción respectivamente.

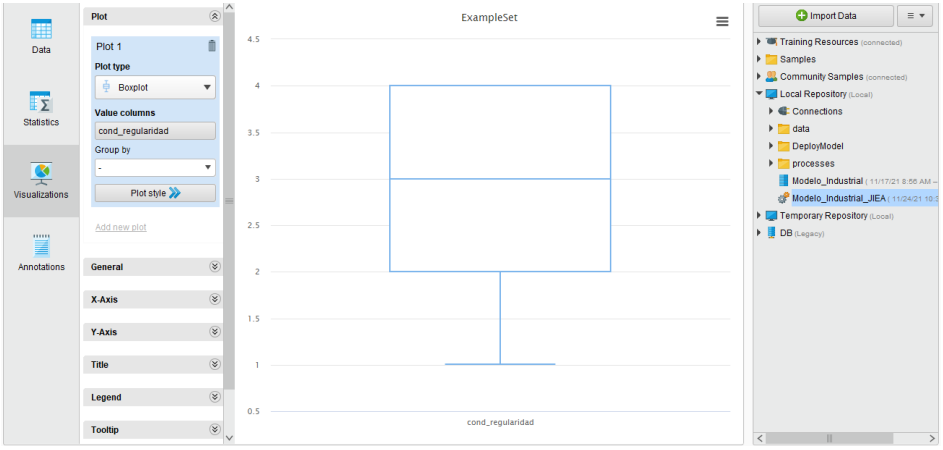


Gráfico 02: Boxplot del atributo cond_regularidad

En cambio para ver el atributo Resultado, que es de tipo categórico nominal, se puede observar en el gráfico siguiente, donde el éxito académico está asociado al código A, que corresponde con los que regularizaron o promocionaron la cátedra, mientras que R indica los que reprobaron la materia y U los que quedaron libres o abandonaron la materia según los datos de las tablas auxiliares.

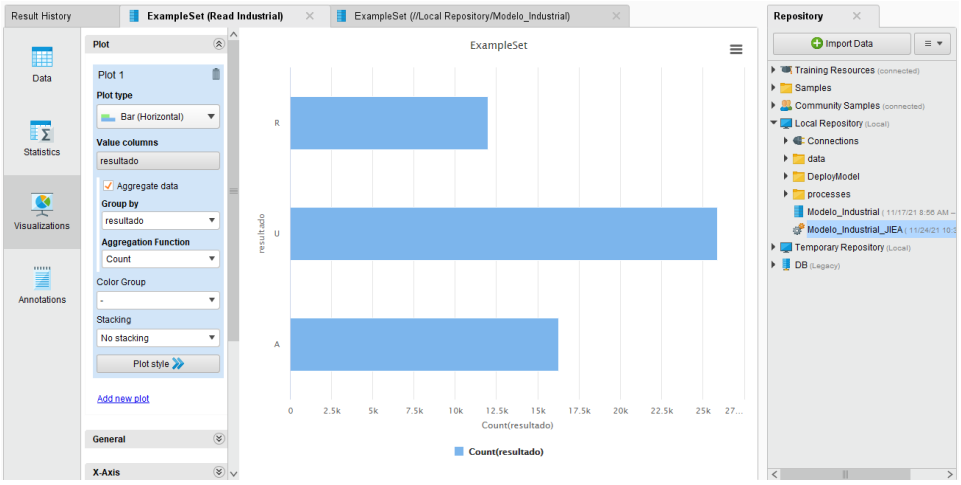


Gráfico 03: Gráfico de barras con el resultado de la cursada

Luego de probar con los modelos: Naive Bayes, Generalized linear model, Logistic Regresión, Fast Large Margin y Decision Tree, se observa que el mejor modelo es el Árbol de Decisión. Esto es debido a que los errores son mínimos para este modelo, y además los tiempos para procesarlos también son mínimos si vemos los gráficos correspondientes.

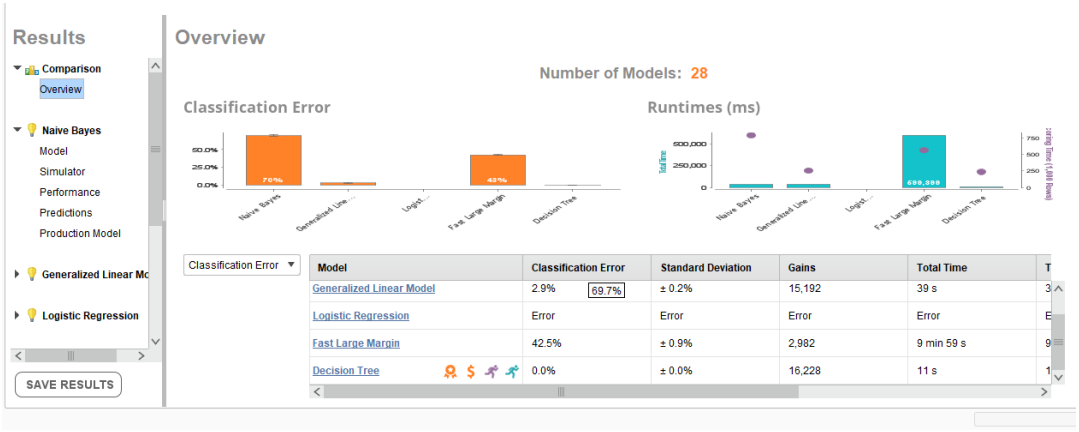


Gráfico 04: Indica que el mejor modelo para este trabajo es el Árbol de Decisión

Árbol de decisión - Simulador

Si se analiza el siguiente gráfico de barras, el mismo no indica que lo más probable es que se obtenga como resultado R: que son los que reprueban la cursada, seguido por U: que son los que abandonaron o quedaron libres, que amerita aclarar que debido al sistema de tutorías y otros programas o actividades, esto se redujo de manera importante en estos últimos diez años.

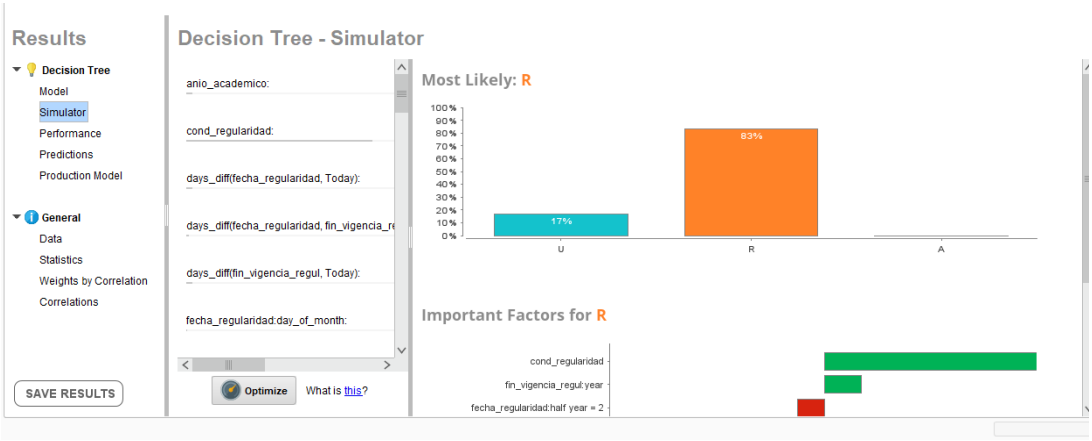


Gráfico 05: Resultados con mayor probabilidad dentro de la Simulación con los atributos del Árbol de Decisión

Ahora si se analiza en el siguiente gráfico los factores que más influyen para que predomine el resultado R, se puede afirmar que es cond_regularidad, acompañado principalmente por la fecha de fin de vigencia de la regularidad, ya que los alumnos no rinden y se les vence la materia, lo que significa que tienen que volver a cursar.

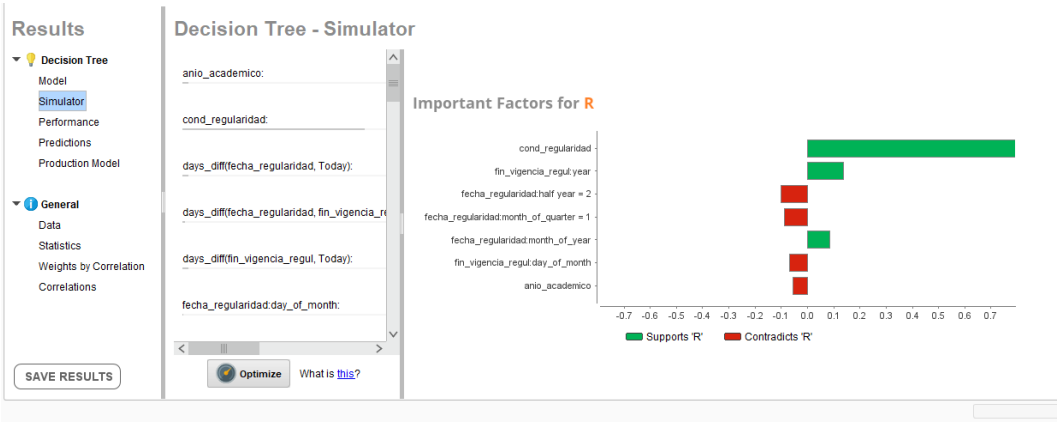
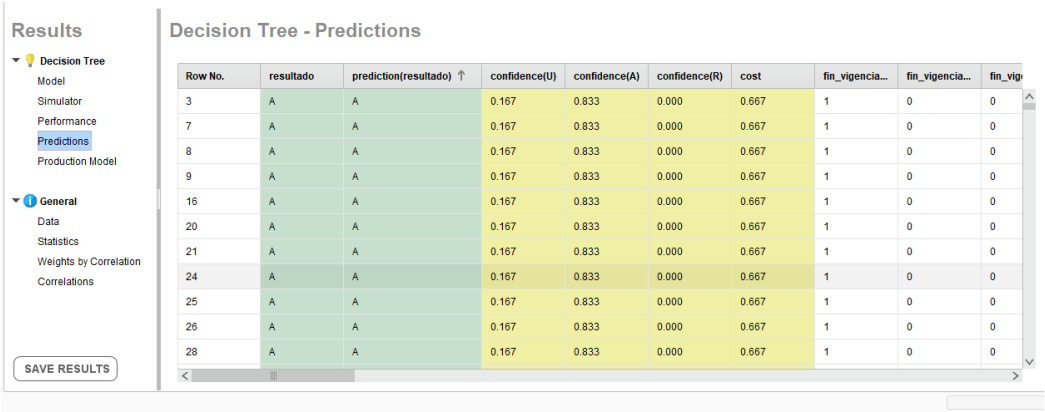


Gráfico 06: Factores más importantes que influyen en los resultados con los atributos del Árbol de Decisión

En la siguiente tabla de las predicciones del atributo resultado, con la confianza y costos asociados.



Row No.	resultado	prediction(resultado)	confidence(U)	confidence(A)	confidence(R)	cost	fin_vigencia...	fin_vigencia...	fin_vig
3	A	A	0.167	0.833	0.000	0.667	1	0	0
7	A	A	0.167	0.833	0.000	0.667	1	0	0
8	A	A	0.167	0.833	0.000	0.667	1	0	0
9	A	A	0.167	0.833	0.000	0.667	1	0	0
16	A	A	0.167	0.833	0.000	0.667	1	0	0
20	A	A	0.167	0.833	0.000	0.667	1	0	0
21	A	A	0.167	0.833	0.000	0.667	1	0	0
24	A	A	0.167	0.833	0.000	0.667	1	0	0
25	A	A	0.167	0.833	0.000	0.667	1	0	0
26	A	A	0.167	0.833	0.000	0.667	1	0	0
28	A	A	0.167	0.833	0.000	0.667	1	0	0

Tabla 02: Tabla de Predicciones del Árbol de Decisión

Para trabajar definir el árbol de decisión, utilizamos la tabla que mencionamos en Selección de datos, luego definimos el atributo Resultado como label con el operador Set Role, para poder aplicar el operador Árbol de decisión, quedando definido el diseño de la siguiente manera.

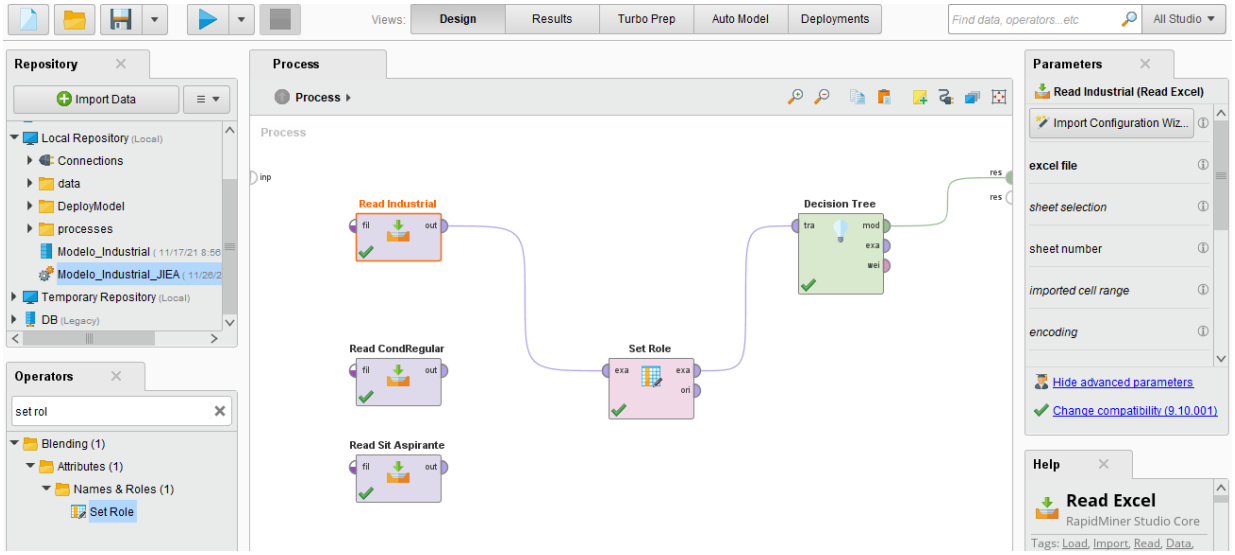


Gráfico 07: Gráfico del Diseño del Árbol de Decisión

A continuación tenemos el Árbol de decisión, que trabaja con las condiciones descritas en el Gráfico 09, donde se ve que se relacionan de manera particular el label que sería el Resultado con el atributo cond_regularidad. En donde si bien los códigos del atributo de condición de regularidad son números enteros, para poder armar las condiciones dentro del árbol de decisión se ve que si cond_regularidad es menor que 2.5 quiere decir si toma los valores 1: Libre o 2: Abandono, lo que implicaría que el resultado sería U. En caso de que cond_regularidad sea mayor o igual a 2.5, esto quiere decir que pueden pasar 2 cosas:

Que sea mayor que 3.5, esto implicaría que puede tener Código 4: Regular o código 5: Promociona, lo que en este caso daría por resultado A, que para nuestro caso se corresponde con el Éxito Académico.

Mientas que si es menor que 3.5 y mayor que 2.5, significa que estaría en el código 3, cuyo resultado sería U.

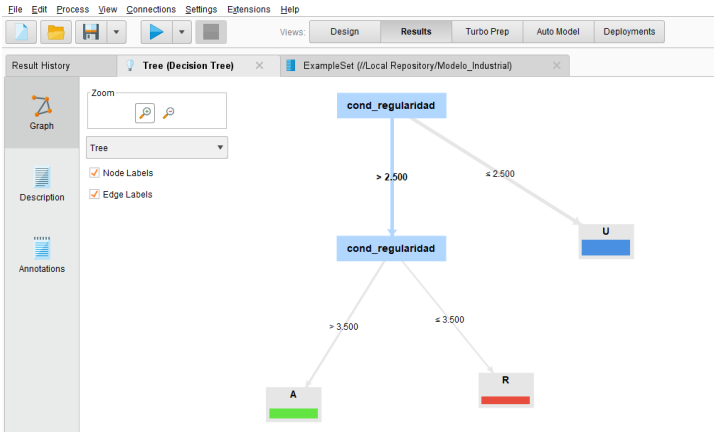


Gráfico 08: Gráfico del Árbol de Decisión - Modelo de Producción

En la Descripción de las condiciones que se cumplen en el árbol de decisión, se ven las cantidades de cada uno de los resultados, discriminadas también por el atributo cond_regularidad.

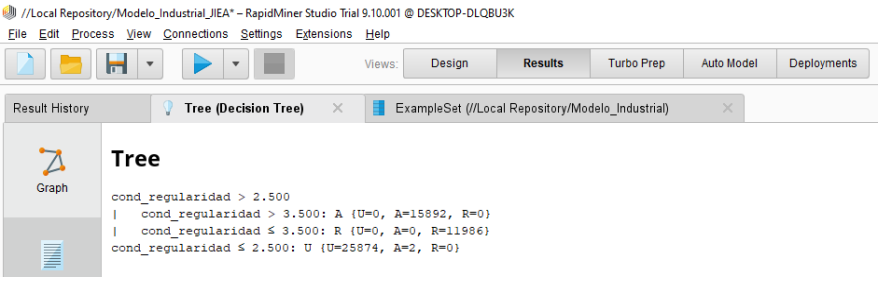


Gráfico 09: Descripción del Árbol de Decisión

Información General - Data

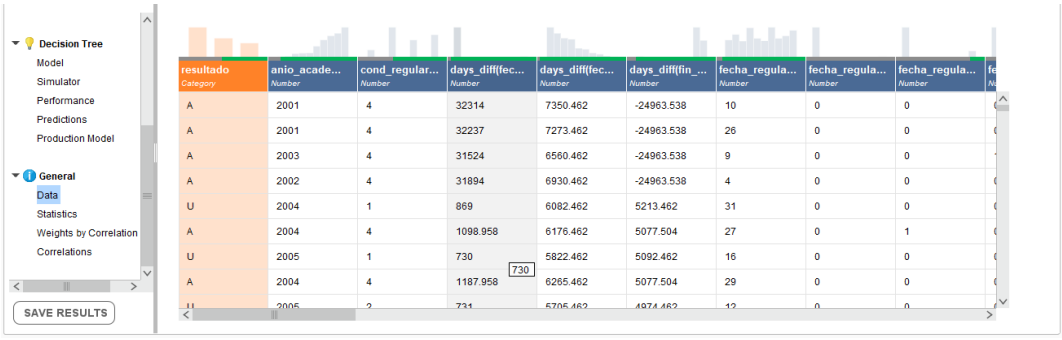


Tabla 03: Tabla Industrial; Atributos seleccionados de la base de datos para realizar el presente análisis y predicción.

El siguiente gráfico muestra cómo se distribuyen los datos con respecto a la condición de regularidad.

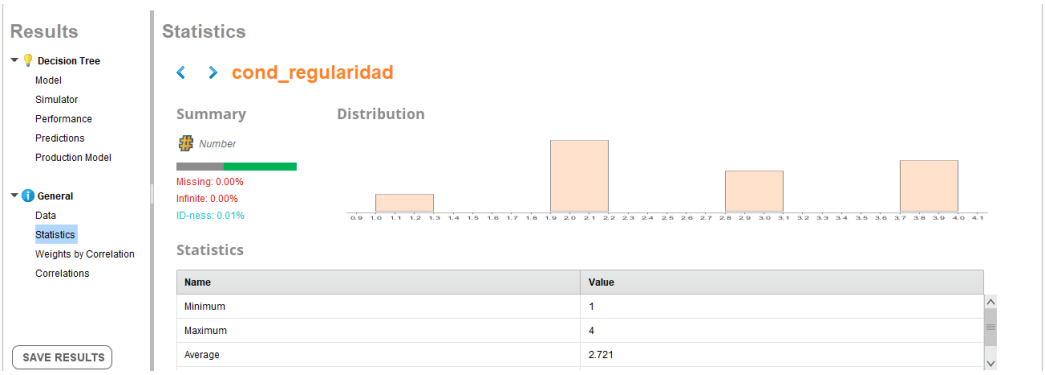


Gráfico 10: Información General - Estadísticas de Condición de Regularidad

En el siguiente gráfico se puede apreciar en el promedio de la diferencia entre fecha de regularización y fecha de fin de vigencia de regularidad son 822 días, lo cual es muy cercano a los 2 años y 3 meses, mientras que el tiempo de la vigencia de regularidad es de 2 años para esta carrera, pero como vence el 31 de marzo se suman tres meses más al promedio. Es decir que los alumnos tienen este tiempo para rendir tres veces para aprobar la materia, si no aprueban en esos intentos, pierden la regularidad de la misma.

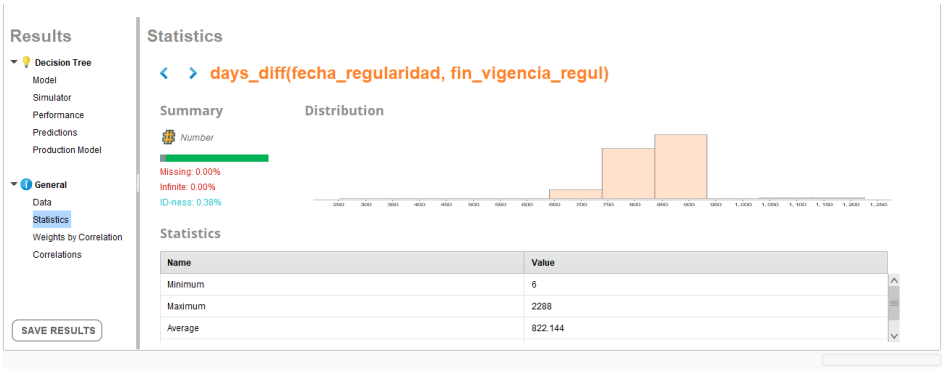


Gráfico 11: Información General - Estadísticas Diferencia entre fecha de regularidad y fin de vigencia de regularidad

Información general - Weights by Correlations. En el siguiente gráfico se ve que el mayor peso o la mayor correlación para el trabajo que estamos realizando lo tiene el atributo cond_regularidad con un peso de 0.626 superior a todos.

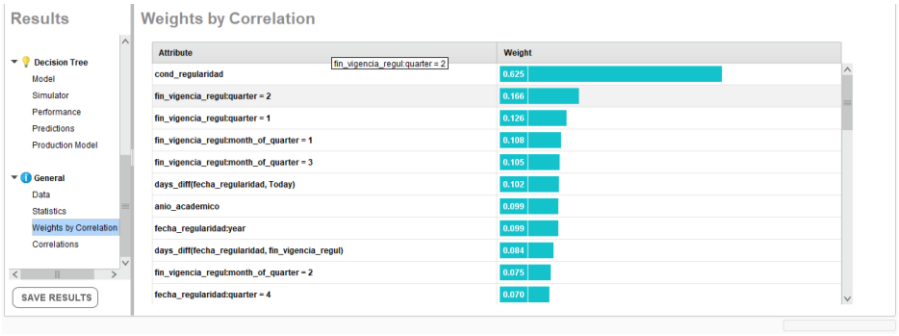


Gráfico 11: Información General - Weights by Correlations

CONCLUSIONES

En primer lugar en la Selección y Depuración de datos, se puede observar que el DataSet en estudio tenía inicialmente datos de todas las carreras y planes, para reducir la dimensionalidad del conjunto de datos se seleccionaron solamente alumnos de la carrera Ingeniería Industrial, con el plan 2001 que es el único plan actualmente de la misma. Luego de la selección, esos atributos quedaron eliminados, mientras que en las filas en el atributo fin_vigencia_regularidad se encontraron fechas con años atípicos como por ejemplo 2090 o 2209, que aparentemente fueron errores de tipeo, por lo que se eliminó un total de 386 filas, es decir, que en este proceso se eliminaron datos erróneos, faltantes o redundantes, que no aportan para realizar el presente trabajo. (Ver Tabla 01 y 02, Gráfico 01)

Luego de probar con los modelos según el Gráfico 04 Naive Bayes, Generalized linear model, Logistic Regresión, Fast Large Margin y Decision Tree, se observa que el mejor modelo para la predicción es el Árbol de Decisión. Esto es debido a que los errores son mínimos para este modelo, y además los tiempos para procesarlos también son mínimos.

Entre las conclusiones más importantes se puede ver que en el Gráfico 02 de boxplot, el atributo cond_regularidad es de tipo numérico, se observa cómo se distribuyen los datos, particularmente se destaca que la mediana toma el valor 3 que son los que desaprobaron la materia, como primer cuartil se obtiene 2 que son los alumnos que abandonaron la cursada, mientras que el mínimo valor es 1, lo cual señala los que quedaron libres en la cursada. También se aprecia que no hay casos de alumnos que hayan promocionado con cond_regularidad = 5, por ese motivo coincide el tercer cuartil con el máximo y con el bigote superior. Aclarando además que los códigos que se consideran éxito académico, los cuales corresponden a los que obtuvieron la regularidad o la promoción respectivamente.

Mientras que en el Gráfico 03, vemos el atributo Resultado, que es de tipo categórico nominal, se puede observar que el éxito académico está asociado al código A, que corresponde con los que regularizaron o promocionaron la cátedra, mientras que R indica los que reprobaron la materia y U los que quedaron libres o abandonaron la materia según los datos de las tablas auxiliares.

Como conclusión final del presente análisis descriptivo inteligente, podemos decir que se analizaron las características más importantes de las variables o atributos seleccionadas para poder evaluar la tendencia al éxito académico de los alumnos de la Carrera de Ingeniería Industrial de la Fac. de Ingenieria UNJu, pero el objetivo para un próximo trabajo sería analizar patrones del éxito académico, basándonos en otros atributos adicionales a los del trabajo actual, como ser: nivel educativo de los padres, la condición socio económica de la familia (planes sociales, estabilidad laboral,), si el alumno trabaja, tiene computadora, wifi en su casa (en época de pandemia), becas, distancia al lugar de estudio, transporte, entre otros, para poder predecir el éxito, fracaso académico o deserción en algunos casos.

BIBLIOGRAFÍA

- Fernández, M. (24 de enero del 2018). Egresan 8 mil ingenieros por año frente a 34 mil graduados de sociales, abogacía y psicología. Recuperado de https://www.infobae.com/educacion/2018/01/24/psicologos-y-abogados-pero-no-ingenieros-en-algunasdisciplinas-clave-se-reciben-menos-de-25-alumnos/
- Fischer, E. (2012). Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios. Tesis de maestría. Universidad de Chile. Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/111188/cffischer_ea.pdf ?sequence=1
- Goicochea, A. (2009). CRISP-DM: Una metodología para proyectos de minería de datos (artículo de blog). Recuperado de https://anibalgoicochea.com/2009/08/11/crispdm-una-metodologia-para-proyectos-de-mineria-de- datos/
- La Red, D., Karanik, M., Giovannini, M. y Scappini, R. (2009). Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios. XI Workshop de Investigadores en Ciencias de la Computación: WICC 2009, 7-8 de mayo. San Juan: Universidad Nacional de San Juan. Recuperado de http://sedici.unlp.edu.ar/handle/10915/ 53320
- Hernandez Orallo, 2005 - Introducción a la Minería de Datos, España: Editorial
- Bowen, O., & Agustín, A. (2016). Análisis de la deserción y permanencia académica en la educación superior aplicando minería de datos. Tesis Doctoral, Universidad Nacional de Colombia-Sede Bogotá.
- Fernández, S. F., Sánchez, J. M., Córdoba, A., & Largo, A. C. (2002). Estadística descriptiva. Esic Editorial.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.
- Kuna, H., García-Martínez, R. V., & Villatoro, F. (2010). Pattern Discovery in RES C.A.F.I. N°199/19 in University Students Desertion Based on Data Mining. IV Meeting on Dynamics of Social and Economic Systems, 2, págs. 275-285.
- Lattuada, M. (2017). Deserción y retención en las unidades académicas de educación superior. Una aproximación a las causas, instrumentos y estrategias que contribuyen a conocer y morigerar su impacto. Debate Universitario, 5(10), 100-113.
- Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. Expert Systems with Applications, 38(12), 14984-14996.
- Sánchez, E. G., & Chinchilla Brenes, S. (2005). Detección de estudiantes en riesgo académico en el Instituto Tecnológico de Costa Rica. Revista Educación, 29(2), 123-138.
- Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. IEEE International Work Conference on Bioinspired Intelligence (IWOBI) , (págs. 1-6).
- Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2015). Finding patterns with a rotten core: Data mining for crime series with cores. Big Data, 3(1), 3-21.
- Ware, C. (2004). Information Visualization - Perception for Design. Morgan-Kaufmann.
- Ware, C. (2008). Visual Thinking for Design. Morgan Kaufman/Elsevier.

- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. Pearson Educación S.A.
- [Documentación – Rapidminer, 2020], Recuperado de: https://docs.rapidminer.com/ Fecha de consulta: 19/11/2021.
- [Ventajas Desventajas Algoritmos de Clasificación], Recuperado de: https://www.youtube.com/watch?v=T8aCfSBlrqU&list=PLJjOveEiVE4Dk48EI7I-67PEleEC5nxc3&index=57 Fecha de consulta 5/11/2021
- [Rapidminer Tutorial], Recuperado de: https://www.youtube.com/watch?v=m74LNl-1J44 Fecha de consulta 8/11/2021
- [Rapidminer Arbol de decisiones], Recuperado de: https://www.youtube.com/watch?v=Z6vp8LPxCSQ Fecha de consulta 10/11/2021
- [Rapidminer Agrupamientos con K-Means], Recuperado de:https://www.youtube.com/watch?v=2532w8T7KnU Fecha de consulta 11/11/2021

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Optimización de la estrategia de monitoreo de un sistema de distribución de agua potable mediante análisis estadístico multivariado

Corimayo, S. N.[1]; Rajal, V. B.[1,2] y Cruz, M. C.[1]*

1 Instituto de Investigaciones para la Industria Química (INIQUI, UNSa - CONICET)

2 Facultad de Ingeniería – Consejo de Investigación de la Universidad Nacional de Salta

*sheicorimayo.96@gmail.com, *mccruz@conicet.gov.ar*

RESUMEN

La calidad del agua potable es crucial para la salud humana, por lo que se debe realizar una vigilancia sanitaria, definiendo puntos estratégicos y frecuencia adecuada según las normativas locales y los recursos disponibles. El objetivo de este estudio fue monitorear la calidad del agua y aplicar un análisis estadístico multivariado para determinar la frecuencia mínima y cantidad óptima de puntos de muestreo para una vigilancia eficaz. Se recogieron datos sobre la calidad microbiológica y fisicoquímica del agua de una red de distribución durante nueve meses. Se aplicó un análisis de componentes principales (ACP) para identificar los parámetros que contribuían significativamente a la variación de la calidad del agua en los sitios. Se utilizó el análisis de conglomerados (AC) según los meses de monitoreo, con el propósito de evaluar si valores similares de temperatura (estaciones) tenían igual características de agua potable. La calidad del agua fue aceptable según la normativa, excepto en el 11 % de las muestras. El ACP permitió identificar tres puntos de muestreo estratégicos a partir de ocho iniciales. El AC mostró que se debe monitorear por lo menos un mes correspondiente a las distintas estaciones; sin embargo, se debe aumentar la frecuencia según la tasa de residencia.

Palabras Clave: Agua potable - análisis multivariado - diseño óptimo de monitoreo - calidad de agua

INTRODUCCIÓN

La calidad del agua es crucial para la salud humana, por lo que el agua potable se protege, controla y gestiona cuidadosamente (Proctor y Hammes, 2015). La calidad de la misma en los sistemas de distribución de agua potable (SDAP) depende de diferentes parámetros fisicoquímicos y microbiológicos y para determinarla, se debe seguir la recomendación de la Organización Mundial de la Salud (OMS, 2017) y la normativa del Código Alimentario Argentino (CAA, 2019).

La estabilidad biológica del agua potable se refiere a la baja o nula variación de su calidad microbiológica entre el punto de producción y el grifo donde es consumida (Prest et al., 2016). Para salvaguardar la estabilidad biológica del agua potable se debe realizar una vigilancia sanitaria desde la fuente de abastecimiento o puntos de almacenamiento hasta el grifo, definiendo puntos estratégicos que sean representativos del sistema, con la frecuencia que la situación lo amerite para garantizar la salud de la población (Organización Panamericana de la Salud, 2013). Para el agua potable, los sitios de muestreo en los diferentes esquemas de vigilancia, se encuentran principalmente en la planta de agua o en el punto final de consumo (Dong et al., 2015). En Argentina, la frecuencia es fijada por la autoridad nacional competente según la cantidad de personas consumidoras, indicando por ejemplo muestras mensuales en una población abastecida de 100.000 habitantes (Ley N° 26221, 2007). Sin embargo, la determinación de la frecuencia de muestreo como los sitios representativos, deben basarse en necesidades de conocimiento realistas y prácticas, y planificarse dentro de los recursos humanos, financieros y técnicos disponibles, así como de las obligaciones legales y políticas (Behemel et al., 2016). Debe tenerse en cuenta que una gran parte de los costes de funcionamiento del monitoreo está directamente relacionada con la frecuencia de muestreo y la cantidad de sitios estudiados (Strobl y Robillard, 2008). Lo más probable es que los limitados recursos financieros no permitan la misma frecuencia de muestreo de todos los parámetros de cada sitio, y tampoco permitan la inclusión de todos los sitios deseables de muestreo.

Para comprender mejor la calidad del agua de los SDAP, los datos de la vigilancia suelen interpretarse aplicando técnicas estadísticas. Estas técnicas permiten identificar los posibles factores que influyen en la calidad del agua (Bu et al., 2010). Los métodos de análisis de datos multidimensionales son muy divulgados en los estudios ambientales que se ocupan de la medición y el seguimiento de la calidad del agua (Oketola et al., 2013). Los análisis más comunes que se utilizan son el análisis jerárquico de conglomerados (AC) o *clustering*, y el análisis de componentes principales (ACP) (Howladar et al., 2017). El análisis de conglomerados (AC) comprende métodos multivariados exploratorios y no supervisados, lo que significa que no requiere una estructura de agrupación conocida *a priori*, que encuentran grupos de datos (Jain, 2010). El objetivo principal de este análisis es unir objetos que son más similares entre sí en una misma clase, teniendo en cuenta algunas propiedades determinadas (Howladar et al., 2017). El ACP es una poderosa técnica de reconocimiento de patrones que intenta explicar la varianza de un gran conjunto de variables intercorrelacionadas y transformarlas en un conjunto más pequeño de variables independientes (no correlacionadas) **y puede identificar asociaciones entre observaciones, variables y entre variables y observaciones** (Bu et al., 2009).

El ACP y el AC pueden aplicarse a un conjunto de variables de calidad del agua y descubrir subconjuntos coherentes y relativamente independientes entre sí (Oketola et al., 2013). Como resultado, estos subconjuntos de datos pueden representar aproximadamente la misma cantidad de información que el conjunto mucho más amplio de las observaciones originales. En este sentido, estas técnicas pueden ser de gran utilidad para intentar reducir la

dimensionalidad del conjunto de datos que se deben obtener en una gestión fiable de la calidad del agua.

Es por ello que el objetivo del presente trabajo fue mejorar la vigilancia sanitaria reduciendo costos y esfuerzos mediante el diseño de un monitoreo inteligente y eficaz de la calidad del agua, utilizando análisis estadísticos multivariados que puedan proporcionar información útil en la selección de cantidad de sitios y tiempos mínimos de muestreo.

METODOLOGÍA

Lugar de estudio

El sitio de estudio se encuentra en el valle de Lerma, en el sector norte de la ciudad de Salta (Argentina). Es un sistema de distribución de agua cerrado perteneciente a un complejo educativo que representa un modelo de mini-ciudad. Además de aulas, cuenta con talleres de obras y de servicios, un jardín de infantes, colegio secundario, unidad sanitaria, gimnasio y comedores y, existe diversidad de materiales y de años de uso de las cañerías. El suministro de agua del establecimiento proviene de dos fuentes: un acueducto (agua subsuperficial) y un pozo propio de 156,9 m, en cuya salida se le realiza cloración, y luego se mezcla y almacena en la cisterna de 10 m de diámetro y 2 m de profundidad, aproximadamente.

Recolección de muestras de agua

Se tomaron muestras de agua de grifos de seis puntos correspondientes a distintos edificios distribuidos a diferentes distancias (P3 a P8), del tanque elevado de distribución (P2), y de la cisterna (P1) (Figura 1), mensualmente desde el mes de marzo a agosto de 2020 y desde noviembre de 2020 a enero de 2021. Durante el primer mes de monitoreo, el establecimiento educativo presentaba asistencia del personal administrativo, estudiantes, niños en la guardería maternal y colonia de vacaciones (existía un consumo considerable de agua). Los meses siguientes fueron muestreados en el transcurso de la cuarentena. Durante esta situación única, hubo una reducción en la concurrencia de personas, lo cual provocó una disminución marcada del consumo de agua y, por lo tanto, un aumento del tiempo de residencia del agua (TR) en las tuberías.

Análisis fisicoquímicos

Se midieron *in situ* los parámetros fisicoquímicos como la temperatura (Temp, °C), el pH, la conductividad (Cond, µS/cm), la salinidad (Sal, %) y la turbidez (Turb, NTU), utilizando una sonda multiparamétrica U-10 HORIBA. Para ello, se utilizaron recipientes de plástico de 500 mL previamente lavados (con agua destilada) y enjuagados dos veces con el agua a estudiar. Se realizó la determinación de la concentración de cloro total (CT, mg/L) y libre (CL, mg/L) en el laboratorio por espectrofotometría utilizando el kit comercial Cloro Total *AQAssay* siguiendo las instrucciones del fabricante. Estas muestras se colectaron en vasos de análisis estériles de 100 mL, se guardaron en contenedores con hielo y se transportaron al laboratorio para su procesamiento posterior en un plazo de dos horas. Además, se registró el TR (expresado en meses) para cada edificio.

Análisis microbiológicos

Las muestras de agua se analizaron de acuerdo a normas nacionales e internacionales (Eaton et al, 2005; OMS, 2017; CAA, 2019) para calidad del agua potable y se verificó si las mismas cumplían con la normativa vigente.

Se recolectó 1 L de agua en recipientes limpios y estériles que contenían tiosulfato de sodio al 1,8 % (p/v) para neutralizar el efecto del desinfectante. Las muestras se preservaron en hielo en una conservadora hasta su traslado al laboratorio y se mantuvieron refrigeradas

hasta realizar los análisis bacteriológicos correspondientes dentro de las seis horas de recolección. Se determinó la presencia de bacterias coliformes totales y termotolerantes mediante el método del Número Más Probable. Se detectó la ocurrencia de organismos patógenos oportunistas con medios específicos y selectivos para *Escherichia coli* (mTEC, Sigma Aldrich) a 44 °C durante 48 horas, *Enterococcus* sp. (ME, Difco), *Pseudomonas aeruginosa* (Cetrimide), *Salmonella* sp. (SS agar) y recuento de bacterias mesófilas aerobias (MAT) en Agar Plate Count (APC) incubados a 37 °C durante 48 h, por el método de filtración por membrana empleando 100 mL de muestra. Además, se realizó un recuento de bacterias heterótrofas de aguas tratadas (HAT) en agar Reasoner 2A (R2A) incubadas a 21 °C durante 7 días. Este medio es recomendado para muestras de agua potable, dado su bajo contenido nutricional, ya que estimula el desarrollo de bacterias estresadas, de crecimiento lento y tolerantes al desinfectante (Eaton et al., 2005).

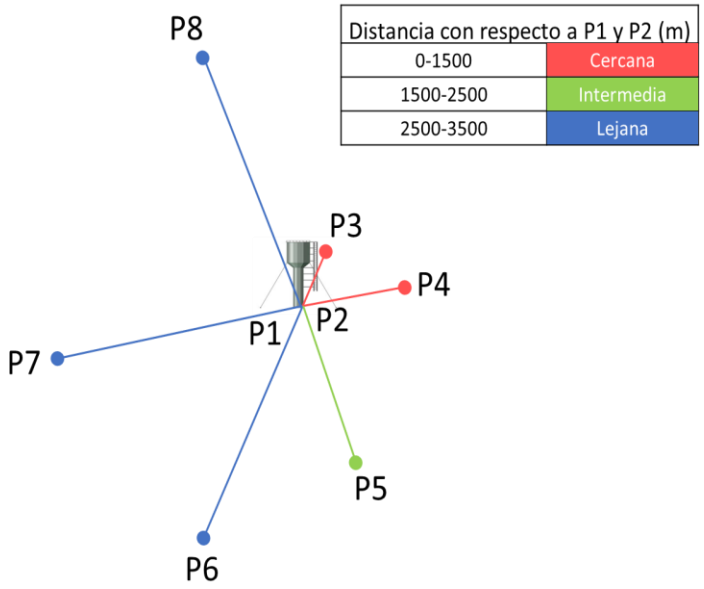


Figura 1. Representación de los sitios de muestreo con su distancia relativa (longitud de radios) al tanque (P2) y cisterna (P1). El color del radio de cada punto indica si se encuentra a una distancia cercana (< 1500 m), intermedia (1500 a 2500 m) o lejana (> 2500 m) de los tanques de almacenamiento.

Análisis de datos

Se realizó un análisis estadístico multivariado para evaluar los componentes dominantes que explican las variaciones en la calidad del agua en los distintos sitios y meses muestreados. Esta evaluación se realizó mediante dos técnicas: Análisis de Componentes Principales y Análisis de Conglomerados. En el ACP, un autovalor proporciona una medida de la importancia del componente; aquellos que son iguales o mayores a 1,0 se consideran significativos (Oketola et al. 2013). Por otro lado, se aplicó el criterio de Liu (2003), para evaluar las contribuciones significativas de cada variable a los componentes principales. Valores de carga mayores a 0,75 se consideran de correlación fuerte, valores entre 0,75 y 0,5 de correlación moderada y se consideran sin correlación los valores menores que 0,5. Se realizó esta metodología con el fin de evaluar los parámetros más significativos que afectan la calidad del agua en los distintos sitios de la red de distribución. Además, se generaron *biplots* **con el fin de visualizar observaciones y variables en un mismo espacio, así fue posible identificar asociaciones entre las observaciones (sitios).** En el AC, los niveles de similitud en los que se unen las observaciones similares se utilizaron para construir un dendrograma (Oketola et al. 2013). Para este estudio, se construyeron dendrogramas utilizando el método de Ward y la distancia euclidiana al cuadrado con el propósito de evaluar si valores similares de temperatura (estaciones) tenían igual calidad de agua potable.

Además, se evaluaron los datos con métodos no paramétricos debido a que no se ajustaban a una distribución normal (p-valor < *0,001*). Se testeó la bondad de ajuste a la distribución normal para todos los datos, aplicando el test-W de Shapiro–Wilks. Se calculó el coeficiente de correlación de Spearman (ρ) para determinar la correlación entre las principales

variables ambientales (Temp del agua, CL y TR) y microbiológicas (recuento de microorganismos en R2A y APC). Se usó el test de Kruskal-Wallis para determinar la variación estacional de las variables monitoreadas en los ocho sitios. En este estudio se compararon varios grupos dados por los meses o sitios, y se determinó si existe una diferencia estadísticamente significativa entre sus medianas.

Todos los análisis estadísticos se realizaron utilizando el programa InfoStat (Di Rienzo et al., 2018).

DESARROLLO

Características fisicoquímicas del agua potable

Se recolectaron mensualmente en total 72 muestras, de ocho sitios de la red de distribución de agua potable durante nueve meses. La conductividad mostró los valores promedios más bajos durante el mes de marzo y noviembre (253 y 254 µS/cm, respectivamente) y no presentó variaciones significativas (*p*-valor > 0,05) durante los otros meses (248 - 287 µS/cm) (Figura 2, A). La turbidez presentó valores de 0 a 3 NTU, por lo cual su rango de variación se encontraba dentro de valores permitidos (máx. 3 NTU) para el agua potable según la legislación vigente (CAA, 2019). Sólo un sitio presentó 11 NTU de turbidez durante el mes de noviembre debido a un largo período de estancamiento de agua. Los porcentajes de salinidad mostraron una mínima variación de 0 a 0,01 % en todos los meses.
El pH tuvo los valores promedios más bajos en el mes de marzo (7,39) (Figura 2, B). Se observó un incremento en los valores de pH (> 8,5) por encima del permitido según el CAA (CAA, 2019) en un sitio durante abril y agosto y, en dos en julio, noviembre y enero.
La temperatura del agua de los sitios provenientes de los edificios (P3 a P8) presentó variaciones estacionales (12,7 – 26,7 °C), lo cual se esperaba ya que la temperatura del agua dentro de las tuberías depende de la temperatura ambiental (Figura 2, C). En cambio, la temperatura del agua almacenada en la cisterna y el tanque (P1 y P2) no presentó variaciones significativas (*p*-valor > 0,05) durante el muestreo (17,8 – 22,9 °C).

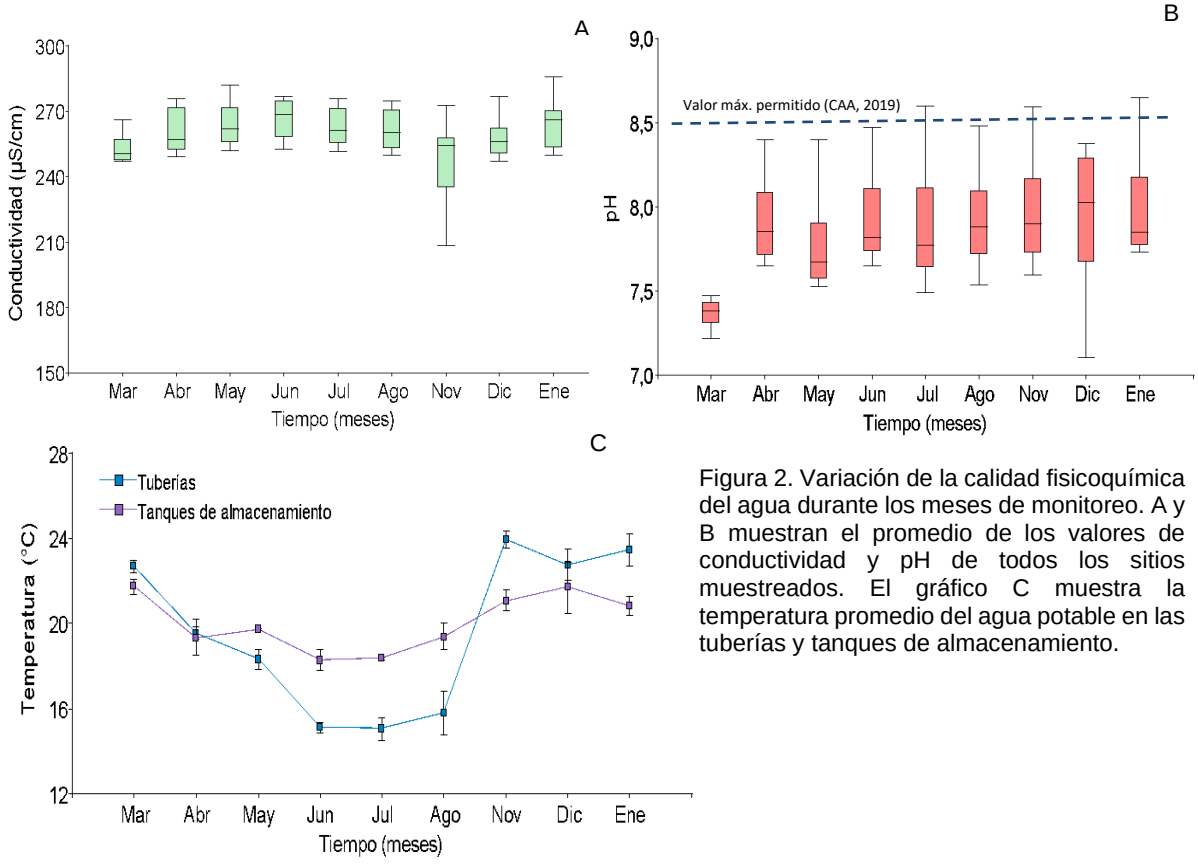


Figura 2. Variación de la calidad fisicoquímica del agua durante los meses de monitoreo. A y B muestran el promedio de los valores de conductividad y pH de todos los sitios muestreados. El gráfico C muestra la temperatura promedio del agua potable en las tuberías y tanques de almacenamiento.

Durante el mes de marzo las concentraciones de CL fueron similares (0,34 – 0,44 mg/L, *p*-valor > 0,05) en todos los sitios muestreados (Figura 3). La concentración de cloro libre del agua varió significativamente en los meses posteriores (0,0 – 1,66 mg/L, *p*-valor < 0,05). El CAA (2019) establece que la concentración de cloro residual en los sistemas de distribución de agua no debe ser menor que 0,2 mg/L. El 11 % (8/72) de las muestras presentaron concentraciones de CL menores al permitido. El CL se correlacionó negativamente con la temperatura (R2 = -0,42, *p*-valor < 0,05) y no se encontró una correlación con el TR (*p*-valor > 0,05).

La mayoría de los sitios, excepto los de almacenamiento, experimentaron un aumento del TR. Hasta el mes de noviembre la mayoría de los edificios muestreados estuvieron hasta seis meses sin consumo de agua. Sólo los sitios P5 y P6 experimentaron un aumento del TR hasta enero de 2021 de 11 meses. El sitio P7, fue el único edificio en el cual el agua se utilizó todos los meses por el personal del lugar, por lo que el TR fue nulo.

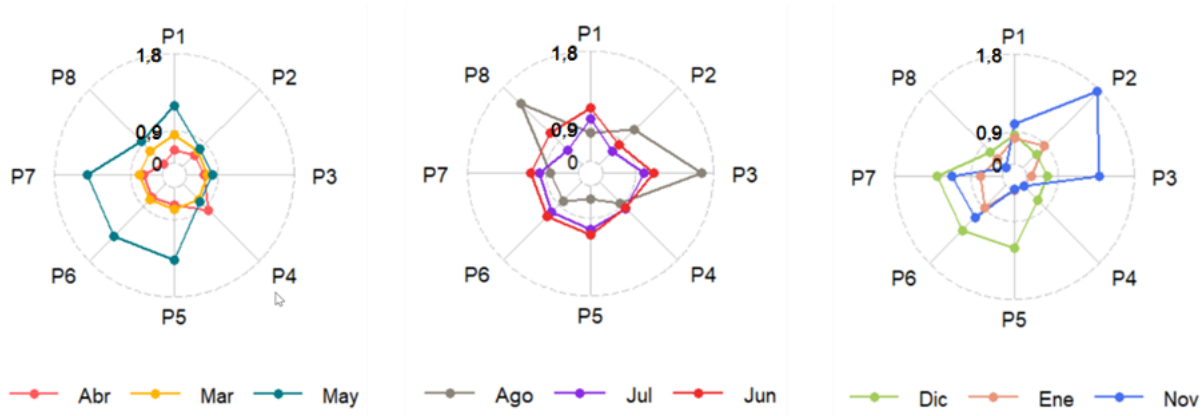


Figura 3. Concentración de cloro libre (mg/L) encontrados en los ocho sitios (P1-P8) durante los meses de muestreo.

Calidad microbiológica del agua potable

Durante los nueve meses de monitoreo, las muestras de agua cultivadas en agar APC, presentaron un valor máximo de 757 UFC/100 mL de muestra (Figura 4). En todos los casos la cantidad de organismos mesófilos totales no superó el valor permitido por el CAA, 2019 (500 UFC/mL). Se observó que la cantidad promedio de colonias de las muestras provenientes de grifo fue de 56,5 UFC/100 mL, mientras que en los tanques de almacenamiento fue de 16,9 UFC/100 mL. No se encontraron diferencias significativas de esta variable, entre las muestras de red y los tanques de almacenamiento (*p*-valor > 0,05). El recuento de microorganismos en este medio fue muy variable a lo largo de los meses y no se correlacionó con la Temp, el CL ni el TR (*p*-valor > 0,05).

En contraste con los resultados de APC, los cultivos en R2A mostraron mayor crecimiento (Figura 5). Se cuantificó hasta 9,7 x 10^3 UFC/100 mL, y se encontró que la cantidad mínima fue de 20,5 UFC/100 mL. Este medio promueve el crecimiento de organismos oligotróficos que se encuentran estresados en un ambiente como el agua potable tratada con desinfectante residual, por lo que se espera mayor cantidad de microorganismos comparado con APC (Eaton et al., 2005). Se determinó una gran variabilidad entre los distintos sitios a lo largo del tiempo (*p*-valor < 0,05). Estos valores se correlacionaron positivamente con el TR (R^2 = 0,46, *p*-valor < 0,05), negativamente con el CL (R^2 = -0,27, *p*-valor < 0,05), y no se encontró una correlación con la Temp (*p*-valor > 0,05). Se determinó que la concentración promedio en los tanques de

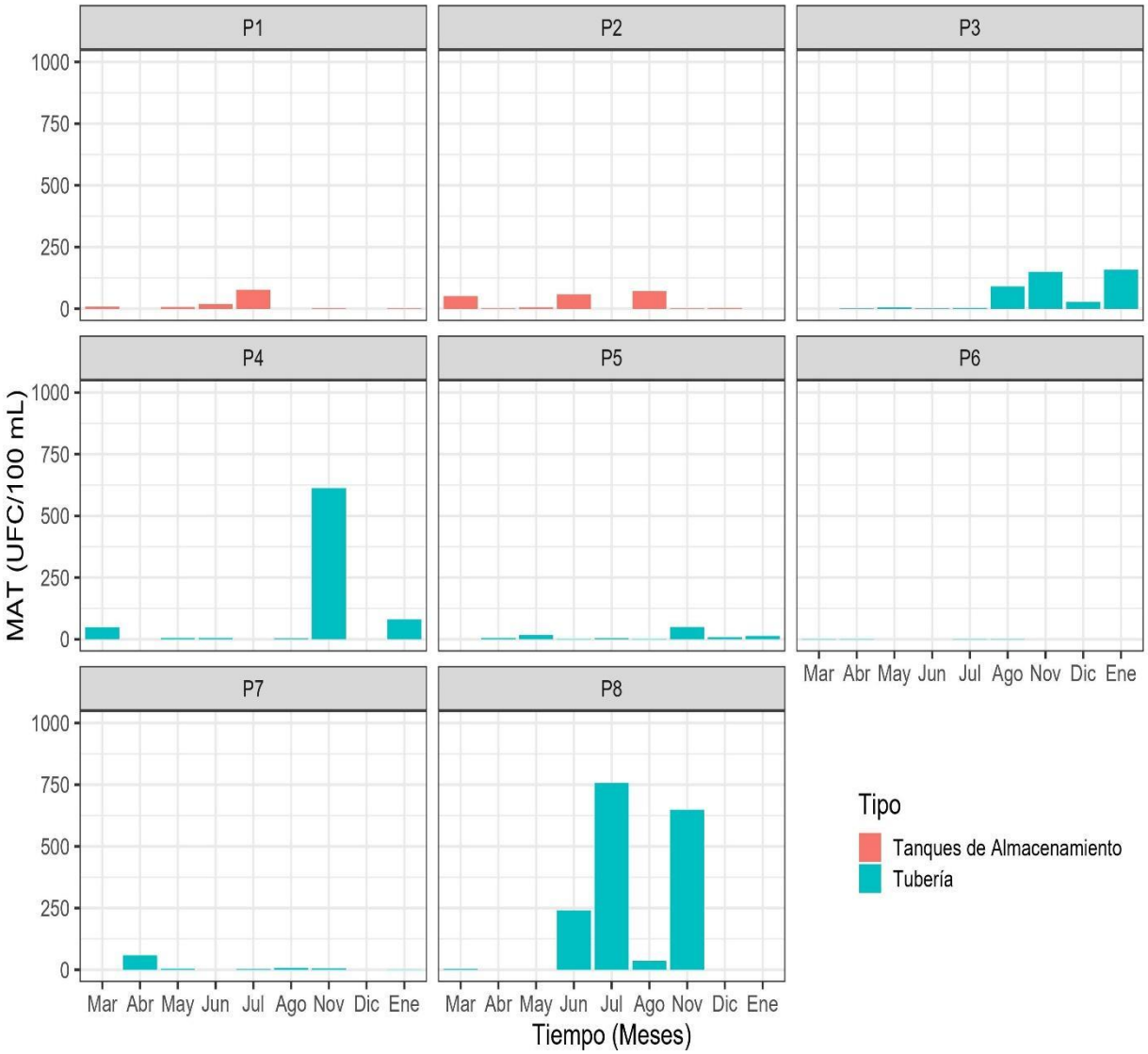


Figura 4. Concentración total de mesófilos aerobios totales (MAT) determinada en Agar Plate Count y expresada en unidades formadoras de colonias (UFC) en 100 mL de muestra, de todos los sitios estudiados durante los meses de muestreo. Se usó la misma escala del eje vertical para facilitar la comparación.

almacenamiento fue de 2,5 x 10^2 UFC/100 mL, un orden menor a la encontrada en la red de distribución (2,1 x 10^3 UFC/100mL). No se observó la presencia de microorganismos indicadores y/o patógenos oportunistas en ninguno de los puntos muestreados.

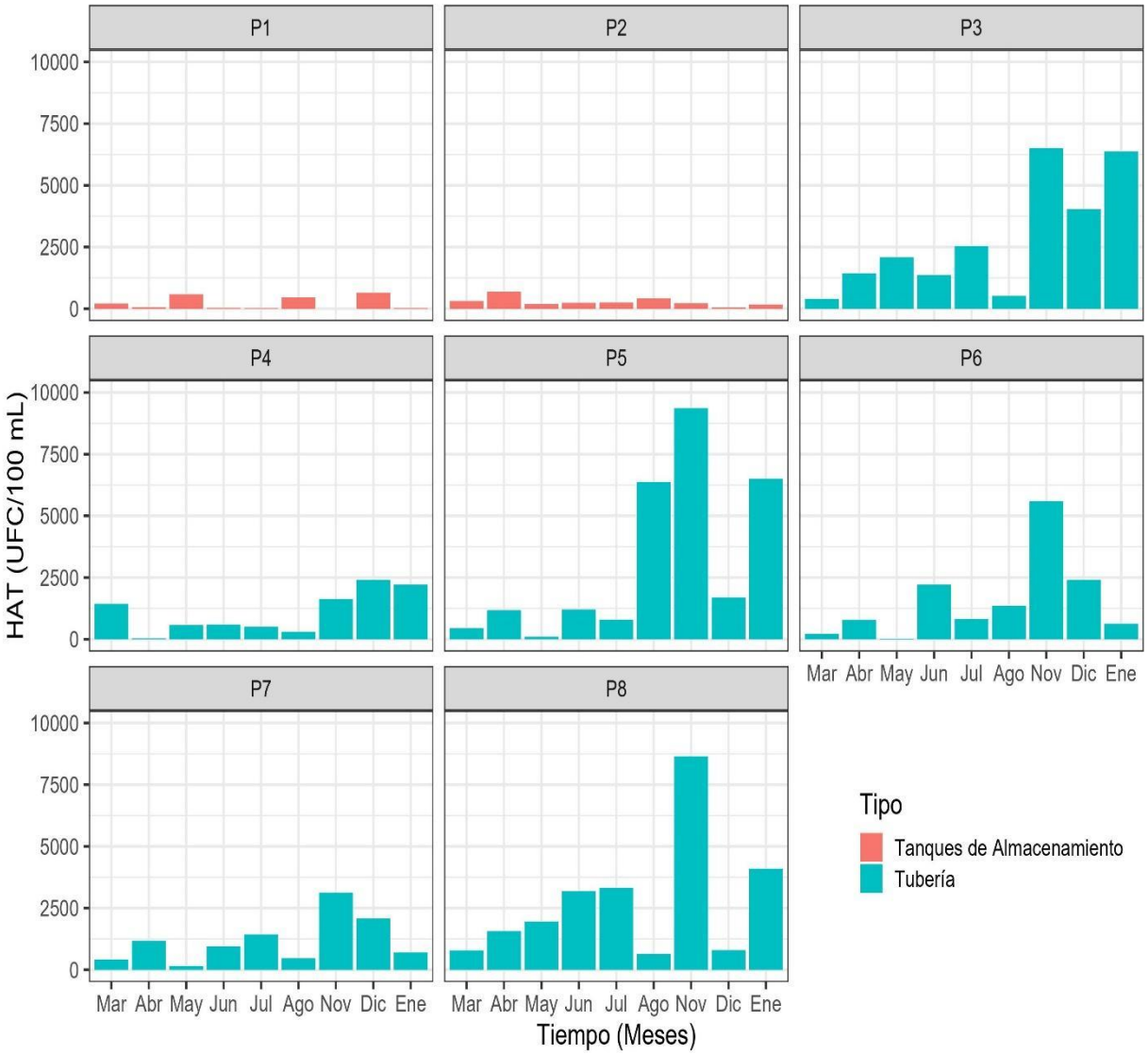


Figura 5. Concentración total de heterótrofos de aguas tratadas (HAT) determinada en el medio R2A, expresada en unidades formadoras de colonias (UFC) en 100 mL de muestra, de todos los sitios estudiados durante los meses de muestreo. Se usó la misma escala del eje vertical para facilitar la comparación.

Análisis de conglomerados

Se realizaron dendrogramas según los meses de muestreo, de las 72 muestras de agua, teniendo en cuenta la Temp, Cond, pH, CL, recuento de HAT y MAT, TR y la distancia de los edificios con respecto a P1 y P2 (Dist) de los ocho sitios en estudio. Los diferentes meses de muestreo del sistema de distribución de agua se agruparon según características de agua potable similares. Este procedimiento de agrupación, generó dos grupos y un mes *outlier*. El grupo I estuvo conformado por los meses que comprenden la estación invernal (junio, julio, agosto), con menores temperaturas (16,2 ± 2,0 °C), y por el mes de mayo, que pertenece a otoño con una Temp promedio de 18,6 ± 1,2 °C. Este grupo demostró que los meses con las menores Temp tienen características similares del agua potable ya que se encontraron más cercanos entre sí, y mayo se distanció dentro del grupo por tener una Temp estacional media

Sin embargo, se encontró que el grupo II, estuvo conformado por meses con diferentes Temp: abril (Temp media) y diciembre, enero y marzo (altas Temp). El mes abril tuvo una Temp promedio de 19,5 ± 0,7, mientras que los otros meses tuvieron una Temp promedio de 22,6 ± 1,5 °C. Los meses de diciembre y enero presentaron una distancia de 1 con respecto a abril, mientras que se encontraron muy distanciados respecto a marzo (Figura 6).

La Temp es una variable que tiene una influencia notable en la calidad del agua. En la mayoría de los casos, las características del agua fueron semejantes para los meses con temperaturas próximas. Marzo se distanció de los otros meses dentro del grupo II, posiblemente debido a que el tiempo de estancamiento fue nulo durante este mes en todos los puntos y presentó los valores de pH más bajos del monitoreo, indicando que estas variables son importantes y deben ser consideradas para el diseño de muestreo. El mes de noviembre se separó del resto ya que además de las elevadas Temp (23,2 ± 1,6), presentó valores altos de HAT y un mayor tiempo de estancamiento del agua en la mayoría de los sitios (durante diciembre, el consumo de agua aumentó en varios sitios). Estos resultados indicaron que, en la vigilancia de la calidad del agua en el tiempo, no es suficiente seleccionar meses que abarquen distintos rangos de Temp, se deben considerar además los TR del agua en los edificios muestreados.

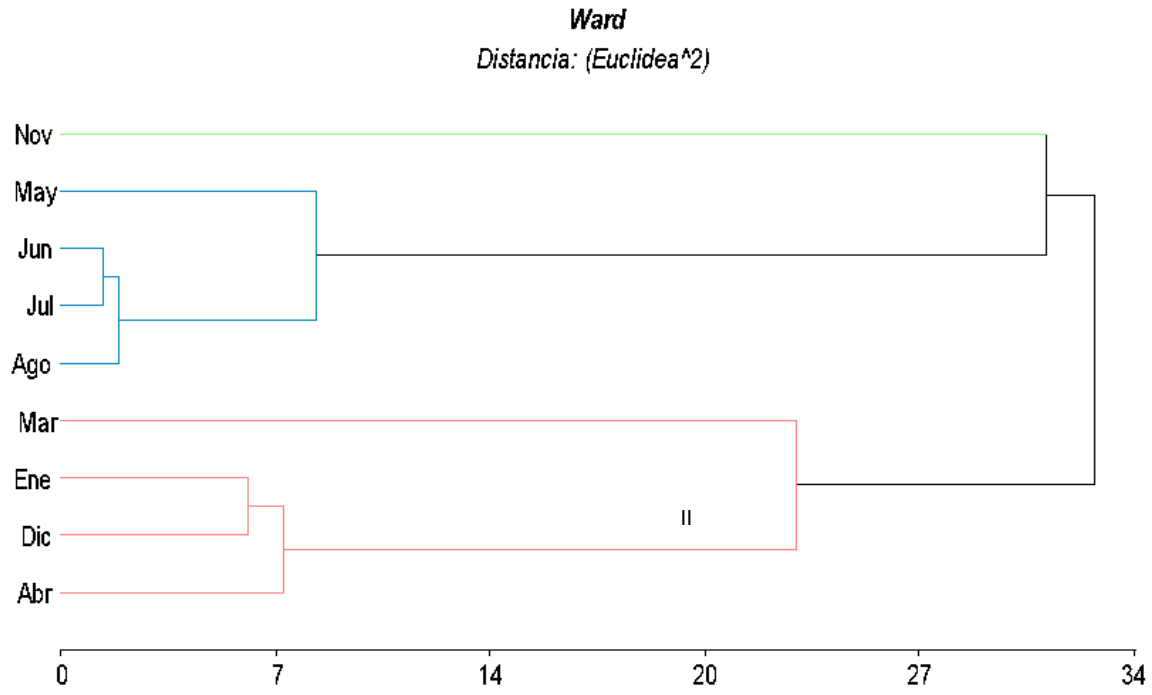


Figura 6. Análisis de conglomerados. Dendrograma de los nueve meses de muestreo basado en las variables de calidad del agua potable del sistema de distribución.

Análisis de componentes principales

El ACP se ejecutó en este estudio para ocho variables, de los ocho sitios, en los nueve meses de seguimiento de la calidad del agua. Las variables analizadas fueron: Temp, Cond, pH, CL, recuento de HAT y MAT, TR y Dist.

Se realizó el ACP con el criterio de clasificación por sitio con el fin de conocer las variables que tuvieron un mayor impacto sobre cada uno de los puntos muestreados. El ACP mostró tres CPs con auto valores superiores a 1 (Tabla 1) que explicaron el 80,0 % de la varianza total. El primer CP explicó el 34,0 % de la varianza total y estuvo relacionado con una variable microbiológica como la HAT. El CP2 estuvo dominado por una variable fisicoquímica (Cond) y representó el 29,0 % de la varianza total. El CP3 explicó el 17,0 % de la varianza total y estuvo mejor representado por una característica ambiental como la Temp.

Tabla 1. Análisis de componentes principales del conjunto de datos estandarizados sobre la calidad del agua para los ocho sitios de muestreo. Las cargas factoriales en negrita son moderadamente significativas (> 0,50).

Variables	Componentes principales		
	CP1	CP2	CP3
Temp	-0,10	-0,08	**0,71**
Cond	-0,11	**0,63**	0,04
pH	0,37	0,21	0,46
CL	-0,20	-0,45	0,34
HAT	**0,53**	-0,16	0,27
MAT	0,41	0,44	0,03
TR	0,43	-0,21	-0,44
Dist	0,41	-0,29	0,34
Autovalores	2,74	2,29	1,34
Variación explicada	0,34	0,29	0,17
Variación acumulada	0,34	0,63	0,80

La representación gráfica de las observaciones y variables indicaron asociaciones entre los distintos sitios (Figura 6). P1 y P2 son los puntos de almacenamiento, tuvieron Temp más altas, menores TR y menores concentraciones de HAT. P6 y P7 son puntos que se encuentran a la misma distancia (lejana) con respecto a P1 y P2, presentaron mayores concentraciones de CL y menor MAT y Cond. El punto P5 se encuentra a una distancia media y presentó un mayor tiempo de residencia y concentración de HAT. El grupo conformado por dos puntos cercanos (P3 y P4) y uno alejado (P8), estuvo fuertemente asociado a valores altos de pH y menor concentración de MAT.

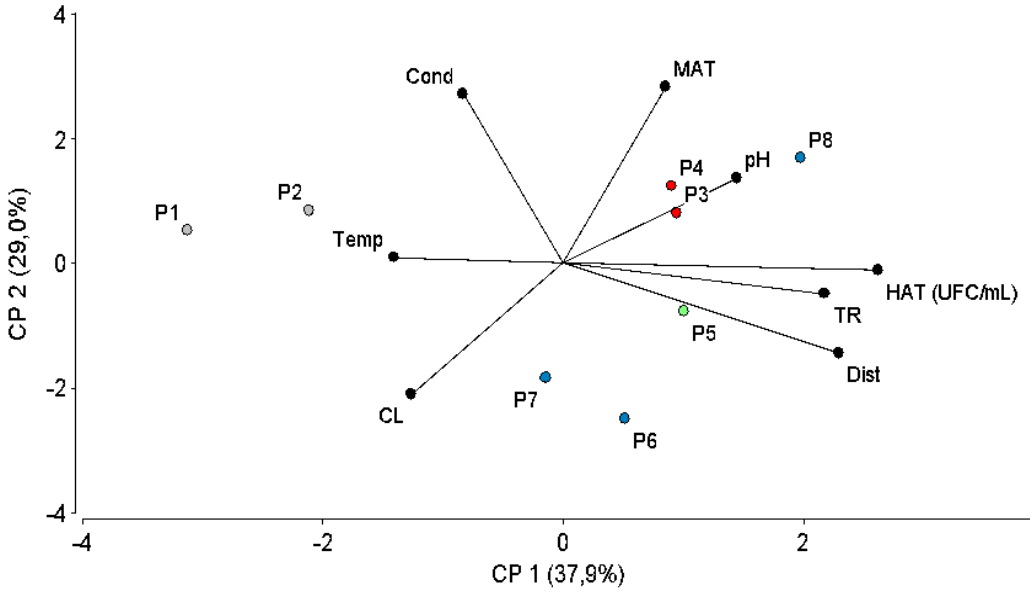


Figura 7. Biplot del análisis de componentes principales en base a las dos primeras componentes para los ocho sitios de muestreo. El color de los sitios (P1-P8) indican si se encuentra a una distancia cercana (rojo), intermedia (verde) o lejana (celeste) de los tanques de almacenamiento (gris).

Este análisis mostró que las muestras de los ocho sitios estuvieron más asociadas entre sí, según su calidad del agua, si se encontraban a la misma distancia del tanque y la cisterna. Sólo un sitio (P8) que es el sitio más alejado de P1 y P2, fue ubicado próximo a P3 y P4, sitios que se encontraron más cercanos a los tanques de almacenamiento.

CONCLUSIONES

- El monitoreo mensual realizado, indicó que en general la calidad fisicoquímica del agua fue aceptable según la normativa vigente. Se observó un incremento en los valores de pH en algunos sitios y una gran variabilidad en las concentraciones de desinfectante a lo largo de los meses.
- Las características microbiológicas del agua potable durante el período de cuarentena permitieron calificar al agua como apta para el consumo humano. Los recuentos de colonias en el medio R2A fueron superiores a los determinados en el medio APC y no se observó la presencia de microorganismos indicadores y/o patógenos oportunistas en ninguno de los puntos muestreados.
- El análisis multivariado demostró ser una herramienta útil para definir puntos claves en la vigilancia espacio-temporal de la calidad del agua potable en un sistema de distribución de agua.
- Es necesario que en el esquema de monitoreo se incluya por lo menos un sitio a diferentes distancias de las fuentes de agua (almacenamiento) y un momento de las distintas temperaturas del año (estaciones). Sin embargo, en el seguimiento temporal de las características del agua potable, es necesario tener en cuenta además la tasa de consumo del agua potable de los edificios de la red, ya que esta variable puede afectar notablemente parámetros físicos y microbiológicos.

Estas consideraciones en el diseño de monitoreo facilitan la evaluación continua de la estabilidad biológica deseada a lo largo de todo el sistema de distribución, ayudando así a garantizar el agua potable segura y de alta calidad para el consumidor.

BIBLIOGRAFÍA

Behmel, S., Damour, M., Ludwig, R., & Rodriguez, M. J. (2016). Water quality monitoring strategies—A review and future perspectives. Science of the Total Environment, 571, 1312-1329.

Bu, H., Tan, X., Li, S., & Zhang, Q. (2010). Water quality assessment of the Jinshui River (China) using multivariate statistical techniques. Environmental Earth Sciences, 60(8), 1631-1639.

CAA, Capitulo XII. BEBIDAS HÍDRICAS, AGUA Y AGUA GASIFICADA. Código Alimentario Argentino. 2019.

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W., (2018). InfoStat versión 2018. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Dong, J., Wang, G., Yan, H., Xu, J., & Zhang, X. (2015). A survey of smart water quality monitoring system. Environmental Science and Pollution Research, 22(7), 4893-4906.

Eaton, A. D., Clesceri, L. S., Rice, E. W., Greenberg, A. E., & Franson, M. A. H. (2005). Standard Methods for the Examination of Water and Wastewater 21st Edition. Washington DC: American Public Health Association.

Howladar, M. F., Al Numanbakth, M. A., & Faruque, M. O. (2018). An application of Water Quality Index (WQI) and multivariate statistics to evaluate the water quality around Maddhapara Granite Mining Industrial Area, Dinajpur, Bangladesh. Environmental Systems Research, 6(1), 1-18.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Ley N° 26221. Boletín Oficial de Argentina, Buenos Aires, Argentina, 02 de marzo de 2007.

Oketola, A. A., Adekolurejo, S. M., & Osibanjo, O. (2013). Water quality assessment of River Ogun using multivariate statistical techniques.

Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Science of the total environment, 313(1-3), 77-89.

OMS, Guía para la calidad del agua. 4ta ed. 2017.

Organización Panamericana de la Salud, Guía rápida para la vigilancia sanitaria del agua. 2013.

Prest, E. I., Hammes, F., van Loosdrecht, M., & Vrouwenvelder, J. S. (2016). Biological stability of drinking water: controlling factors, methods, and challenges. Frontiers in microbiology, 7, 45.

Proctor, C. R., & Hammes, F. (2015). Drinking water microbiology—from measurement to management. Current Opinion in Biotechnology, 33, 87-94.

Strobl, R. O., & Robillard, P. D. (2008). Network design for water quality monitoring of surface freshwaters: A review. Journal of environmental management, 87(4), 639-648.

AGRADECIMIENTOS

Agradecemos a la Dra. Mónica Aparicio González de la Universidad Nacional de Salta, por su participación en el procesamiento microbiológico de las muestras.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Aplicación de la regresión lineal múltiple para el análisis multivariante de parámetros operativos en los procesos de obtención de carbonato de litio en el NOA.

Thames Cantolla, Martin; Valdez, Silvana K.; Orce Schwarz, Agustina

Facultad de Ingeniería, INBEMI; Universidad Nacional de Salta, Salta

core.mtc@hotmail.com, +54 9 387 466 9838

RESUMEN

En el noroeste argentino, la producción de carbonato de litio tiene lugar a partir de salmueras con concentraciones económicas de litio. El método de producción que aplican las empresas mineras depende en gran medida de la composición química de la salmuera que utilizan como materia prima. La mayoría de los métodos involucran etapas de concentración, purificación y precipitación química de sales. La composición química de las salmueras difiere entre salares e incluso entre salmueras pertenecientes a un mismo salar. El objetivo de este trabajo es analizar cómo influye la composición química de las salmueras junto a otras variables que intervienen en la producción de carbonato de litio y subproductos, como ser: grado de evaporación, temperaturas de entrada y salida de salmueras y humedad de filtrados. Para ello se realizaron simulaciones en Aspen Plus v11 de 3 procesos productivos que aplican empresas mineras empleando como materia prima 4 salmueras pertenecientes a diferentes salares de la región. Se utilizó la regresión lineal múltiple para encontrar una vinculación entre estas variables y las cantidades de producto y subproductos que podrían obtenerse. Los resultados obtenidos pueden ser empleados por las nuevas empresas mineras productoras de litio, para la toma de decisiones estratégicas.

Palabras Clave: litio, análisis multivariante, procesos, regresión lineal, NOA.

INTRODUCCIÓN

Carbonato de litio en el NOA

En los últimos años, el litio ha adquirido una enorme relevancia y su consumo se ha incrementado significativamente debido a su uso en baterías recargables en dispositivos de comunicación móvil y principalmente en vehículos eléctricos en donde la industria automotriz es el principal demandante con casi un 40% del mercado, demanda que, según proyecciones, alcanzaría a más de dos tercios de la demanda mundial para el año 2025 [1,2].
Desde la visión productiva, los salares sudamericanos enriquecidos en litio presentan una gran disponibilidad de recursos y costos operativos competitivos [1,2]. En Sudamérica la cuarta reserva mundial de litio se localiza en Argentina (país integrante del "triángulo del litio" junto a Bolivia y Chile) representando un desafío local y regional, superar un esquema de extracción minera de litio de alto valor tecnológico [3-5].
Según datos de la Secretaría de Minería de Salta, en el país existen alrededor de 50 proyectos de extracción de litio en desarrollo, y 12 de ellos están ubicados en nuestra provincia. De estos últimos, 2 se encuentran en fase avanzada para estar en producción para el 2022. Es por esto que el gobierno nacional, tiene altas expectativas puestas en Salta, y espera que el país se convierta en el segundo productor mayoritario de litio a nivel mundial en el año 2022 [5].
Una vez establecida la factibilidad operativa del salar, se procede al desarrollo o selección del proceso productivo para obtener el producto deseado, en este caso, carbonato de litio (Li_2CO_3) [4-6]. El desarrollo o selección del proceso productivo se encuentra íntimamente vinculado con la calidad de la salmuera a tratar (concentración de iones presentes). En un proceso típico de obtención de carbonato de litio, se encuentran presentes las etapas de: evaporación, purificación, encalado, tratamiento con resinas, entre otros [7,8].
Encontrar una vinculación entre las variables de entrada y salida puede resultar beneficioso para la empresa minera, ya que permitiría encontrar el funcionamiento óptimo del proceso, obteniendo de esta manera un alto rendimiento sin la necesidad de incurrir en un gasto innecesario de recursos [8-10].
Sin embargo, la elección del método de extracción puede ser una de las actividades más críticas y problemáticas de la minería. El objetivo de elegir un método de procesamiento es maximizar las ganancias de la empresa y la recuperación de recursos minerales, y proporcionar un entorno seguro al elegir el método correcto con la menor cantidad de problemas entre las alternativas factibles [9-11]. Esta elección es una tarea compleja y es necesario considerar muchos factores, como la calidad de la materia prima, el clima, la tecnología disponible, etc.
Particularmente, los salares en donde se encuentran concentraciones económicas de litio para su explotación, presentan propiedades y características diversas entre sí. Aspecto que es válido incluso en salares que pertenecen a una misma zona geográfica [10,11,13].
Aunque la experiencia y el juicio de la ingeniería aún brindan información importante para elegir los métodos de extracción, generalmente solo a través de un análisis detallado de los datos disponibles se pueden distinguir los matices de cada depósito. La elección del mejor método productivo, debe garantizar que todos los factores se consideren con su correspondiente nivel de importancia, sin embargo, debe tenerse en cuenta que en algunas ocasiones no se cuenta con la persona o el grupo adecuado de especialistas para seleccionar el mejor método [8,10,13].
El objetivo de este trabajo es analizar cómo influye la composición química de las salmueras junto a otras variables que intervienen en la producción de carbonato de litio y subproductos, como ser: grado de evaporación, temperaturas de entrada y salida de salmueras y humedad de filtrados. Para ello se realizaron simulaciones en Aspen Plus v11 de 3 procesos productivos que aplican empresas mineras empleando como materia prima 4 salmueras pertenecientes a diferentes salares de la región. Se utilizó la regresión lineal múltiple para encontrar una vinculación entre estas variables y las cantidades de producto y subproductos que podrían obtenerse. Para ello se realizaron simulaciones en Aspen Plus v11 de 3 procesos productivos que aplican empresas mineras empleando como materia prima 4 salmueras pertenecientes a diferentes salares de la región. Se utilizó la regresión lineal múltiple para encontrar una vinculación entre estas variables y las cantidades de producto y subproductos que podrían

obtenerse. Este modelo representa una optimización de un modelo presentado anteriormente, en el que se aplicó la regresión lineal simple. Los resultados obtenidos pueden ser empleados por las nuevas empresas mineras productoras de litio, para la toma de decisiones estratégicas.

Aspen Plus.

El software llamado Aspen (Sistema Avanzado para Ingeniería de Procesos) es utilizado en diferentes industrias para la simulación de procesos químicos mediante su representación con diagramas de flujo [8, 9]. Este software originado en 1970 puede ser empleado en casi todas las áreas de la ingeniería de procesos, desde la etapa del diseño hasta el análisis de costos y rentabilidad. Cuenta con una biblioteca de elementos y compuestos químicos, como así también de equipos típicos de la industria como ser: columnas de destilación, intercambiadores de calor, separadores y reactores, entre otros [14-17].

Al utilizar un simulador de procesos, resulta posible:

- Predecir el comportamiento de un proceso químico.
- Analizar diferentes escenarios al modificar variables.
- Optimizar un proceso, ya sea por etapa o en su totalidad.
- Implementar mejoras o agregar etapas a un proceso.

De forma general se puede indicar que los pasos a seguir para realizar una simulación en Aspen Plus son los siguientes [14-19]:

1) Definir el flowsheet del proceso, con sus unidades de operación, corrientes de entrada y salida.
2) Establecer los componentes químicos en el proceso.
3) Seleccionar los modelos termodinámicos (presentes en el banco de datos de Aspen) para representar las propiedades físicas.
4) Definir los caudales de las corrientes.
5) Definir las condiciones de operación de los equipos.
6) Vincular los equipos mediante las corrientes de material.
7) Verificar que todos los datos requeridos por el simulador para ejecutar la simulación se encuentren cargados correctamente y en las unidades correspondientes.
8) Ejecutar la simulación.

Finalmente, entre las principales funciones que podemos aprovechar de este simulador, se encuentran [17-22]:

- Generación de gráficos y tablas.
- Estudio de casos.
- Dimensionado y evaluación económica de equipos.
- Estimación de propiedades químicas y termodinámicas.
- Optimización de procesos.
- Ajustes de datos experimentales.
- Determinación de consumos energéticos.
- Análisis de curvas de funcionamiento.

En la Figura 1 se presenta una vista de la pantalla de carga de datos del simulador.

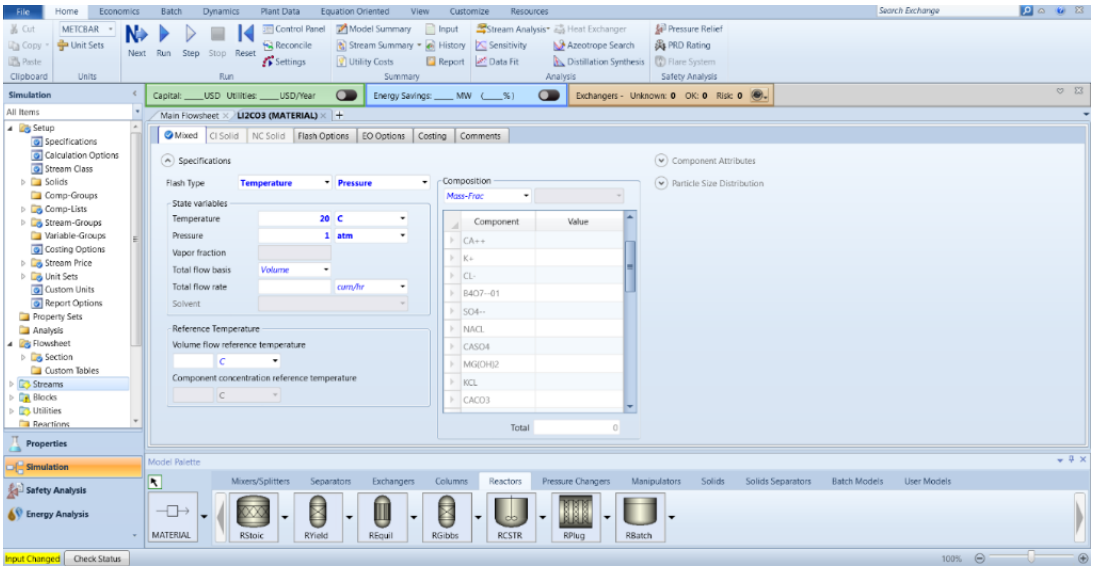


Figura 1. Pantalla de carga de datos en Aspen Plus v11

Regresión lineal múltiple

El método de regresión lineal es una práctica estadística utilizada en diversas áreas como la economía, política, psicología, entre otros. Esta metodología permite analizar la relación entre un conjunto de variables.
El análisis de regresión lineal múltiple permite establecer la relación entre una variable dependiente *Y,* y un conjunto de variables independientes (X_1, X_2, ... X_K). El análisis de regresión lineal múltiple, a diferencia del simple, se acerca más a la situación de análisis real, porque los fenómenos y procesos químicos son generalmente complejos, por lo que deben ser explicados en la mayor medida posible por una serie de variables que directa o indirectamente se encuentran involucradas. El objetivo de estos modelos es tratar de explicar la relación que existe entre la variable dependiente y el conjunto de variables independientes [23]-[25].
Mediante la aplicación de esta herramienta se puede predecir el comportamiento de la variable de respuesta (dependiente) partir de los valores conocidos de las variables explicativas (independientes). La Ecuación 1 es la que describe a la regresión lineal múltiple [29,30].

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (1)$$

Donde:
Y = variable de respuesta.
X_n = variables explicativas.
b_n = coeficientes de las variables explicativas.
b_0 = constante de la regresión.

En la regresión múltiple las pruebas y análisis que se aplican para determinar la relación y el grado de asociación o vinculación entre variables siguen la misma metodología que la aplicada en la regresión lineal simple. La diferencia radica que en la regresión lineal múltiple las variables del modelo son seleccionadas una a una. Esto es así ya que el objeto es encontrar aquellas variables que mejor expliquen a la variable dependiente.

De esta manera, cada variable independiente es ponderada, indicando su contribución relativa a la predicción conjunta. El procedimiento del análisis de regresión asegura una predicción máxima a partir del conjunto de variables independientes [25]-[27].

El procedimiento resumido a seguir sería:

1) se inicia el modelo sin ninguna variable explicativa
2) se agregan de a una cada variable explicativa, verificando que sean significativas.

3) Se verifica la bondad del ajuste del modelo
4) Se repite el paso 2 y 3 con todas las variables seleccionadas
5) El modelo final es aquel que contiene sólo a aquellas variables significativas que permiten una alta bondad de ajuste.

Debe tenerse en cuenta que el análisis de regresión lineal debe utilizarse sólo cuando las variables dependientes e independientes son métricas. No obstante, ante eventuales condiciones de operación, se podrá utilizar variables no métricas a partir de la transformación de los datos en variables ficticias [24]-[27].
Es importante destacar que, el éxito de la regresión lineal múltiple radica en la correcta asignación de las variables dependiente e independientes [23]-[25].

METODOLOGÍA

Para poder determinar cómo influye la composición inicial de un salar en la selección del método productivo para producir carbonato de litio, se realizaron simulaciones en Aspen Plus v11, empleando 3 procesos diferentes de producción que actualmente emplean empresas mineras de la zona y a su vez, utilizando como materia prima 4 salmueras de la región. De acuerdo a esto, se siguieron 6 etapas que se describen a continuación:

Selección de los procesos a analizar

En primer lugar, se determinaron los procesos a estudiar. Para lograr una aplicación práctica, se seleccionaron 3 procesos que son utilizados por empresas litieras del medio. Se los nombraron como Proceso 1, Proceso 2 y Proceso 3. Cada uno de ellos presenta diferencias en el modo de recuperar litio de la salmuera. El Proceso 1 consiste en una concentración por evaporación y purificación por reactivos. Los Procesos 2 y 3, consisten en una purificación mediante adsorción con resinas y la obtención intermedia de subproductos. Al ser confidenciales, impiden a los autores brindar mayores detalles sobre ellos.

Selección de materia prima

Las salmueras presentan distintas composiciones incluso en pozos de un mismo salar, por ello se consideró necesario no limitar la simulación a una única salmuera sino utilizar 4 distintas provenientes de diferentes salares de la región. A las mismas se las nombró como Salmuera A, B, C y D.

Tabla 1. Composiciones de las 4 salmueras empleadas

Composición inicial (fracción másica)	**Salmuera A**	**Salmuera B**	**Salmuera C**	**Salmuera D**
$[Li^{+}]$	0,00069	0,00066	0,00094	0,00018
$[Mg^{++}]$	0,00627	0,00152	0,00265	0,00398
$[SO_4^{--}]$	0,00672	0,00845	0,00200	0,00034
$[Na^{+}]$	0,06348	0,09186	0,08930	0,07340
$[K^{+}]$	0,00656	0,00682	0,00568	0,00117

En la Tabla 1 se presentan las concentraciones, expresadas en fracción másica, de algunos iones presentes en las salmueras utilizadas para este trabajo. Estos iones a su vez, se encuentran presentes entre los productos y subproductos que son posible obtener con los diferentes procesos estudiados. Cabe destacar que la Salmuera D, presenta la menor concentración de litio con respecto al resto.
Toda la información de las salmueras, y de las condiciones operativas de los procesos, fue provista por el banco de datos del Instituto de Beneficios de Minerales (INBEMI) de la Universidad Nacional de Salta, Facultad de Ingeniería.

Selección de variables

Se establecieron las variables a analizar. En esta etapa se consideró realizar dos modelos por separado a efectos de poder estudiar los procesos productivos con mayor detalle; el primer modelo que permita determinar la cantidad de producto final (Li_2CO_3) a obtenerse y el segundo modelo que permita determinar la cantidad de subproductos que pueden obtenerse, ambos en función de los parámetros operativos:

Tabla 2. Variables dependientes e independiente del modelo de regresión

Variable dependiente 1	Variable dependiente 2	Variables independientes
✓ Cantidad de producto final (Li_2CO_3), en toneladas/hora.	✓ Generación de subproductos, en toneladas/hora	✓ evaporación (grado de concentración) ✓ temperaturas de entrada y salida de la salmuera. ✓ nivel de humedad de filtrados.

Primera simulación en Aspen

Una vez determinados los procesos a analizar, seleccionadas las salmueras a estudiar, y fijadas las variables dependientes e independientes que se desean estudiar, se procedió a realizar la simulación en Aspen Plus [14-22]. Para los 3 procesos se cargaron:

- ✓ Las unidades de medida.
- ✓ La composición química inicial de la salmuera a estudiar.
- ✓ Los flujos de materia prima.
- ✓ Equipos intervinientes por etapa y su eficiencia.
- ✓ Las corrientes de entrada y salida por equipo.
- ✓ Temperaturas, presiones y humedad de los filtrados.

Estas primeras simulaciones se ejecutaron para cada uno de los procesos y cada una de las salmueras, obteniéndose las tablas de resultados iniciales en donde se presentaban los resultados operativos para las condiciones iniciales cargadas. En esta etapa se determinó a priori las cantidades de producto y subproductos que podrían obtenerse.

Simulación iterativa

Se ejecutó la simulación en Aspen Plus de forma iterativa, asignando distintos valores al nivel de evaporación (concentración) de la salmuera. Para realizar esto de forma automática, se empleó la función “Sensitivity Analysis” que trae incorporada el simulador. Esta función permite que se ejecute la simulación, para cada uno de los valores que el usuario establece para una o un conjunto de variables de entrada [14-22]. Cabe destacar que, en este punto, se estableció un intervalo de valores de evaporación de [0%, 60%] ya que dichos valores contemplan los valores mínimo y máximo, aplicados por las empresas mineras.
Mediante esta función, se ejecutaron en promedio 10.000 iteraciones para cada uno de los procesos y cada una de las salmueras.

Análisis de datos y modelización

Una vez realizadas las simulaciones, se volcaron los datos a una planilla de cálculo de Excel, en donde fue posible ordenarlos y analizarlos a partir de las herramientas propias del software entre las cuales es posible utilizar la regresión lineal múltiple.

RESULTADOS

Al aplicar la regresión lineal múltiple a cada uno de los procesos y salmueras, fue posible determinar la vinculación de las distintas variables explicativas y las de respuesta.
De esta forma, se pudo establecer 2 tipos de relaciones entre las variables dependientes y las independientes para cada uno de los procesos. Tal como se esperaba, la composición de las salmueras influye en el grado de vinculación entre variables, por lo que se consideró adecuado presentar las ecuaciones de regresión lineal por cada proceso y salmuera por separado.
En total se han generado 24 regresiones lineales múltiples, sin embargo, por cuestiones de espacio, solo se presentarán algunos de los resultados obtenidos. Los mismos se pueden observar en la Tabla 3 y Tabla 4 en donde se presentan los datos de regresión para la vinculación de la cantidad de producto final en función de los parámetros operativos de la Salmuera C con el Proceso 3 y la vinculación de la cantidad de subproductos generados en función de parámetros operativos de la Salmuera D con el Proceso 1, respectivamente.

Tabla 3. Modelo de regresión: Cantidad de producto final en función de parámetros operativos

Tipo de vinculación	Cantidad de producto final en función de parámetros operativos	
Tipo de Proceso	Proceso 3	
Tipo de salmuera	Salmuera C	
Coeficiente de determinación R^2		0,7796
Error típico		0,0361
Variable de simulación	**Parámetro en la regresión**	**Coeficientes**
-	Intercepción	$3{,}78E^{+00}$
Evap1	Variable X_1	$-3{,}41E^{-01}$
Evap2	Variable X_2	$-1{,}18E^{-04}$
TempSalmIN	Variable X_3	$8{,}50E^{-08}$
HFiltro1	Variable X_4	$-5{,}28E^{-01}$
HFiltro2	Variable X_5	$-8{,}22E^{-02}$
TempSalmOUT	Variable X_6	$2{,}14E^{-07}$

A partir de esto, es posible escribir el modelo de regresión lineal para las toneladas de Li_2CO_3, tal como se observa en la Ecuación 2:

$$t\ Li_2CO_3 = 3.78E^{+00} - 3{,}41E^{-01}Evap1 - 1{,}15E^{-04}Evap2 + 8{,}50E^{-08}TempSalmIn - 5{,}28E^{-01}Hfiltro1 - 8{,}22E^{-02}Hfiltro2 + 2{,}14E^{-07}TempSalmOut \quad (2)$$

Cabe destacar que los modelos de regresión correspondientes a la cantidad de producto final en función de parámetros operativos de los otros procesos y salmueras presentaron valores similares para cada variable independiente.

Tabla 4. Modelo de regresión: Cantidad de subproductos generados en función de parámetros operativos.

Tipo de vinculación	Cantidad de subproductos generados en función de parámetros operativos.	
Tipo de Proceso	Proceso 1	
Tipo de salmuera	Salmuera D	
Coeficiente de determinación R^2		0,9756
Error típico		3,5355
Variable de simulación	**Parámetro en la regresión**	**Coeficientes**

-	Intercepción	$2{,}17E^{+01}$
Evap1	Variable X_1	$1{,}30E^{+02}$
Evap2	Variable X_2	$1{,}30E^{+02}$
HFiltro1	Variable X_3	$-4{,}49E^{+00}$
HFiltro2	Variable X_4	$-2{,}46E^{-02}$
HFiltro3	Variable X_5	$-2{,}47E^{-02}$
TempSalmIN	Variable X_6	$-5{,}31E^{-04}$
TempSalmOUT	Variable X_7	$2{,}02E^{-04}$

A partir de esto, es posible escribir el modelo de regresión lineal para las toneladas de Subproductos, tal como se observa en la Ecuación 3:

$$t\ Subproductos = 2{,}17E^{+01} + 1{,}30E^{+02}Evap1 + 1{,}30E^{+02}Evap2 - 4{,}49E^{+00}Hfiltro1 - 2{,}46E^{-02}Hfiltro2 - 2{,}47E^{-02}Hfiltro3 - 5{,}31E^{-04}TempSalmIn + 2{,}02E^{-04}TempSalmOut \quad (3)$$

Al igual que en el caso anterior, se destaca que los modelos de regresión correspondientes a la cantidad de subproductos generados en función de parámetros operativos de los otros procesos y salmueras presentaron valores similares para cada variable independiente.

En todos los casos, las variables consideradas resultaron estadísticamente significativas (valor p <0,05) lo que permite pensar que es posible pronosticar el valor de las variables de respuesta mediante las variables explicativas seleccionadas para estos modelos estadísticos.

CONCLUSIONES

Se pudo modelar exitosamente los 3 procesos empleando 4 salmueras diferentes mediante el simulador Aspen Plus. Con ello se pudo determinar el nivel de rendimiento de cada una de ellas, tanto para productos como subproductos.

Al plantearse modelos individuales tanto para la cantidad de carbonato de litio a producir como de la cantidad de subproductos a generarse, permiten que se pueda tener una visión más amplia al momento de tomar decisiones estratégicas. De los modelos obtenidos, se observa que en el caso la determinación de la cantidad de Li_2CO_3, todas las variables intervinientes presentan un nivel de significancia uniforme; por el contrario, en los modelos para determinar la cantidad de subproductos generados, es destacable mencionar que se ha observado un gran impacto del grado de evaporación (concentración) que se alcanza.

Si bien no es recomendado generalizar estos modelos para todas las salmueras y procesos productivos, la sencillez matemática para la regresión lineal múltiple, hacen de esta metodología una herramienta que puede aplicarse a sistemas complejos como las salmueras y simular procesos de obtención de carbonato de litio, con una buena aproximación e incluso poder interpretar los resultados fácilmente.

La incorporación de múltiples variables independientes, hacen de estos modelos matemáticos una herramienta que permita tomar decisiones contemplando la mayor cantidad de variaciones posibles en el sistema.

Cabe destacar que es posible incorporar más variables independientes a los modelos, sin embargo, el nivel de iteraciones que requerirá el simulador Aspen Plus, aumentará de manera exponencial, traduciéndose en mayor tiempo para la ejecución de las simulaciones.

A futuro se plantea la incorporación de otras variables tanto propias de los procesos productivos como así también de aquellas externas al sistema, permitiendo así alcanzar un análisis global con una interpretación óptima.

BIBLIOGRAFIA

[1]. Castello, A., Kloster, M. (2015). Industrialización del Litio y Agregado de Valor Local: *Informe Tecno-Productivo*. CIECTI, CABA.

[2]. Investiga, Ciencia y Tecnología UNLP. (2019). Litio: un tesoro escondido en la Puna Argentina. *Informe especial*. Disponible en: https://investiga.unlp.edu.ar/especiales/litio-17104 [Accedido el 20/09/2021]

[3]. Diario El Tribuno (2019). El litio salteño llevará al país al segundo lugar de producción en el mundo. Recuperado de: https://www.eltribuno.com/salta/nota/2019-2-5-0-0-0-el-litio-salteno-llevara-al-pais-al-segundo-lugar-de-produccion-en-el-mundo [Accedido el 15/08/2020].

[4]. Comercio y Justicia. (2019). Argentina será el segundo mayor productor global de litio en 2022. Recuperado de: https://comercioyjusticia.info/economia/argentina-sera-el-segundo-mayor-productor-global-de-litio-en-2022/ [Accedido el 28/10/2021].

[5]. Manrique, A. (2014). Explotación del litio, producción y comercialización de baterías de litio en Argentina. *Universidad Nacional de Mar del Plata, Facultad de Ingeniería. E-Book*, ISBN 978-987-544-641-0

[6]. Universidad Nacional de la Plata. (2020). El litio en la Argentina: visiones y aportes multidisciplinarios desde la UNLP. Buenos Aires.

[7]. Calvo, E.J. (2019). Litio, un recurso estratégico para el mundo actual. *Instituto de Química Física de los Materiales, Medioambiente y Energía (INQUIMAE)*, UBA-Conicet. 28, 17-23.

[8]. Visintin, A. (2021). Avances actuales y perspectivas de futuro en torno a las tecnologías de litio en Argentina. *In-Genium*, 1, 103-111.

[9]. Ortiz Sánchez, O. (2018). Modelo analítico para evaluar un yacimiento mineral aplicación en un proyecto de minado superficial. *Revista del Instituto de Investigación de la Facultad de Geología, Minas, Metalurgia y Ciencias Geográficas*, 21 (42), 27-34.

[10]. Revista Panorama Minero. (2018). Litio y tecnología: Cómo obtener más valor de las salmueras. Recuperado de: https://panorama-minero.com/litio/litio-y-tecnologia-como-obtener-mas-valor-de-las-salmueras/ [Accedido el 19/04/2021].

[11]. Flexer, V., Baspineiro, C.F. & Galli, C.I. (2018). Lithium recovery from brines: A vital raw material for green energies with a potential environmental impact in its mining and processing. *Science of The Total Environment*, 639, 1188-1204, ISSN 0048-9697.

[12]. Bravo, V. (2019). Algo sobre el litio. *Documento de trabajo*. Fundación Bariloche. Departamento de Economía Energética.

[13]. Ruiz Peyré, F., Dorn, F. (2020). Aprovechamiento del litio en la Argentina – Realidades, desafíos y perspectivas en un mundo globalizado. *Scripta Nova*. Revista Electrónica de Geografía y Ciencias Sociales, 24, 632.

[14]. Espínola Lozano, F. (2017). Tutorial de Aspen Plus, Introducción y modelos simples de operaciones unitarias. Universidad de Jaén.

[15]. Adornado, A., Soriano, A. & Bungay, V. (2017). Assessment of Aqueous Lithium-based Salt Solutions as Working Fluid for Absorption Chillers using Aspen Plus. *Asean Jorunal of Chemical Engineering*, 17 (2), 51-59

[16]. Chu, J. (2005). Simulation of industrial catalytic reforming process by developing user's module on ASPEN PLUS platform. *Journal of Chemical Industry and Engineering*, 56. 1714-1720.

[17]. Schefflan, R. (2011). Teach Yourself the Basics of Aspen Plus. *Ed. John Wiley & Sons*, Inc.

[18]. Sandler, S. (2015). *Using Aspen Plus in Thermodynamics Instruction: A Step-by-Step Guide*. 1st Edition. *Ed. John Wiley & Sons*, Inc.

[19]. Adams, T (2018). *Learn Aspen Plus in 24 Hours*. McGraw-Hill Education: New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto.

[20]. Al-Malah, K. (2016). *Aspen Plus: Chemical Engineering Applications*. Ed. John Wiley & Sons, Inc.

[21]. Hamza, A., Wajid, A.K. & Zabid, U. (2021). Modeling of an Integrated System Based on Solar Heat Source, Organic Rankine Cycle and Water/Lithium Bromide Absorption Chiller in Aspen Plus. *Faculty of Mechanical Engineering. GIK Institute of Engineering Sciences and Technology.*

[22]. Mouad Hachhach, H.A., Mounir Hanafi, O.A. & Tarik, C. (2019). Simulation and Sensitivity Analysis of Molybdenum Disulfide Nanoparticle Production Using Aspen Plus. *International Journal of Chemical Engineering.*

[23]. Hair, J.F.; Anderson, R.Jr.; Tatham, R.; Black, W. (1999). *Análisis multivariante*. 5º Edición. Prentice Hall Liberia. Madrid.

[24]. Rodríguez Jaume, M.J., Mora Catalá, R. (2001). Análisis de Regresión Múltiple. En Universitat d´Alacant/Universidad de Alicante, Servicio de Publicaciones (Ed.), *Estadística informática: casos y ejemplos con el SPSS* (pp 3-17).

[25]. Ortiz Sánchez, O. (2012). Modelo de regresión de dos etapas aplicado a un cargador de roca en una operación minera superficial. *Revista del Instituto de Investigación*, 15 (29), 117-124.

[26]. Ortiz Sánchez, O. (2018). Modelo analítico para evaluar un yacimiento mineral aplicación en un proyecto de minado superficial. *Revista del Instituto de Investigación de la Facultad de Geología, Minas, Metalurgia y Ciencias Geográficas*, 21 (42), 27-34.

[27]. Montero Granados, R. (2016). Modelos de regresión lineal múltiple. *Documentos de Trabajo en Economía Aplicada*. Universidad de Granada. España.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis comparativo de modelos de machine learning para la predicción de la presión pulmonar en respiradores artificiales

Ramos, Joaquín Ignacio; González, José Nery; Rodríguez, Mariela; Farfán, José Humberto

Facultad de Ingeniería - UNJu. San Salvador de Jujuy[1]

Datos de contacto: joa193@fi.unju.edu.ar[1]; 39807718@fi.unju.edu.ar[1]; hfarfan@gmail.com[1]; mariela.rodriguez@fi.unju.edu.ar[1]

RESUMEN

Los respiradores artificiales bombean oxígeno a los pulmones de un paciente sedado a través de un tubo en la tráquea. Pero la ventilación mecánica es un procedimiento intensivo para los médicos, una limitación que se mostró de manera prominente durante los primeros días de la pandemia de COVID-19.

Al mismo tiempo, desarrollar nuevos métodos para controlar los ventiladores mecánicos es prohibitivamente caro. Los simuladores de alta calidad podrían reducir esta barrera.

Los simuladores actuales se entrenan como un conjunto, donde cada modelo simula una configuración de pulmón único. Sin embargo, los pulmones y sus atributos forman un espacio continuo, por lo que se debe explorar un abordaje paramétrico que considere las diferencias en los pulmones del paciente.

La Universidad de Princeton junto con el equipo de Google Brain crearon una competencia en la plataforma Kaggle, donde pusieron a disposición de todos los usuarios un conjunto de datos extraídos de un respirador en funcionamiento junto con un pulmón artificial.

En el presente trabajo se realiza un análisis comparativo de tres modelos de inteligencia artificial para la predicción de presión pulmonar medida por el respirador y así lograr alcanzar un mejor control de la ventilación mecánica.

Palabras Clave: respiradores artificiales, ventiladores mecánicos, análisis exploratorio, Data Mining.

INTRODUCCIÓN

Los respiradores usan un ventilador para bombear oxígeno hacia los pulmones de un paciente sedado a través de un tubo en la tráquea. El actual trabajo se centra en la situación abordada mediante la competencia propuesta por Google Brain en donde se procede a realizar simulaciones de un respirador conectado al pulmón de un paciente.

Es sabido que cuando un paciente tiene dificultad para respirar los médicos son sometidos a un proceso intensivo de utilizar respiración mecánica y estas dificultades se hicieron más frecuentes y notorias con la llegada de las primeras cepas de COVID-19 [1]. Es por esto que los costos de desarrollar métodos para controlar los ventiladores mecánicos son probablemente altos.

A partir de la disponibilidad del dataset provisto por el equipo de Google Brain y la Universidad de Princeton se parte de la idea de que en base a sus características, las redes neuronales y el aprendizaje profundo pueden usarse para predecir y generalizar mejor la presión de aire necesaria para mantener al paciente con el correcto flujo de aire en los pulmones. Si se tiene éxito se podrá allanar caminos para algoritmos que se adapten a los respiradores y a los pacientes y que a la vez hagan la tarea un poco más fácil para los médicos.

METODOLOGÍA

Como metodología adoptada para este estudio se emplea la metodología KDD (Descubrimiento de Conocimiento en Base de Datos). KDD está basada en un bien definido proceso KDD de múltiples pasos, para el descubrimiento de conocimiento en grandes colecciones de datos. El proceso KDD es iterativo por naturaleza, y depende de la interacción para la toma de decisiones, de manera dinámica [2] a fin de encontrar y descubrir patrones o reglas que expliquen el comportamiento de los datos de acuerdo a la información registrada en el dataset provisto para la predicción de la presión de los ventiladores. KKD se basa en 5 pasos [3] y [4] que se detallan en el siguiente apartado (Figura 1).

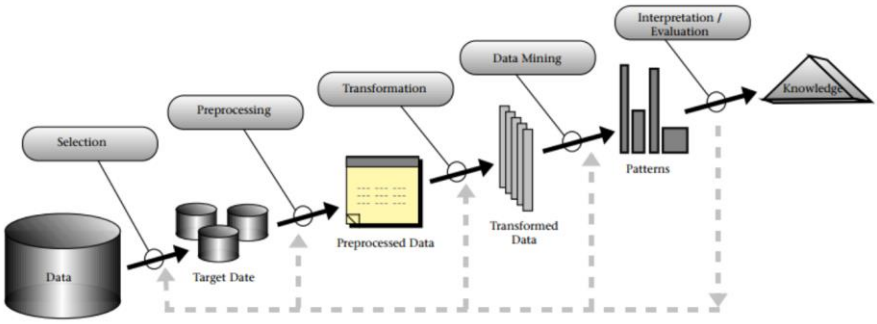


Figura 1. Visión general de los pasos que componen el proceso KDD

DESARROLLO

Objetivo

El objetivo es evaluar modelos para predecir la presión de las vías respiratorias a partir de numerosas series de tiempo de respiraciones y de entradas de control del respirador.

Impacto social de la competencia en kaggle

Si la competencia tiene éxito, ayudará a superar la barrera de los costos de desarrollar nuevos métodos para controlar los ventiladores mecánicos. Esto allanará el camino para algoritmos que se adapten a los pacientes y reduzcan la carga para los médicos durante estos nuevos tiempos y más allá. Como resultado, los tratamientos con ventilador pueden estar más disponibles para ayudar a los pacientes a respirar.

Descripción de los datos

Los datos del respirador utilizados en esta competencia se produjeron utilizando un respirador "open source" modificado conectado a un pulmón de prueba de fuelle artificial a través de un circuito respiratorio.

El respirador "People's Ventilator Project (PVP)" es un respirador de control de presión de código abierto y de bajo costo diseñado para una dependencia mínima de piezas médicas especializadas para adaptarse mejor a la escasez de la cadena de suministro.

El siguiente diagrama ilustra la configuración, con las dos entradas de control resaltadas en verde y la variable de estado (presión de las vías respiratorias) para predecir en azul. La primera entrada de control es una variable continua de 0 a 100 que representa el porcentaje que la válvula solenoide inspiratoria está abierta para dejar entrar aire al pulmón (es decir, 0 está completamente cerrado y no se deja entrar aire y 100 está completamente abierto). La segunda entrada de control es una variable binaria que representa si la válvula exploratoria está abierta (1) o cerrada (0) para dejar salir el aire.

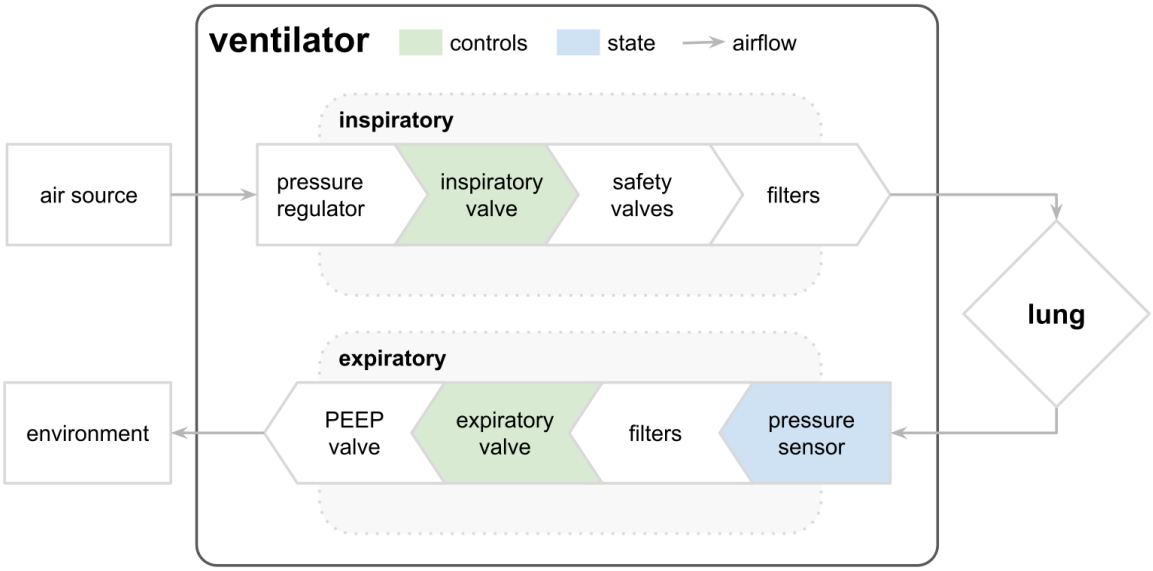


Figura 2. Diagrama de flujo de aire del respirador artificial.

Cada serie de tiempo representa una respiración de aproximadamente 3 segundos. Los archivos están organizados de manera que cada fila es un paso de tiempo en una respiración y da las dos señales de control, la presión resultante en las vías respiratorias y los atributos relevantes del pulmón, que se describen a continuación.

Los datos cuentan con

- Cantidad de registros: 6.035.000
- Columnas: 8

Columna	Descripción
Id	Id del registro
breath_id	id de la respiración actual
R	atributo pulmonar que indica qué tan restringida está la vía aérea (en cmH2O / L / S). Físicamente, este es el cambio de presión por cambio de flujo (volumen de aire por tiempo). Intuitivamente, uno puede imaginarse inflar un globo con una pajita. Podemos cambiar R cambiando el diámetro de la pajilla, siendo R más alto más difícil de soplar.
C	atributo de pulmón que indica qué tan moldeable es el pulmón (en mL / cmH2O). Físicamente, este es el cambio de volumen por cambio de presión. Intuitivamente, uno puede imaginar el mismo ejemplo de globo. Podemos cambiar C cambiando el grosor del látex del globo, con un C más alto que tiene un látex más delgado y más fácil de soplar.
time_step	marca de tiempo de la respiración actual
u_in	la entrada de control para la válvula solenoide inspiratoria. Rango de valores: desde 0 hasta 100.
u_out	la entrada de control para la válvula de solenoide exploratoria. Rango de valores: 0 o 1.
pressure	La presión de las vías respiratorias medida en el circuito respiratorio, medida en cmH2O.

Tabla 1. Descripción de las columnas presentes en el dataset.

Paso 1: Selección

En el paso de selección, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento [5]. Para poder realizar un análisis estadístico o entrenar un modelo será necesario reducir en gran parte el dataset de forma que la información que quede resultante de la selección sea representativa de toda la demás información restante que se deja de lado.

Para el caso del dataset descrito se poseen 6.036.000 registros en total de los cuales se realiza una selección de los primeros 32.000 ya que se tiene en cuenta los siguientes factores:

- Una respiración completa está compuesta por la inhalación y la exhalación, ambos provistos gracias a la simulación de un pulmón artificial por el cual ingresa y sale la corriente de aire.

- El proceso completo de una respiración está integrado en la totalidad de 80 registros consecutivos, empezando desde el primero de todos con el primer step time de la primera inhalación correspondiente a la inhalación número uno.

- Todas las respiraciones se aproximan a una misma forma para el patrón de la presión generado por cada respiración, siendo mucho mayor la presión en la etapa de inhalación para luego bajar abruptamente durante la exhalación

Por lo tanto al seleccionar los primeros 32.000 registros estamos seleccionando un total de 400 respiraciones que se usarán para la extracción del conocimiento y de esta forma evitar problemas de rendimiento y de limitación de hardware en lo que respecta al tiempo de procesamiento de las operaciones que se realizan.

Paso 2: Preprocesamiento / Limpieza

En la etapa de preprocesamiento/limpieza (data cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo [5].

El dataset provisto por parte de la competencia de Kaggle y del equipo de Google Brain está en perfectas condiciones en cuanto a datos desconocidos se refiere ya que no posee ningunos datos missing o empty en ninguno de sus campos, por lo tanto, para limpiar los datos del dataset lo que se realizó fue la aplicación de otros tipos de operadores como son la selección de atributos en donde se seleccionan solamente un subset de atributos y se eliminan los demás los cuales son *id* y *breath_id* ya que los dos no aportan información al valor de la predicción. También se eliminaron aquellos registros que presentan valores de presión negativos para que estos no influyan en el resultado final.

Paso 3: Transformación / Reducción

Análisis exploratorio

Matriz de correlación

Attribut...	id	breath_id	R	C	time_st...	u_in	u_out	pressure
id	1	1.000	0.002	0.007	-0.000	-0.002	-0.000	-0.002
breath_id	1.000	1	0.002	0.007	-0.000	-0.002	-0.000	-0.002
R	0.002	0.002	1	-0.096	-0.015	-0.148	-0.008	0.016
C	0.007	0.007	-0.096	1	0.005	0.151	0.004	-0.037
time_step	-0.000	-0.000	-0.015	0.005	1	-0.352	0.839	-0.525
u_in	-0.002	-0.002	-0.148	0.151	-0.352	1	-0.417	0.308
u_out	-0.000	-0.000	-0.008	0.004	0.839	-0.417	1	-0.615
pressure	-0.002	-0.002	0.016	-0.037	-0.525	0.308	-0.615	1

Figura 3. Matriz de correlación de todos los atributos

Correlación de atributos dependiente “pressure” (label)

attribute	weight ↓
u_out	0.615
time_step	0.525
u_in	0.308
C	0.037
R	0.016
id	0.002
breath_id	0.002

Tabla 2. Correlación de los atributos respecto del label (atributo “pressure”)

Gráfico de la presión a través del tiempo de una respiración

El siguiente gráfico de la Figura 4 representa la presión de una sola respiración (80 registros) a través del tiempo que son aproximadamente 3 segundos, el color azul indica que la válvula de salida del respirador está cerrada y cuando es color verde indica que la válvula de salida del respirador está abierta.

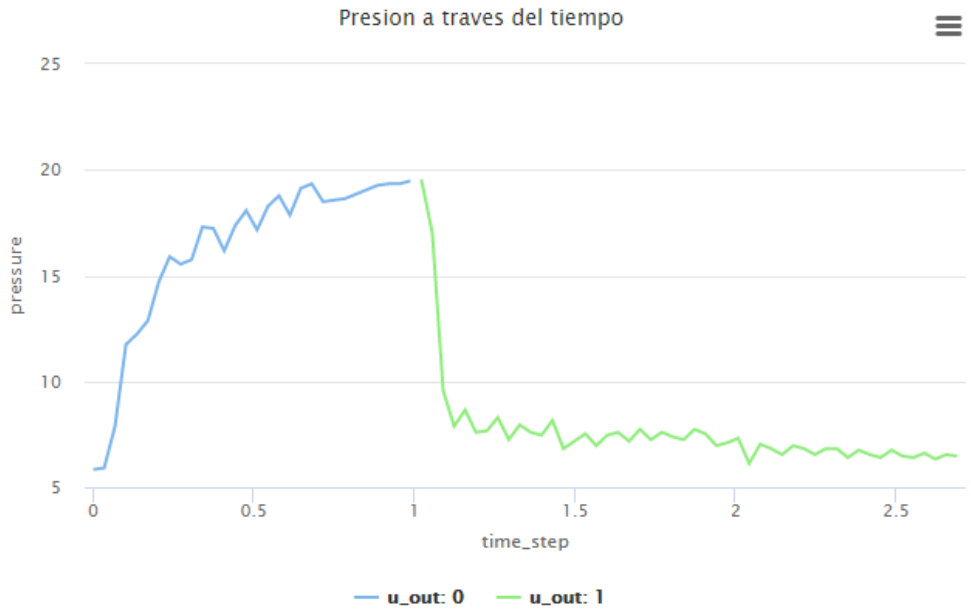


Figura 4. Presión a través del tiempo de una respiración. Coloreado según atributo u_out.

Se observa claramente la relación entre la presión y el estado de la válvula de salida.

La presión del usuario aumenta cuando la válvula de salida está cerrada (valor 0) y la presión del usuario disminuye cuando la válvula está abierta (1), lo que implica una relación inversa entre ambos atributos.

La relación inversa encontrada en la matriz de correlación (Figura 3) explica el valor negativo (-0.615) que aparece en la intersección de los dos atributos (pressure y u_out) en la matriz de correlación presentada anteriormente.

Paso 4: Minería de datos

Métrica de predicción

Se utilizará como métrica de evaluación el error absoluto medio entre las presiones pronosticadas y reales de cada respiración. La puntuación viene dada por:

$$|X - Y|$$

donde X es el vector de la presión predicha e Y es el vector de las presiones reales en todas las respiraciones del equipo de prueba.

Según la consigna de la competencia en Kaggle los mejores resultados tendrían que tomar en cuenta los parámetros del pulmón (R y C) para alcanzar mejores predicciones.

Modelos de predicción

Para intentar predecir la presión de los pulmones del usuario del respirador se decidió probar tres modelos y evaluar su efectividad de predicción.

Los modelos son:

- Regresión lineal
- Árbol de decisión
- Red neuronal

Para evaluar los resultados de cada modelo se utilizará la técnica de validación simple y validación cruzada.

Para la validación simple se realizó una distribución de los datos de la siguiente manera:

- 80% de los datos de entrenamiento
- 20% de datos de prueba.

Por otro lado, la validación cruzada o cross validation lo que realiza es evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones (Figura 5). Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Para nuestro dataset lo que se hizo fue seleccionar 10 folds (pliegues de subsets), por lo tanto el número de iteraciones que se tendrá lugar en el proceso de entrenamiento será el mismo valor que el de pliegues.

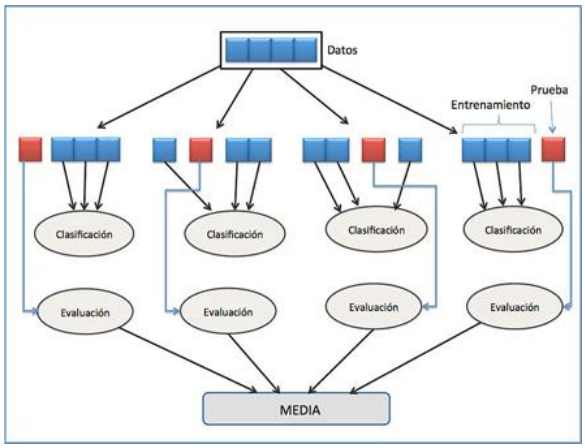


Figura 6: Método de validación cruzada

Modelo de regresión lineal

<u>Con Validación Simple:</u>

Resultados:

root_mean_squared_error

```
root_mean_squared_error: 6.373 +/- 0.000
```

absolute_error

```
absolute_error: 3.985 +/- 4.974
```

Figura 6: Resultado del modelo de regresión lineal con validación simple

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
R	0.003	0.002	0.008	1.000	1.632	0.103
C	-0.032	0.002	-0.068	1.000	-13.737	0
time_step	-0.309	0.096	-0.029	0.308	-3.219	0.001
u_in	0.041	0.003	0.065	0.851	11.915	0
u_out	-9.446	0.156	-0.562	0.298	-60.638	0
(Intercept)	17.939	0.128	?	?	140.320	0

Figura 7: Coeficientes del modelo de regresión lineal

Modelo generado:

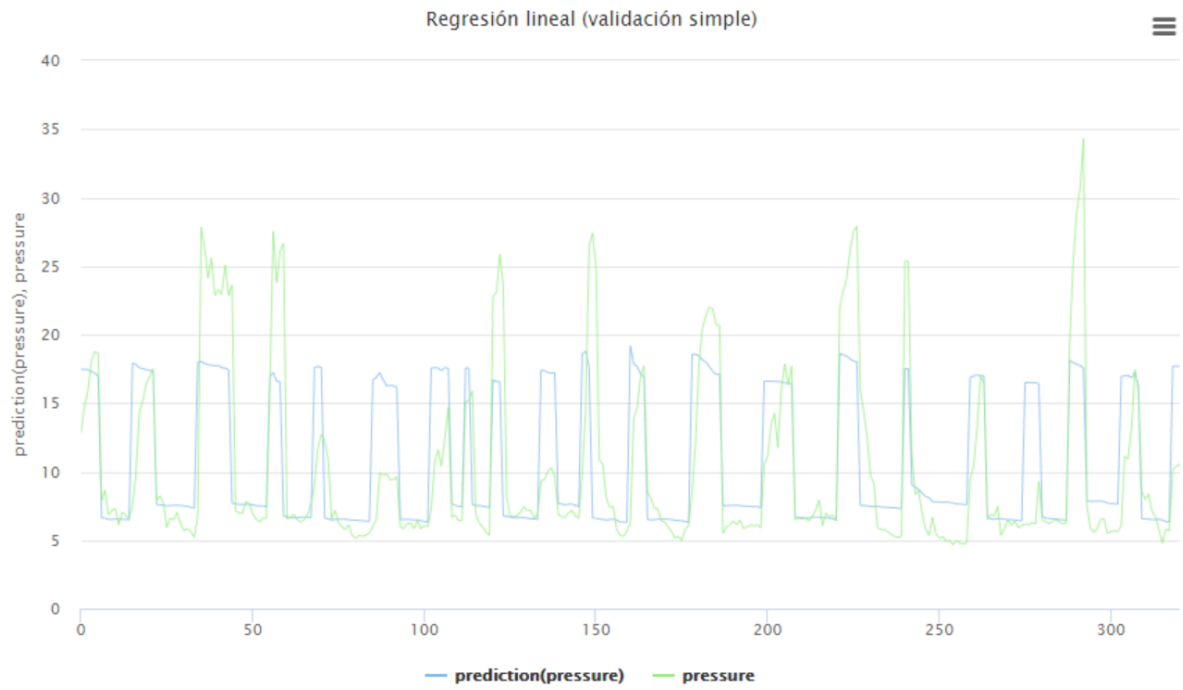


Figura 8: Resultado gráfico de la predicción del modelo de regresión lineal con validación simple

Se observa que el modelo de regresión lineal no llega a predecir los picos más altos de presión en cada respiración.

<u>con Validación Cruzada:</u>

Resultados:

root_mean_squared_error

root_mean_squared_error: 6.408 +/- 0.111 (micro average: 6.409 +/- 0.000)

absolute_error

absolute_error: 4.038 +/- 0.076 (micro average: 4.038 +/- 4.977)

Figura 9: Resultado del modelo de regresión lineal con validación cruzada

Modelo generado:

LinearRegression

```
  0.003 * R
- 0.031 * C
- 0.286 * time_step
+ 0.041 * u_in
- 9.528 * u_out
+ 17.946
```

Figura 10: Coeficientes del modelo de regresión lineal

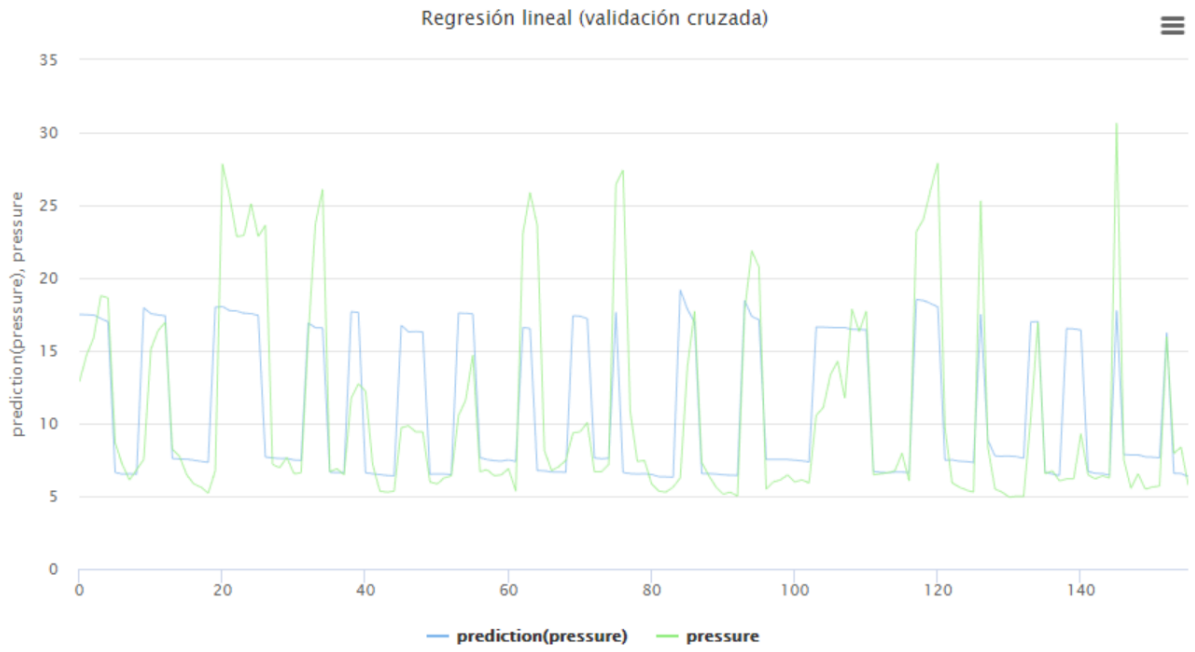


Figura 11: Resultado gráfico de la predicción del modelo de regresión lineal con validación cruzada

El modelo de regresión no llega a predecir correctamente los picos de presión de cada respiración

Modelo de árbol de decisión con validación simple

Resultados de validación simple

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 4.116 +/- 0.000
absolute_error: 2.123 +/- 3.526
relative_error: 17.28% +/- 33.84%
```

Figura 12: Resultado del modelo de árbol de decisión con validación simple

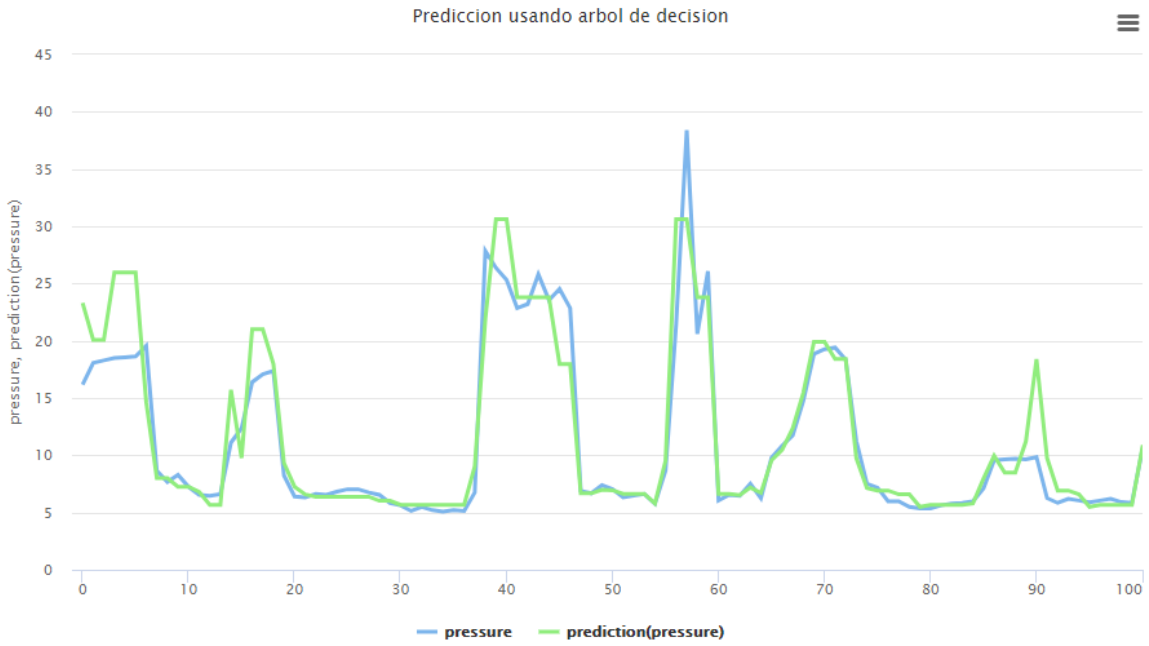


Figura 13: Comparación gráfica de la predicción del modelo de árbol de decisión con validación simple

Se observa que el modelo de árbol de decisión realiza una mejor predicción de los picos de presión en la respiración del usuario.

con Validación cruzada

Resultados de árbol de decisión con validación cruzada

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 4.065 +/- 0.158 (micro average: 4.068 +/- 0.000)
absolute_error: 2.099 +/- 0.058 (micro average: 2.099 +/- 3.485)
relative_error: 16.83% +/- 0.71% (micro average: 16.83% +/- 32.35%)
```

Figura 14: Resultado del modelo de árbol de decisión con validación cruzada

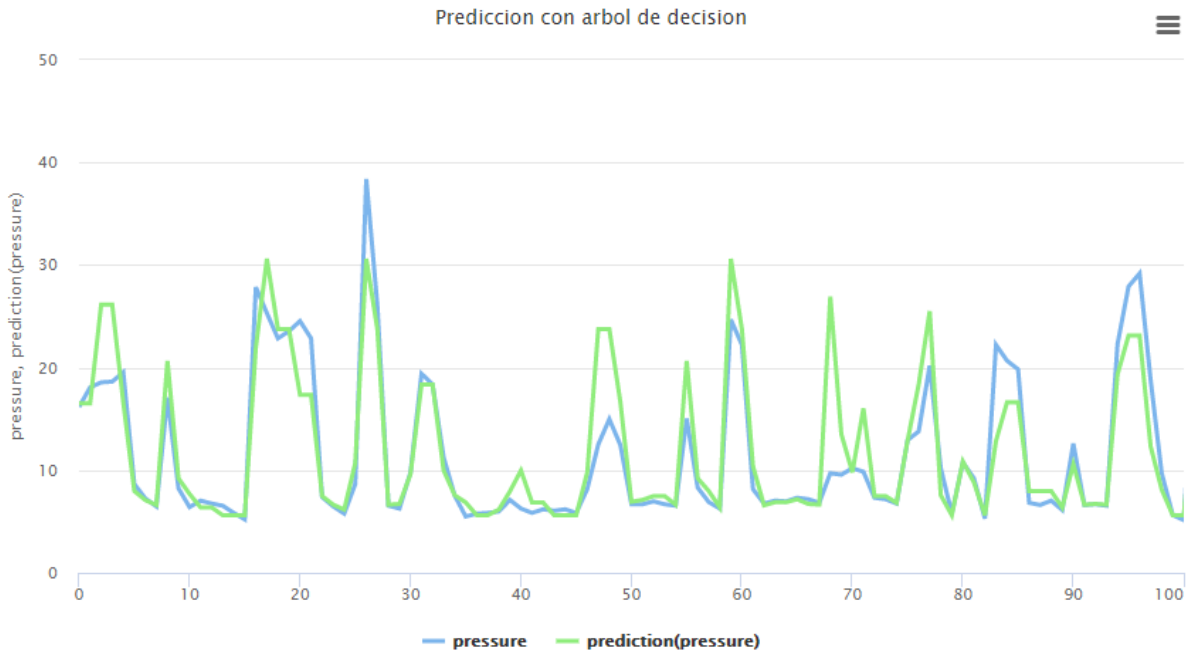


Figura 15: Comparación gráfica de la predicción del modelo de árbol de decisión con validación cruzada

Modelo de redes neuronales

Parámetros

Figura 16: Parámetros del modelo de redes neuronales

Capa oculta	Número de neuronas
uno	10
dos	10
tres	5

Tabla 3: Diseño de la red neuronal

Resultados de redes neuronales con validación simple

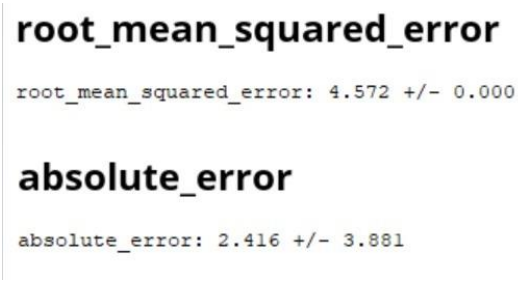


Figura 17: Resultados de modelo de redes neuronales con validación simple

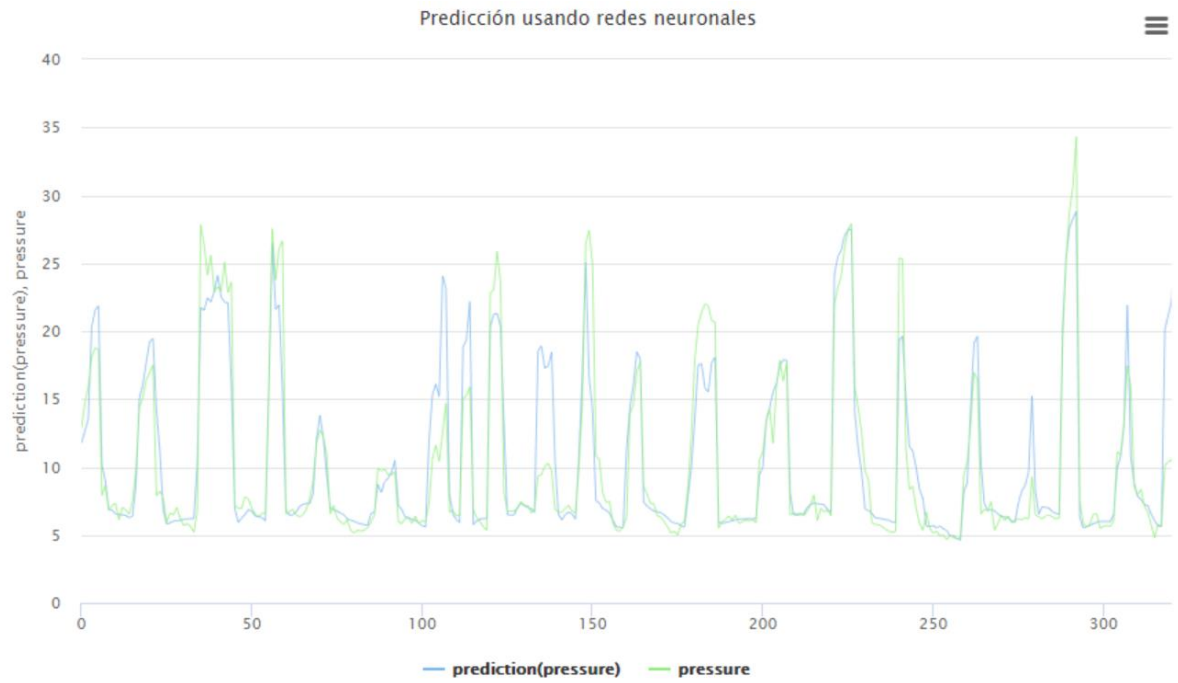


Figura 18: Comparación gráfica de la predicción del modelo de redes neuronales con validación cruzada

Resultados de redes neuronales con validación cruzada

root_mean_squared_error

```
root_mean_squared_error: 5.249 +/- 0.755 (micro average: 5.297 +/- 0.000)
```

absolute_error

```
absolute_error: 3.144 +/- 0.655 (micro average: 3.144 +/- 4.263)
```

Figura 19: Resultados de modelo de redes neuronales con validación cruzada

Paso 5: Interpretación / Evaluación

Comparación de validación simple y cruzada

	Error absoluto	
Modelo	**Validación simple**	**Validación cruzada**
Regresión lineal	3.985 ± 4.974	4.038 ± 0.076
Árbol de decisión	2.12 ± 3.526	2.099 ± 0.058
Redes neuronales	2.416 ± 3.881	3.144 ± 0.655

Tabla 4: Comparación de resultados de modelos de machine learning utilizando validación simple y cruzada

CONCLUSIONES

La validación cruzada otorgó un menor resultado en el error absoluto de todos los modelos (excepto el árbol de decisión), esto se lo asocia a que para la validación cruzada se utilizaron

muchos más registros para las pruebas, lo que podría ocasionar que el modelo presente una menor precisión en sus resultados a medida que se realicen más pruebas.

Además la incertidumbre de cada resultado es menor en la validación cruzada porque esta técnica realiza las pruebas con un número mayor de datos ya que realiza subparticiones de todos los registros en distintos grupos y efectúa las pruebas un determinado número de veces con los grupos resultantes.

También se puede observar que donde mayores problemas tuvieron los modelos fue en casos donde la presión del usuario llegaba a picos bastantes grandes.

El modelo que presenta mayor nivel de exactitud en sus resultados de ambas validaciones (simple y cruzada) es el modelo de árbol de decisión. Este modelo permite aproximar mejor la curva de presión a través del tiempo presente en estos datos.

El modelo de redes neuronales y el árbol de decisión se acercaron bastante en los resultados de la validación simple, pero la validación cruzada demostró que el modelo de redes neuronales en realidad no presenta ese valor debido a que se reduce su efectividad de predicción al realizar pruebas adicionales.

Como trabajo a futuro se propone investigar la aplicación de modelos de series de tiempo como ARIMA para comparar su desempeño con los resultados de los modelos utilizados en este documento.

BIBLIOGRAFÍA

[1] Google Brain - Ventilator Pressure Prediction, Overview. Obtenido de https://www.kaggle.com/c/ventilator-pressure-prediction/overview/description, consultado en Noviembre 2021.

[2] SK Gupta, V. B. (1997). ―A proposal for Data Mining Management System. Sk Wasan.‖

[3] Hernandez Orallo, (2005). ―Introducción a la Minería de Datos‖. España. Editorial Pearson Educación S.A.

[4] José Farfán, Mariela Rodriguez. (2020). ―Técnicas de Minería de Datos‖. Obtenido de https://jhfarfan.wixsite.com/datamining/tecnicas-dm-1, consultado en Noviembre 2021.

[5] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis de series temporales de hechos delictuales

Mariela Rodriguez, Nazarena Laureano, Jose H. Farfán

Institución: Facultad de Ingeniería, Facultad de Humanidades y Ciencias Sociales.

Datos de contacto: mariela.rodriguez@fi.unju.edu.ar, nazarenalaureano8@gmail.com, jhfarfan@fi.unju.edu.ar

RESUMEN

El análisis de la criminalidad en la Argentina es una tarea que se lleva adelante hace varias décadas y evoluciona a medida que se cuentan con nuevas técnicas y tecnologías.

Los algoritmos de clasificación y predicción de la minería de datos han permitido resolver diferentes problemas ramas de las ciencias y apoyan a la toma de decisiones. En este sentido, el presente trabajo realiza la predicción de series temporales de hechos delictuales ocurridos en el periodo del 2.018 al 2.021 en la provincia de Jujuy, Argentina, utilizando algoritmos de regresión lineal, árbol de decisión y redes neuronales, y se analiza el comportamiento de cada uno de ellos. Para medir el rendimiento, se utilizó el error cuadrático medio que genera cada uno de los modelos.

.

Palabras Clave: series temporales, redes neuronales, hechos delictuales

INTRODUCCIÓN

Los organismos de seguridad registran los hechos delictuales con la finalidad de generar acciones tendientes a la prevención del delito. Este registro varía de acuerdo a la modalidad, fechas, horas, tipos de eventos ocurridos, permitiendo formar un aprendizaje longitudinal e histórico.

La provincia de Jujuy registra los hechos delictuales que son informados al Sistema Nacional de Información Criminal a través del Centro de Información y Análisis Criminal, dependiente del Ministerio de Seguridad de la provincia. El Sistema Nacional de Información Criminal (SNIC) tiene como propósito la recopilación, gestión y difusión de las estadísticas criminales en el territorio nacional, conforme lo establecido por la Ley N° 25.266. El mismo comenzó a registrar los hechos delictuosos ocurridos en el territorio nacional desde enero de 1999. [1]

Según el SNIC en el año 2020 la provincia de Jujuy registró una disminución de 9,7% en la tasa de hechos delictivos, con un descenso generalizado en la mayoría de los códigos delictuales. Los delitos que registran aumentos significativos son: “Homicidios dolosos”, “Otros delitos contra la libertad”, “Tentativas de robo”, “Tentativas de hurto”, “Portación ilegal de armas de fuego”, “Contrabando simple”, “Otros delitos previstos en leyes especiales” [2]. La disminución de registro de hechos delictuales no solo influyó en la provincia sino que a nivel mundial afectados por los niveles de confinamientos que se tuvo en el año 2020. Desde la fase 1 de confinamiento que logró la restricción de actividades casi de forma total incidiendo en una notable disminución de delitos a otras fases que permitieron una mayor cantidad de actividades y por ende el crecimiento de delitos.

La ocurrencia de los delitos se conforma de un complejo conjunto de circunstancias en las que confluyen nuevas realidades sociales en el marco de una sociedad cada día más compleja por un lado, así como nuevas dinámicas delictivas y complejos riesgos por otro, todo ello sin olvidar las nuevas demandas dirigidas a los servicios públicos de seguridad que están motivando que las estrategias utilizadas por la Policía para dar respuesta a este escenario securitario deban adaptarse a la realidad actual. [3]

El análisis de la ocurrencia de hechos a través del tiempo tiene sus fundamentos en la técnica estadística de series temporales que son utilizados en campos como son la econometría. Sin embargo, este estudio es muy útil para otros campos como el análisis delictual. Por otro lado, el avance en las técnicas de minería de datos permite realizar el estudio de las series temporales con algoritmos que se utilizan para descripción y predicción de sucesos como árboles de clasificación y redes neuronales

METODOLOGÍA

Introduzca aquí el texto correspondiente a la metodología en ARIAL 11.

2.1. Datos

El presente informe se realizó en base a hechos delictuales ocurridos en la provincia de Jujuy, el dataset cuenta de 143053 registros desde el 2018 al 2021. Los registros comprenden desde el 1 de junio de 2018 al 31 de agosto de 2021, siendo un total de 23617 registros para el 2018, 51416 en el 2019, 39646 en el 2020 y 28374 en el 2021. De acuerdo a la figura 1 se muestra la variación de los datos, se observa que existen periodos en los que los hechos tienen una media estable, en días determinados del año ocurren picos de hechos que, posteriormente al día siguiente redunda en bajas pronunciadas. Desde el 18 de marzo al 19 de abril existe una considerable baja de hechos delictuales, que según los registros que se analizan son caracterizados como atípicos, este periodo corresponde al confinamiento estricto debido al COVID-19. Para ese mismo año entre las fechas del 5 de junio al 11 de octubre se denota que la curva va en ascendencia. Por último es destacable que, según los datos ocurren el día 22 de octubre ocurre una cantidad considerable de hechos llegando a superar los datos de

los tres años de análisis, se estima que a partir de esta fecha es que la cantidad de hechos se mantiene una media similar a los años anteriores.

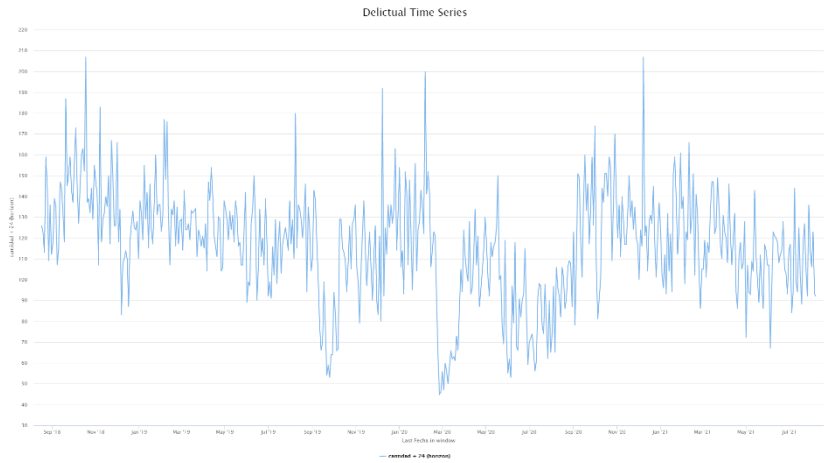


Figura 1: Serie temporal de hechos delictuales

2.2 Construcción del Modelo

Para la construcción del modelo que permite predecir la serie de tiempo en el que desarrollan los delitos se utilizó una comparación de un algoritmo de regresión lineal, un algoritmo de clasificación “Gradient Boosted” y un algoritmo de red neuronal Deep Learning

Regresión Lineal

Los modelos de regresión capturan cómo una o más variables objetivo varían con una o más variables de atributo. Se pueden utilizar para predecir los valores de las variables de destino utilizando los valores de las variables de atributo. Un modelo de regresión lineal que contiene múltiples variables atributos puede describirse como:

$$y = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i \quad (1)$$

donde (xi,yi) denotan la ith observación de x e y, ϵi representa el ruido aleatorio contribuido al ith observación de y.[4]

Gradient Boosted Tree

El algoritmo Gradient Boosted (GBDT) construye un árbol de decisión a la vez para adaptarse al residuo de los árboles que lo preceden. GBDT tiene la característica que cuenta con alta precisión y rápido entrenamiento [5].

GBDT permite especificar la función de pérdida como los modelos de aprendizaje base a pedido. Dada una función de pérdida específica Ψ (y, f) y / o una base de aprendizaje personalizado h (x, θ), la solución a las estimaciones de los parámetros puede ser difícil de obtener.

El algoritmo es de aprendizaje supervisado, por ello debe proporcionarse un objetivo. Se cuenta con el conjunto de datos (x, y)Ni = 1, donde x = (x1,⋯, xd) son las variables de entrada explicativas e y son las variables de salida. El objetivo es reconstruir la dependencia funcional desconocida x → fy con el estimador f^ (x), de modo que alguna función de pérdida especificada Ψ (y, f) se minimice[6]:

$$\hat{f}(x) = \arg\min_{f(x)} \Psi(y, f(x)) \quad (2)$$

Deep Learning

Las redes neuronales con retroalimentación multicapas (Deep Learning) son aquellas que disponen de un conjunto de neuronas agrupadas en varios (2, 3, etc.) niveles o capas. En estos casos, una forma para distinguir la capa a la que pertenece una neurona, consistiría en fijarse en el origen de las señales que recibe a la entrada y el destino de la señal de salida. Normalmente, todas las neuronas de una capa reciben señales de entrada desde otra capa anterior (la cual está más cerca a la entrada de la red), y envían señales de salida a una capa posterior (que está más cerca a la salida de la red)[7]. El algoritmo de Deep Learning se entrena con el descenso de gradiente estocástico utilizando el método de back-propagation. Internamente la red puede contener una gran cantidad de capas ocultas que consisten en neuronas con funciones de activación. Las funciones avanzadas, como la tasa de aprendizaje adaptativo, el recocido de tasas, el entrenamiento de impulso, la deserción y la regularización L1 o L2 permiten una alta precisión predictiva.

DESARROLLO

Medidas de Evaluación

Validación Cruzada

La validación cruzada es una técnica para evaluar los modelos de aprendizaje automático. Para su validación subdivide el conjunto de datos y por cada subconjunto cuenta con datos de entrenamiento y datos de evaluación. Una vez entrenado el modelo se aplica un proceso para evaluar con los datos destinados para tal fin. El rendimiento del modelo se mida al final de la fase de prueba. Este método estadístico ayuda a comparar y seleccionar el modelo en el aprendizaje automático aplicado.
La validación que se utiliza en este trabajo es de K iteraciones. Para tal fin se divide los datos en en k subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto $(k-1)$ de los subconjuntos como datos de entrenamiento. El proceso es repetido durante k-interacciones, con cada uno de los posibles subconjuntos de datos de prueba.
El método es muy preciso evalúa a partir de las k combinaciones de datos de entrenamiento y de prueba [8][9].

Error cuadrático medio RMSE

El error cuadrático medio es un criterio de evaluación para modelos que utilizan regresión de datos. El RMSE es una medida de ajuste absoluto e indica cuan cerca están los puntos de datos observados de los valores modelados en la predicción. Se asume que se cuenta con n muestras de errores del modelo ϵ calculados como $(y_i - \hat{y}_i)$ i = 1, 2,..., N). También se asume que el conjunto de muestra de error $\epsilon=(y_i - \hat{y}_i)$ es insesgado. El RMSE se calcula para el conjunto de datos como [10]:

$$\text{RMSE} = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (3)$$

Se supone que los errores son insesgados y siguen una distribución normal. Los valores más bajos de RMSE indican un mejor ajuste. RMSE es una buena medida de la precisión con que el modelo predice la respuesta.

Resultados

Inicialmente se subdivide el conjunto de datos de series temporales para convertirlos en un dataset que pueda ser trabajado con los algoritmos de regresión lineal, de gradiente mejorado GBDT y Deep Learning. Los datos fueron divididos en 48 grupos que avanza cada dos días para la evaluación de datos.

El dataset se validó de forma cruzada utilizando un k=10 para los tres algoritmos utilizados y la medida de rendimiento es por medio del error medio cuadrado RMSE.
Evaluando el dataset con el algoritmo de regresión lineal con un mínimo de tolerancia de 5%. Para el algoritmo GBDT se utilizó 100 árboles de decisión, con una profundidad máxima de 5 con un mínimo de filas de 10. Por último para el algoritmo de Deep learning se utilizó dos capas de 50 neuronas cada una.
De acuerdo a la tabla 1 el algoritmo de Regresión Lineal es el que cuenta con menor dispersión de datos, la media es la más cercana a la media original. Según la figura 2 se tiene un error bajo de predicción entre el periodo de agosto del 2018 a setiembre del 2019. En el periodo de setiembre a octubre del 2019 se incrementó el error por la caída brusca de cantidad de hechos. En el periodo de marzo a abril del 2020 la predicción es mayor a la cantidad de hechos ocurridos, que corresponde al periodo de confinamiento debido al COVID-19 y por último el periodo de agosto del 2020 la cantidad de hechos presenta una curva de crecimiento al igual que la curva de predicción.

Tabla 1: Dispersión de datos

Algoritmo	Min	1º cuartil	Media	3º cuartil	Max
DataSet Original	45	101	118	133	207
Linear Regresion	80	109	117	124	147
Gradient Boosted	71	106	116	124	168
Deep Learning	74	111	122	132	170

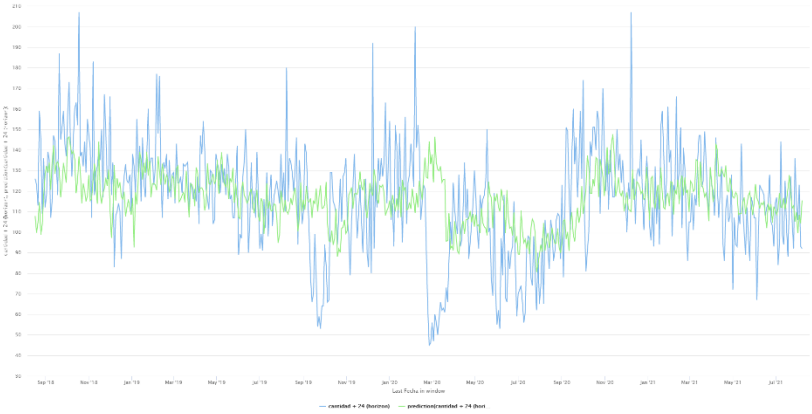


Figura 2: Predicción de serie temporal de hechos delictuales con Regresión Lineal

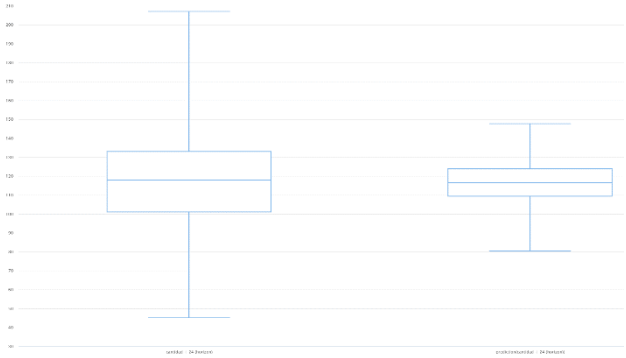


Figura 3: Diagrama de caja y bigotes con algoritmo de Regresión Lineal

Los resultados del modelado con el algoritmo GBDT según tabla 1 muestra que cuenta con mayor dispersión que el algoritmo de regresión lineal, siendo el primer cuartil, media y tercer cuartil menores a los originales y en la figura 4 se ve que la curva de predicción en el mayor rango de tiempo se encuentra por debajo de la curva original. Sin embargo, los periodos de septiembre a octubre del 2019 y de marzo a abril del 2020 no se lograron a predecir.

Figura 4: Predicción de serie temporal de hechos delictuales con GBDT

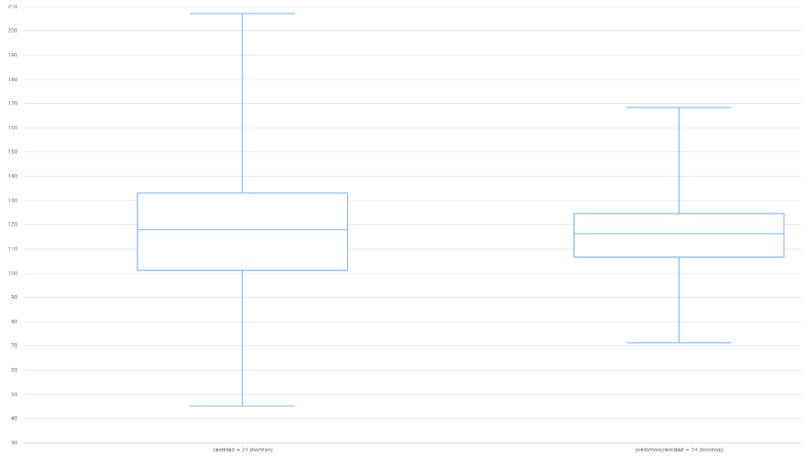


Figura 5: Diagrama de caja y bigotes con algoritmo GBDT

El modelo de predicción de datos con el algoritmo de Deep Learning es el que tiene mayor dispersión en sus valores, que se ve reflejado en la tabla 1. Se puede observar en la figura 6 que la curva de predicción se coloca por arriba de la original, que coincide con los valores del 1º y 3º cuartil que se posicionan por arriba del 1º y 3º cuartil original respectivamente.

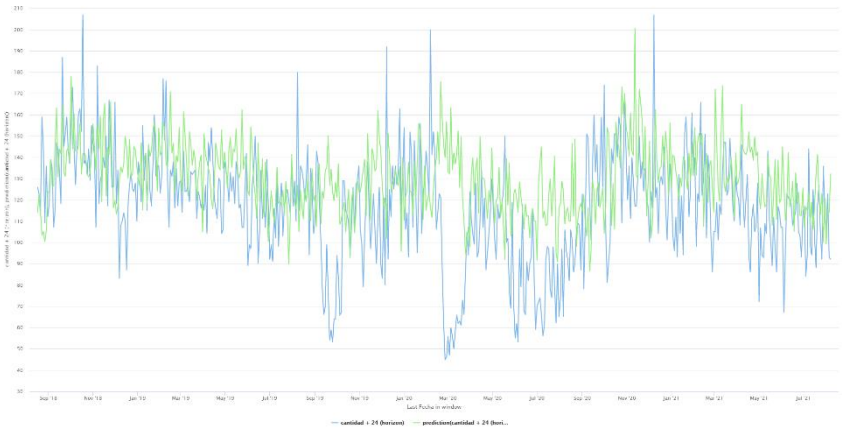


Figura 6: Predicción de serie temporal de hechos delictuales con Deep Learning

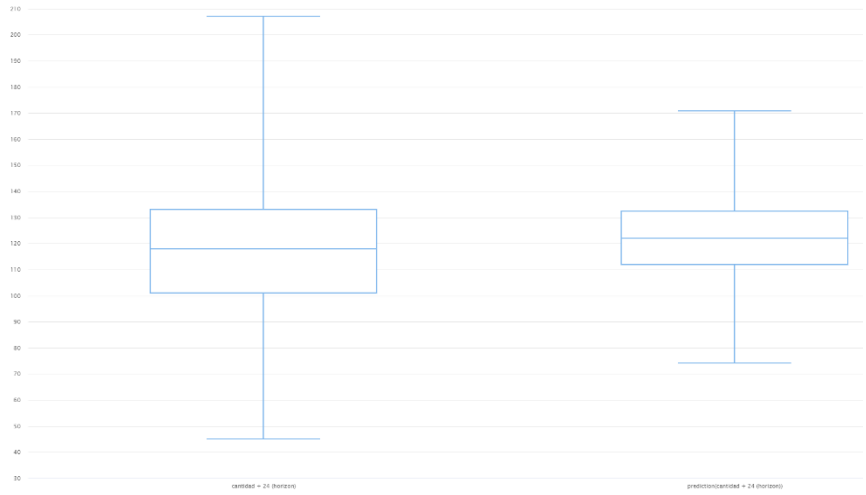


Figura 7: Diagrama de caja y bigotes con algoritmo de Regresión Lineal

Al realizar la comparación del RMSE, la medida de performance elegido para analizar las series temporales de hechos delictuales, permite distinguir que el comportamiento del algoritmo de regresión lineal es superior a GBDT y Deep Learning.

Tabla 2: Medida de Performance – RMSE

Algoritmo	Root mean squared error	Relative error
Linear Regresion	26.012 +/- 3.255	19.63% +/- 3.43%
Gradient Boosted	27.164 +/- 3.779	20.70% +/- 4.21%
Deep Learning	31.282 +/- 4.365	25.54%+/-5.39%

CONCLUSIONES

El análisis de la serie temporal de los hechos delictivos es una tarea pendiente a nivel regional que merece ser estudiado académicamente y operativamente. Existen técnicas estadísticas propias del análisis temporal pero fue un desafío utilizar algoritmos de machine learning para este análisis. La no estacionariedad del dataset en el periodo del 2018 al 2021 es un hecho poco casual influido por las medidas de confinamiento que se tomaron para mitigar el COVID-19 que provocan que los algoritmos cuenten con un RMSE elevado. El algoritmo de regresión lineal que cuenta una fórmula sencilla supera a algoritmos de mayor robustez como son GBDT y Deep Learning.

Este primer análisis permite generar futuros análisis más detallados como son la subdivisión de periodos que puedan ser estacionarios y análisis por cada tipo de delitos. Otro análisis de interés, es medir el nivel de resiliencia que se tiene de los hechos delictivos en general como también por cada tipo de ellos como son los delitos contra la propiedad y delitos contra las personas en el periodo de confinamiento por COVID-19.

BIBLIOGRAFIA

[1] Sub Secretaria de Estadística Criminal (2019), Robos y Tentativas de Robo del 2001 al 2018. Consultado en setiembre de 2021: https://estadisticascriminales.minseg.gob.ar/reports/ROBOS_Y_TENTATIVAS_DE_ROBO_2001_al_2018.pdf

[2] Sub Secretaria de Estadística Criminal (2021), Informe de Estadísticas Criminales 2020. Consultado en setiembre 2018:https://estadisticascriminales.minseg.gob.ar/reports/InformeSNIC2020.pdf

[3] Daniel Salafranca Barreda, Manuel Rodríguez Herrera, “MODELO SDIK: UN SISTEMA ANALÍTICO PARA LA PREDICCIÓN DEL DELITO”, http://www.pensamientopenal.com.ar/system/files/2017/07/miscelaneas45513.pdf

[4] Nong Ye (2014). Data mining Theories, Algorithms, and Examples. Taylor & Francis Group

[5] Si Si, Huan Zhang y Cia, Gradient Boosted Decision Trees for High Dimensional Sparse Output http://proceedings.mlr.press/v70/si17a/si17a.pdf

[6] Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232.

[7] Matich, Damián Jorge (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Universidad Tecnológica Nacional.

[8] Rodriguez Murillo, Natalia (2019). Análisis de validación cruzada bajo diferentes condiciones de ruido. Tecnológico Nacional de México en Celaya.

[9]Joanneum, F. H. (2005) Cross-Validation Explained. Obtenido de Institute for Genomics and Bioinformatics - Graz University of Technology.

[10] T. Chai & R. R. Draxler (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literatura. University Research Court, College Park, University of Maryland, College Park.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Consumo de gaseosas de segundas marcas

Arce, Cristian Alejandro; Arias, María Candela; Gutiérrez, Nicole Aylen; Reyes, Héctor Arnaldo; Vélez, Rodrigo Alejandro

Institución: Facultad de Ingeniería- Universidad Nacional de Salta. Salta.

Datos de contacto:
Cax459@gmail.com (+54) 387-5180155
Candelitaarias6@gmail.com (+54) 388-6555201
reyeshectorarnaldo@gmail.com. (+54) 387-5482966
nicolexeneixe@gmail.com (+54) 387-6192465
rodrigovelez68@gmail.com (+54) 387-5372321

RESUMEN

En este trabajo de investigación se estudiará las influencias de dos factores en la elección de gaseosas para el consumo de familias salteñas. Estos dos factores son la marca y el sabor, se determinará cuál es el factor que más influye en la elección del producto aplicando conocimientos y criterios aprendidos en la materia Estadística Experimental.

Principalmente se ordenarán los datos recolectados para mejorar su utilización, luego se implementará un modelo que explique el comportamiento de los datos con sus respectivas condiciones. Se verificarán las condiciones y se realizará una prueba adicional para validar dicho modelo. Finalmente se realizarán comparaciones entre los tratamientos del factor que haya sido el más influyente para que de esta forma saber las preferencias de la comunidad salteña.

INTRODUCCIÓN

En este proyecto analizaremos gaseosas de segunda línea que consumen cotidianamente las familias salteñas, para ello tendremos en cuenta la marca y el sabor de las gaseosas. Las marcas seleccionadas son Talca, Secco y Manaos, mientras que los sabores son cola, manzana, naranja y pomelo. Se eligieron estas marcas considerando que tienen una competencia más pareja entre las mismas, descartando así las marcas más preferibles como ser Coca-Cola, Pepsi, Paso de los Toros, entre otras. La elección de los sabores fue llevada a cabo teniendo en cuenta que son los más elegidos para acompañar almuerzos y cenas, también porque son comunes a todas las marcas.

De este análisis podremos determinar si la elección de las gaseosas depende de la marca (si la elección va por un hecho de calidad ya conocida) o simplemente por la preferencia del sabor de la gaseosa.

METODOLOGÍA

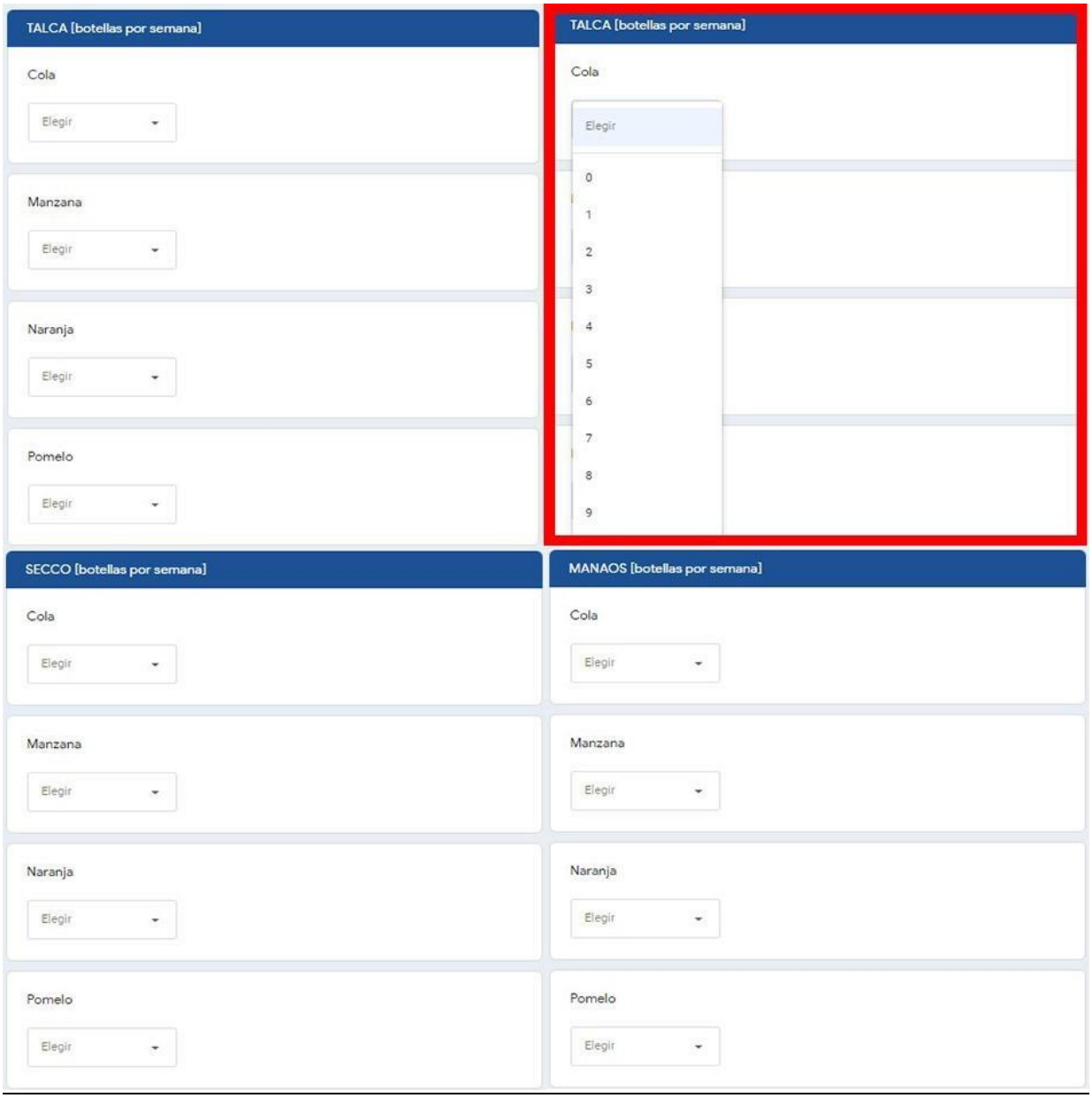
Se realizó una encuesta para la obtención de datos (observación) acerca de la cantidad aproximada de botellas de gaseosa que consume una familia, así como también la marca de las cuales hemos seleccionado (Talca, Secco, Manaos) y el sabor (Cola, Manzana, Naranja, Pomelo). Se solicitó que se ingresará la cantidad de botellas de gaseosas que suele consumir en una semana, junto con el sabor de la misma.

En la parte analítica plantearemos un diseño factorial destinado al estudio del mercado. El objetivo es saber si la elección de la gaseosa depende de la marca o del sabor, teniendo en cuenta la elección de las marcas y sabores para el estudio ya mencionados en la introducción. El diseño elegido es un diseño especial desarrollado para investigar más de un factor a la vez, en esta investigación los factores serían la marca y el sabor. Aquí la observación obtenida por medio de la encuesta proporciona información sobre todos los factores, y es factible ver las respuestas de un factor en diferentes niveles de otro factor en el mismo experimento. La respuesta a cualquier factor observado en diferentes condiciones indica si los factores actúan en las unidades experimentales de manera independiente. La interacción entre los factores ocurre cuando su actuación no es independiente, esto se analizará más adelante.

En resumen, el diseño factorial consiste en realizar todas las combinaciones posibles de los niveles de varios factores.

Datos de la encuesta

Para la recopilación de datos se hizo la siguiente encuesta, con las siguientes preguntas:



Dentro del recuadro rojo, las respuestas proporcionadas al encuestado.

DESARROLLO

Por medio de las encuestas se obtuvo la cantidad de botellas de gaseosas que consume una familia salteña por semana. La cantidad total de familias que respondieron la encuesta fue de 104.

Para determinar si la elección de la gaseosa depende de la marca o del sabor, principalmente se ordenaron los datos obtenidos, por lo que cada combinación sabor/marca sea el resultado de sumar la cantidad de botellas que ingresaron los encuestados, de esta manera se obtuvo el siguiente cuadro:

Tabla de datos

Marca \ Sabor	Cola	Manzana	Naranja	Pomelo
Talca	19	40	15	37
Secco	21	36	13	58
Manaos	24	43	9	26

Modelo Estadístico Lineal del experimento

$$y_{ijk} = \mu + \alpha_i + \beta_i + (\alpha\beta)_{ijk} + \varepsilon_{ijk}$$

$$\mu = media\ general$$

$$\alpha_i = efecto\ que\ produce\ el\ factor\ a$$

$$\beta_i = efecto\ que\ produce\ el\ factor\ b$$

$$(\alpha\beta)_{ijk} = interacción$$

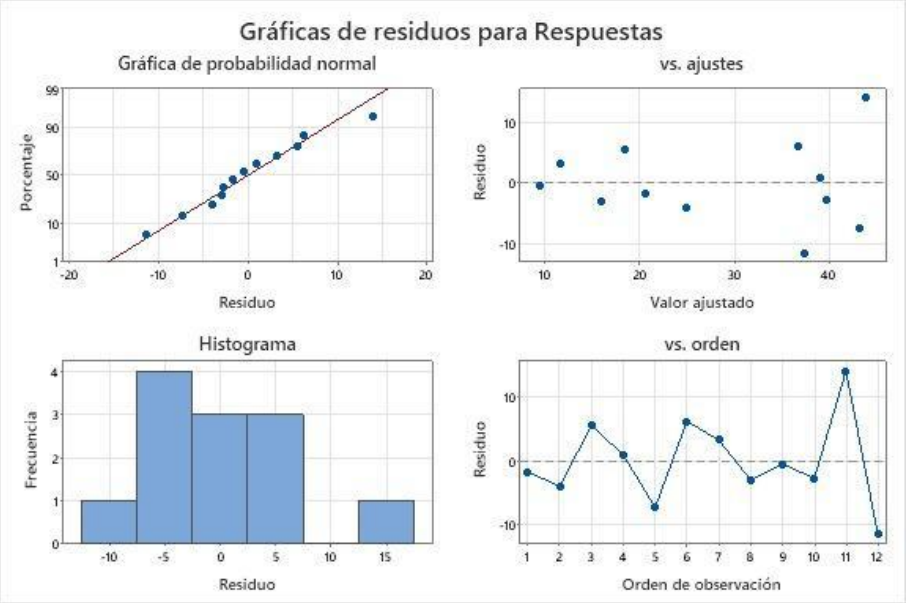
$$\varepsilon_{ijk} = error\ aleatorio$$

Para realizar este modelo, se debe cumplir que:

$$\varepsilon_{ijk} \sim N(0; \sigma^2)$$

Para verificar que el error tenga una distribución normal, se realiza la gráfica de residuos:

Grafica de residuos



(Kuehl, 2001)

A partir de la gráfica de probabilidad normal se observa que se tiene una distribución normal, ya que los residuos se aproximan a una recta, por lo que se tiene una distribución no sesgada con respecto a la distribución normal estándar.

En la parte superior derecha, residuos vs ajustes, se puede comprobar el supuesto de que los residuos tienen una varianza constante.

Mediante el histograma podemos concluir que no existen valores atípicos y que es aproximadamente simétrico.

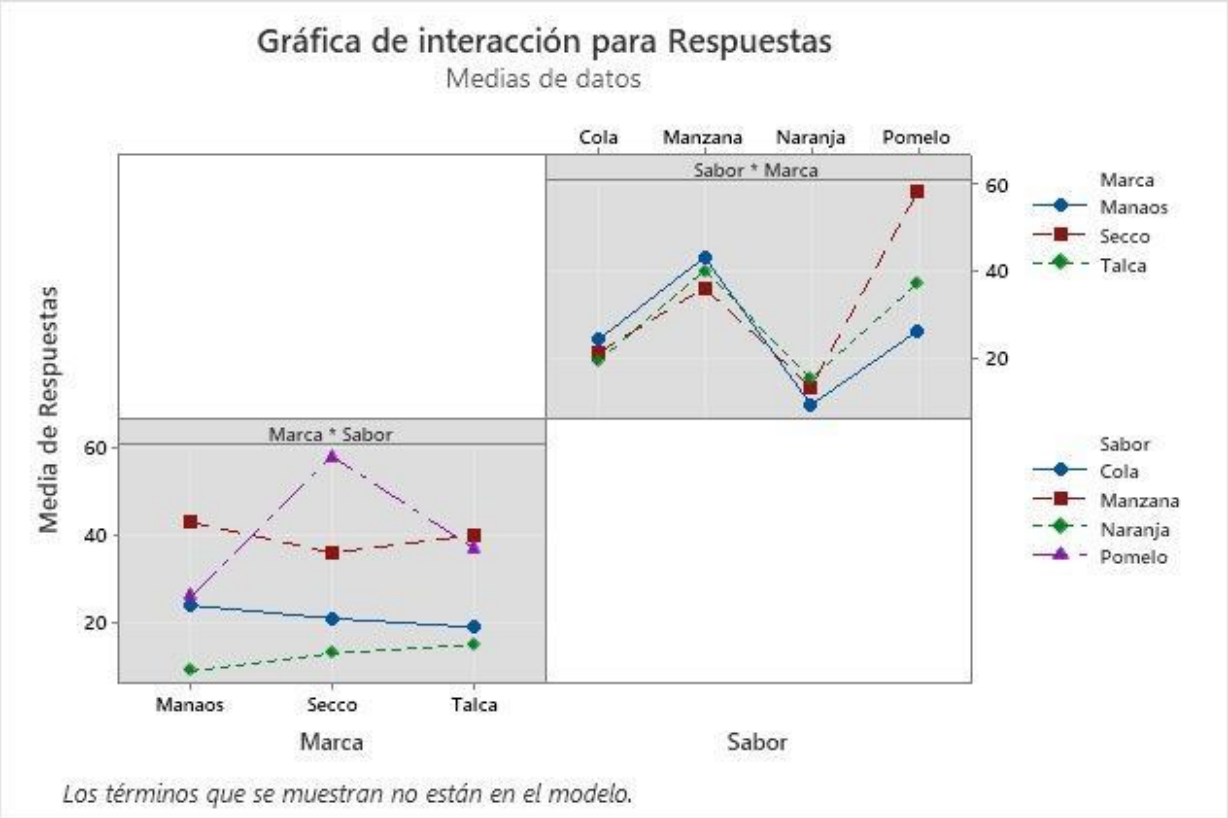
En la gráfica de residuos vs. Orden, se puede comprobar el supuesto de que los residuos no están correlacionados entre sí.

Para continuar con el estudio el modelo lineal deberá satisfacer esta ecuación, en donde no se aprecia las interacciones.

$$y_{ijk} = \mu + \alpha_i + \beta_i + \varepsilon_{ijk}$$

Se verificará que no existan las interacciones realizando la observación de la gráfica de interacción para respuestas.

Grafica de interacciones



(Kuehl, 2001)

Se considera que si hay interacciones, por lo que se deberá realizar la prueba de Tukey de la no aditividad.

Prueba de Tukey de aditividad

$$y_{ijk} = \mu + \alpha_i + \beta_i + \varepsilon_{ijk}$$

Análisis de Varianza

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Marca	87,16666667	2	43,5833333	0,52562814	0,616104318	5,14325285
Sabor	1732,25	3	577,416667	6,9638191	0,02216407	4,757062663
Error	497,5	6	82,9166667			
Total	2316,916667	11				

(Kuehl, 2001)

$$P_j = \sum_i^a \sum_j^b y_{ij} * Y_i * Y_j - Y_{..}(SCA + SCB + \frac{Y^2}{ab}) = 10868{,}5$$

$$SC(no\ aditividad) = \frac{p^2}{ab * SCA * SCB} = 65{,}192399$$

Análisis de varianza de aditividad

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Marca	87,16666667	2	43,5833333	0,52562814	0,616104318	5,14325285
Sabor	1732,25	3	577,416667	6,9638191	0,02216407	4,757062663
Error	497,5	6	82,9166667			
No aditividad	65,192399	1	65,192399	0,75400478	0,424919527	6,61
Residual	432,3076	5	86,46152			
Total	2316,916667	11				

(Kuehl, 2001)

$$Ho: (\alpha\beta)_{ijk} = 0$$

$$Ha: (\alpha\beta)_{ijk} \neq 0$$

$$Region\ de\ Rechazo \rightarrow Fo > Fcritco$$

$$como\ Fo < Fcritico \rightarrow 0{,}75 < 6{,}61$$

Por lo que no se rechaza la hipótesis nula, entonces **no hay interacción** entre los factores.

Entonces se continúa con el análisis.

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Marca	87,16666667	2	43,5833333	0,52562814	0,616104318	5,14325285
Sabor	1732,25	3	577,416667	6,9638191	0,02216407	4,757062663
Error	497,5	6	82,9166667			
Total	2316,916667	11				

(Kuehl, 2001)

Para el factor Marca:

$$Ho: \underline{\mu}_{talca} = \underline{\mu}_{secco} = \underline{\mu}_{manaos} = 0$$

$$Ha: \underline{\mu}_{talca} \neq \underline{\mu}_{secco} \neq \underline{\mu}_{manaos}$$

$$Fo = \frac{CMA}{CME} = 0{,}525; \quad Fcritico = 5{,}1432$$

$$Region\ de\ Rechazo \rightarrow Fo > Fcritco$$

$$como\ Fo < Fcritico \rightarrow 0{,}525 < 5{,}14$$

Por lo tanto, no se rechaza la hipótesis nula, entonces no hay efecto significativo del factor marca.

Para el factor Sabor:

$$Ho: \underline{\mu}_{Cola} = \underline{\mu}_{Manzana} = \underline{\mu}_{Naranja} = \underline{\mu}_{Pomelo} = 0$$

$$Ha: \underline{\mu}_{Cola} \neq \underline{\mu}_{Manzana} \neq \underline{\mu}_{Naranja} \neq \underline{\mu}_{Pomelo}$$

$$Fo = \frac{CMA}{CME} = 6{,}96; \quad Fcritico = 4{,}757$$

$$Region\ de\ Rechazo \rightarrow Fo > Fcritco$$

$$como\ Fo > Fcritico \rightarrow 6{,}96 > 4{,}757$$

Por lo tanto, se rechaza la hipótesis nula, entonces hay efecto significativo del factor sabor.

Se realizaría comparaciones entre los distintos sabores para conocer cual o cuales son las que más difieren. Se empleará el siguiente método.

Comparaciones aplicando el método de Tukey

Cuadro 3.11 Método de Tukey para todas las comparaciones por pares

Para un grupo de k medias de tratamiento, se calcula la diferencia honestamente significativa como:

$$\text{DHS}\ (k;\ \alpha_E) = q_{\alpha,k,v}\sqrt{\frac{s^2}{r}} \qquad \textbf{(3.44)}$$

donde $q_{\alpha,k,v}$ es el estadístico estandarizado de Student para un grupo de k medias de tratamiento en un arreglo ordenado. Los valores críticos de la tasa de error con respecto al experimento, α_E, y los v grados de libertad, se pueden encontrar en la tabla VII del apéndice.

Intervalos de confianza simultáneos de $100(1 - \alpha)\%$

Las estimaciones de los intervalos simultáneos de dos lados para el valor absoluto de todas las diferencias por pares, $\mu_i - \mu_j$ para toda $i < j$ son:

$$|\bar{y}_i - \bar{y}_j| \pm \text{DHS}(k,\ \alpha_E) \qquad \textbf{(3.45)}$$

Prueba de desigualdades $100(1 - \alpha)\%$ confiables

Se establece que dos medias de tratamientos no son iguales, $\mu_i - \mu_j \neq 0$, si:

$$|\bar{y}_i - \bar{y}_j| > \text{DHS}(k,\ \alpha_E) \qquad \textbf{(3.46)}$$

(Kuehl, 2001, p. 108)

Comparaciones por parejas de Tukey: Sabor

Agrupar información utilizando el método de Tukey y una confianza de 95%

Sabor	N	Media	Agrupación	
Pomelo	3	40.3333	A	
Manzana	3	39.6667	A	
Cola	3	21.3333	A	B
Naranja	3	12.3333		B

Las medias que no comparten una letra son significativamente diferentes.

Pruebas simultáneas de Tukey para diferencias de las medias

Diferencia de Sabor niveles	Diferencia de medias	EE de diferencia	IC simultáneo de 95%	Valor T	Valor p ajustado
Manzana - Cola	18.33	7.43	(-7.43; 44.09)	2.47	0.164
Naranja - Cola	-9.00	7.43	(-34.76; 16.76)	-1.21	0.643
Pomelo - Cola	19.00	7.43	(-6.76; 44.76)	2.56	0.147
Naranja - Manzana	-27.33	7.43	(-53.09; -1.57)	-3.68	0.039
Pomelo - Manzana	0.67	7.43	(-25.09; 26.43)	0.09	1.000
Pomelo - Naranja	28.00	7.43	(2.24; 53.76)	3.77	0.035

Nivel de confianza individual = 98.66%

(Kuehl, 2001)

Hay diferencias significativas entre los siguientes sabores:

Contraste 4: Naranja - Manzana

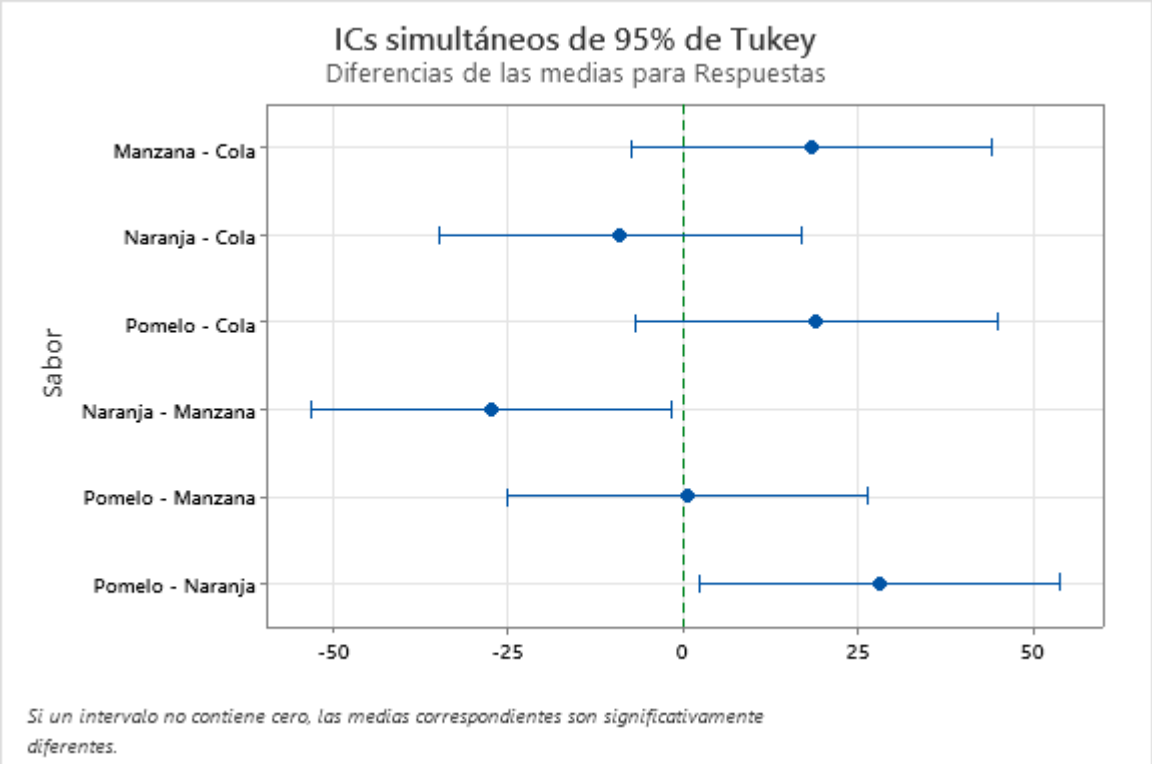
Contraste 6: Pomelo - Naranja

Como en el contraste entre Manzana y Cola, el intervalo de confianza incluye el cero cabe la posibilidad de que no haya diferencia entre esos sabores.

Como en el contraste entre Naranja y Cola, el intervalo de confianza incluye el cero cabe la posibilidad de que no haya diferencia entre esos sabores.

Como en el contraste entre Pomelo y Cola, el intervalo de confianza incluye el cero cabe la posibilidad de que no haya diferencia entre esos sabores.

Como en el contraste entre Pomelo y Manzana, el intervalo de confianza incluye el cero cabe la posibilidad de que no haya diferencia entre esos sabores.



(Kuehl, 2001)

Observando los intervalos de confianza simultáneos de 95% de Tukey, se infiere en que los sabores más elegidos son pomelo y manzana.

CONCLUSIONES

A partir del trabajo realizado, hemos llegado a la conclusión de que no importa la marca de la gaseosa, ya que todas tienen el mismo efecto, o dicho de otra forma la marca no tiene efecto en el tratamiento. Mientras que el factor sabor si es influyente por lo que se observaron diferencias entre ellas a partir del método de a par de Tukey. Con esto se obtuvo que los sabores más diferenciados, es decir los más consumidos por la comunidad encuestada, son pomelo y manzana. En definitiva, se puede inferir en que la elección de la bebida gasificada depende más del sabor que de la marca.

En el ámbito industrial, las marcas elegidas que son Talca, Secco y Manaos pueden utilizar esta información para tomar medidas como ser el mejoramiento de la calidad del sabor o realizar una mayor producción de los sabores más elegidos.

BIBLIOGRAFÍA

Kuehl, R. O. (2001). *Diseño de experimentos: Principios estadísticos de diseño y análisis de Investigación*.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis de tiempos de demora de recargas virtuales entre distintas empresas

Nuñez Duran Alex Luis, Narvaez Alan Marcelo y Basquez Mateo

Institución: Facultad de Ingeniería; Universidad Nacional de Salta. Salta Capital.

Datos de contacto: alexnunez4252@gmail.com, alannarvaez9910@gmail.com , 154meb@gmail.com

RESUMEN

En este proyecto se compararon dos empresas que ofrecen a pequeños comercios el servicio de ventas de saldo virtual, se buscó aquella empresa que brinde el servicio de mayor velocidad para realizar las recargas.

Se planteó un diseño de experimentos unifactoriales tomando como factor la empresa y las compañías telefónicas como bloques que dividirán las muestras para una mayor determinación a la hora de llegar a una conclusión. En primera instancia se determinará si en verdad hay diferencias de velocidad haciendo uso de herramientas estadísticas como el análisis de varianzas y pruebas de contraste para que, en caso de existir tal diferencia, escoger la mejor empresa.

Palabras clave: Recargas virtuales, tiempo, diseño unifactorial, empresa.

INTRODUCCIÓN

¿Cómo se realiza una recarga?

Un integrante del grupo tiene un kiosco en donde entre otras cosas ofrece la venta de saldo virtual. Para realizar una recarga al cliente se le pide el número de teléfono, la compañía a la que pertenece (personal, claro, etc.) y el monto a cargar. Se ingresan los datos a la máquina y en unos instantes se aprueba la operación, aunque la recarga no es inmediata, tarda unos segundos en llegar el mensaje de confirmación de recarga exitosa al celular. Las empresas de las que se dispone son las llamadas ReVirtual y Tcargo.

I Máquinas para cargas virtuales

El proyecto consiste en comparar los tiempos de demora en consolidar una recarga de saldo a celulares luego de ser aprobada la misma en el dispositivo de recarga, se medirán en función del tipo de empresa que brinda el servicio de venta de saldo. Además del factor empresa se tomará en cuenta la compañía telefónica del número a cargar, como esta no es una variable que se pueda manipular a la hora de realizar una recarga, se usarán como bloques.

METODOLOGÍA

Diseño Experimental

Está relacionado básicamente con el planeamiento de la recolección de los datos. Por lo general, un experimento es realizado por una o varias de las razones siguientes: Identificar las principales causas de variación en la respuesta, encontrar las condiciones que permitan alcanzar un valor ideal en la respuesta, comparar las respuestas a diferentes niveles de factores controlados por el investigador y/o construir modelos que permitan obtener predicciones de la respuesta.

Definiciones Básicas

Variable Respuesta: es la variable en estudio, aquella cuyos cambios se desean estudiar. Es la variable dependiente.

Factor: es la variable independiente. Es la variable que manipula el investigador, para estudiar sus efectos sobre la variable dependiente.

Nivel Del Factor: es cada una de las categorías, valores o formas específicas del factor.

Experimento Unifactorial: es aquel en el que se estudia un solo factor.

Tratamientos: Conjunto de condiciones experimentales que serán impuestas a una unidad experimental en un diseño elegido. En experimentos unifactoriales, un tratamiento corresponde a un nivel de factor. En experimentos multifactoriales, un tratamiento corresponde a la combinación de niveles de factores.

Unidad Experimental: es la parte más pequeña de material experimental expuesta al tratamiento, independientemente de otras unidades.

Error Experimental: Describe la variación entre las unidades experimentales tratadas de forma idéntica e independiente. Orígenes del error experimental: variación natural entre unidades experimentales, variabilidad en la medición de la respuesta, imposibilidad de reproducir idénticas condiciones del tratamiento de una unidad a otra, interacción de tratamientos, cualquier factor externo.

Mediciones: Son los valores de la variable dependiente, obtenidos de las unidades experimentales luego de la aplicación de tratamientos.

Diseño de bloques completos aleatorizados

La bloquización, un método para reducir la variación del error experimental, agrupa las unidades experimentales en bloques para comparar tratamientos en un medio más homogéneo.

Cualquier factor que afecta la variable de respuesta y que varía entre las unidades experimentales aumenta la varianza del error experimental y disminuye la precisión de los resultados del experimento.

El uso de bloques estratifica las unidades experimentales en grupos homogéneos, o unidades parecidas. Una buena elección de los criterios de bloquización disminuye la variación entre las unidades dentro de los bloques en comparación con las unidades de diferentes bloques

En él diseño de bloques completos aleatorizados cada tratamiento se asigna al azar a un número igual de unidades experimentales en cada bloque y es posible hacer comparaciones más precisas entre los tratamientos dentro del conjunto homogéneo de unidades experimentales en un bloque.

Aleatorizar el diseño:

La asignación aleatoria de tratamientos a las unidades experimentales está restringida de manera que cada tratamiento debe presentarse el mismo número de veces dentro de cada bloque.

Una permutación aleatoria del orden en el que se colocan los tratamientos en las unidades de cada bloque proporciona una asignación aleatoria de los tratamientos a las unidades, Se selecciona una permutación al azar para cada bloque ya que se requiere una aleatorización separada para cada uno de ellos.

Modelo estadístico y análisis para el diseño de bloques completos aleatorizados:

El modelo lineal para un experimento en un diseño de bloques completos aleatorizado requiere un término que represente la variación identificable en las observaciones como consecuencia de los bloques. La respuesta de la unidad con el i-ésimo tratamiento en el j-ésimo bloque se escribe como:

$y_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$

$i = 1,2,\ldots,a \quad ; \quad j = 1,2,\ldots,b$

Donde:

y_{ij}: *valor observado*

μ: *media general*

τ_i: *efecto fijo de los tratamientos*

ρ_j: *efecto fijo de los bloques*

ε_{ij}: *error experimental*

Se supone que los efectos del tratamiento y del bloque son aditivos, lo que quiere decir que no existe interacción entre tratamientos y bloques; también se supone que los errores experimentales son independientes (independencia justificada a través de la asignación aleatoria de los tratamientos a las unidades experimentales), con medias cero y varianza común σ^2

Si aplicamos el *Método de los Mínimos Cuadrados*, para estimar los parámetros:

$\tau_i = y_{i.} - y_{..}$

$\rho_j = y_{.j} - y_{..}$

$\varepsilon_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}$

Cada componente del modelo contribuye a la variabilidad total. La partición de la Suma de Cuadrados Total involucrará tres fuentes de variación.

$y_{i.} - y_{..}$ $\quad y_{.j} - y_{..}$ $\quad y_{ij} - y_{i.} - y_{.j} + y_{..}$

$$\sum_i \sum_j (y_{ij} - y_{..})^2 = a \sum_i (y_{i.} - y_{..})^2 + b \sum_j (y_{.j} - y_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.} - y_{.j} + y_{..})^2$$

SCT = SCA + SCB + SCE

Siendo:

a: n° de niveles del factor

b: n° de bloques

Tabla de Análisis de varianza para un experimento con un diseño de bloques completo aleatorizado:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F calculada
Tratamientos	SCA	a - 1	CMA = SCA/(a-1)	Fo = CMA/CME
Bloques	SCB	b - 1	CMB = SCB/(b-1)	Fo = CMB/CME
Error experimental	SCE	(a-1)*(b-1)	CME = SCE/(a-1)*(b-1)	
Total	SCT	a*b - 1		

- Si Fo > F ((a-1) ; (a-1)*(b-1)) Rechazo la Hipótesis Nula, por ende se concluye que existe diferencia en la variable respuesta según el tratamiento aplicado.
- Si Fo > F ((b-1) ; (a-a)*(b-1)) Rechazo la Hipótesis Nula, por ende se concluye que existe diferencia en la variable respuesta según el bloque.

DESARROLLO

Variable Respuesta: Tiempo de demora en concretarse la recarga (medido en segundos)
Factor: Empresa proveedora de ventas de saldo
Niveles del Factor: Re Virtual y Tcargo
Bloques: Compañía de celular
Niveles del Bloque: Personal, Claro y Movistar

Diseño de experimento:

$$y_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0; \sigma^2)$$

y_{ij}: valor observado (tiempo)

μ: media general

τ_i: *efecto fijo del tratamiento (empresa)*

ρ_j: *efecto fijo del bloque (compañía)*

ε_{ij}: *error experimental*

condiciones:

*aditividad Tratamiento-Bloque (o sea que no hay interacción)

*para el error experimental: independencia de los datos, homogeneidad de varianzas y normalidad con media igual a 0.

Medición

Se llenará una tabla con los tiempos de demora (medido en segundos) que pasa entre que se aprueba la recarga en la maquina hasta que llega el mensaje de confirmación en el celular.

Valores obtenidos en segundos

		Compañía		
		Personal	Claro	Movistar
Empresa	Tcargo	44,4	1	4,63
		48,02	1,19	2,81
		44,4	1,09	4,77
		44,35	2,09	2,36
		45,01	1,19	1,89
	Revirtual	48,26	2,32	3,71
		47,11	1,93	3,58
		48,74	2,13	2,62
		47,47	1,83	5,63
		47,08	2,22	5,7

Son 5 réplicas para cada tratamiento, se escogió en forma aleatoria (usando el software Excel) el orden en que se efectuaron las recargas dentro de cada bloque. Para el experimento se dispuso de 3 celulares, cada uno con un chip de distinta compañía, se usaron las tres principales de aquí del norte (Personal, Claro y Movistar). Las cargas realizadas fueron todas de $10, este monto fue el escogido ya que es la carga mínima que es compatible para las 3 compañías.

Todas las mediciones se realizaron en una misma tarde. El tiempo de espera entre cada recarga para un mismo celular fue de 15 min (tiempo mínimo exigido por las máquinas para recargar un mismo número dos veces), por lo que se tardó alrededor de 2hs y media en tomar todas las mediciones.

Análisis de varianzas

H_0: $\overline{\mu_1} = \overline{\mu_2} = \cdots = \overline{\mu_a} = 0$

H_a: $\overline{\mu_i} \neq 0$ *para algún i*

α=0,05

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Empresa	1	14,9	14,88	12,08	0,002
Compañia	2	12781,6	6390,79	5188,17	0,000
Error	26	32,0	1,23		
Falta de ajuste	2	4,5	2,24	1,95	0,164
Error puro	24	27,6	1,15		
Total	29	12828,5			

Vp<α ➔Rechazo la Ho para el factor Empresa con lo cual se concluye si hay efecto por parte de ese factor. Los tiempos necesarios para consolidar la recarga son distintos para cada empresa.

Contraste

prueba de contraste simple F0.

α	0,05
r	15
t	2
N	30
CME	1,14799

contraste $C_1 = 1\overline{Y_{Tcargo}} - 1\overline{Y_{ReVirtual}}$

$H_0: C = 0$
$H_a: C \neq 0$

$SCC = \frac{r(\sum_{i=1}^{t} k_i \bar{y}_i)^2}{\sum_{i=1}^{t} k_i^2}$ $F_0 = \frac{CMC}{CME}$ $F_0 \sim F_{\alpha,1,N-t}$

	Coeficientes		Promedio							
contraste	k1	k2	Tcargo	ReVirtual	C	SCC	gl	CMC	F0	Fc
C1	1	1	16,6133333	18,022	-1,40866667	14,88256333	1	14,8825633	12,9640183	4,19597182

Conclusion: F0>Fc por lo que **Rechazo la Ho**, por la estimacion del contraste c1=-1,41<0 se puede suponer que los tiempos de demora son mayores para la empresa "ReVirtual".

Análisis de condiciones:

Prueba de aditividad de Tukey

$$P = \sum_{i}^{a}\sum_{j}^{n} y_{ij} * Y_{i.} * Y_{.j} - Y_{..} * \left(SCA + SCB + \frac{Y_{..}^{2}}{a * b}\right)$$

$$SC(no\ aditividad) = \frac{P^2}{a * b * SCA * SCB}$$

$Y_{i.}$: suma de toda la fila i.
$Y_{.j}$: suma de toda la columna J.
Y...: suma total
SCA y SCB: suma de cuadrados del factor A y B
SC(no aditividad): suma de cuadrados de la no aditividad
a,b: cantidad de niveles del factor A Y b

SCA	14,8825633
SCB	12781,5702
Y..	519,53
a	2
b	3
P	-18688616
SC(no adit)	306013849

Ho: No adit.=0 (Si hay aditiv)
Ha: No hay aditividad

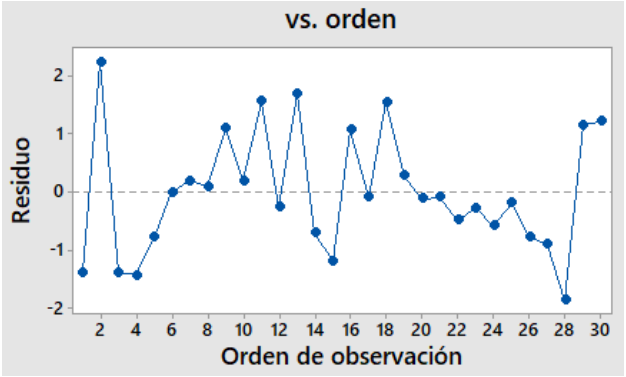
ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Empresa	14,88256333	1	14,88256333	12,96401827	0,001435609	4,259677273
Compañía	12781,57021	2	6390,785103	5566,934471	9,80753E-33	3,402826105
Dentro del grupo	32,02	29	1,14799			
No Aditividad	306013848,5	1	306013848,5	-28,0000029		4,195971819
Residual	-306013816,5	28	-10929064,88			
Total	12828,47954	29				

F<Fc ==> **No Rechazo H0** (si hay aditividad) por lo que se concluye que **no hay interacción**.

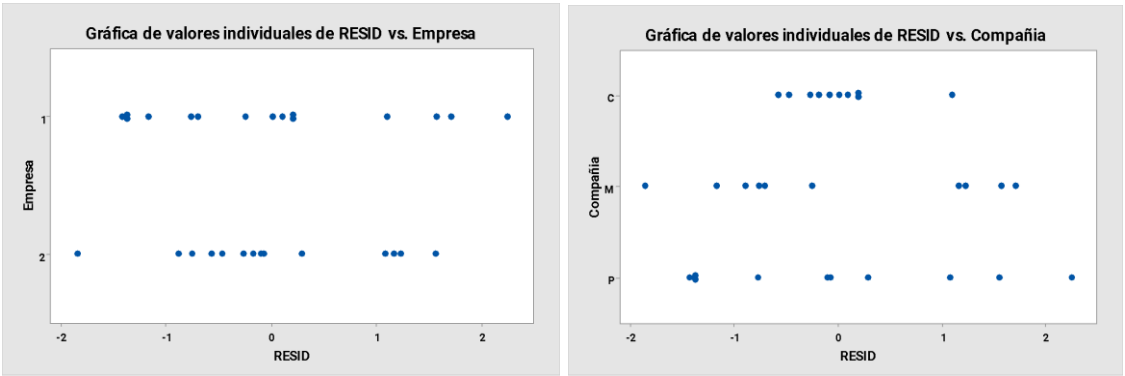
Análisis de residuales

1) La generación de residuos debe tener un comportamiento aleatorio y ser independientes e idénticamente distribuidas.



II Gráficos de análisis de residuos, software Minitab

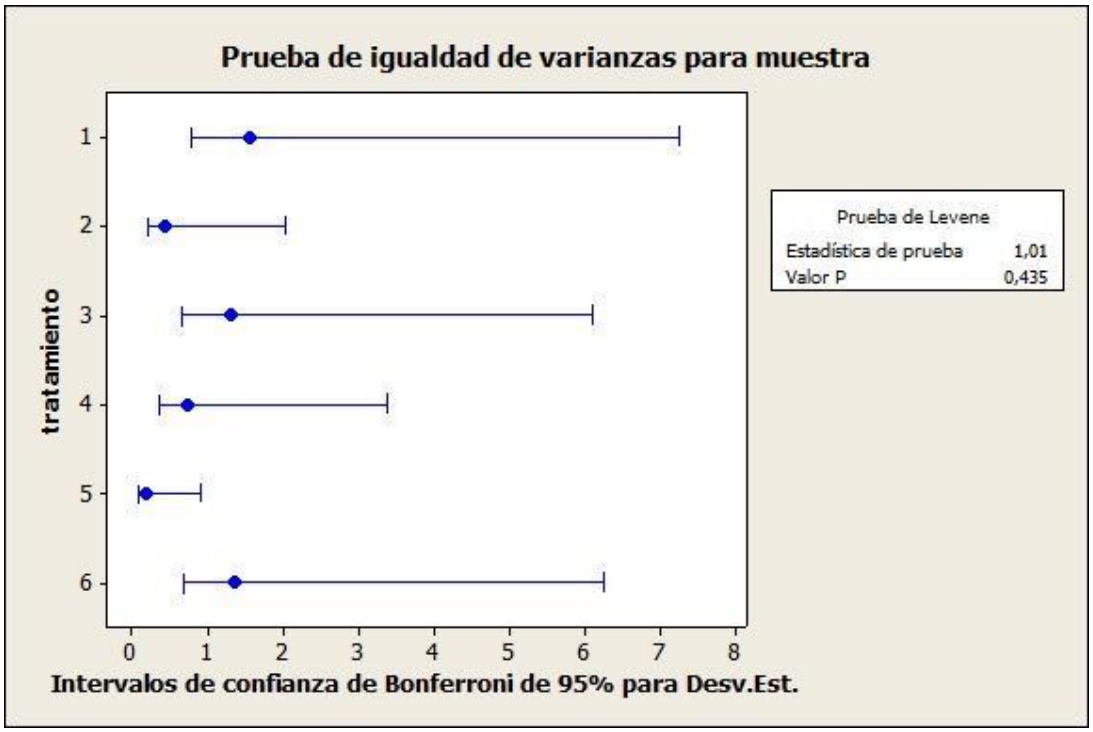
No se observa un patrón definido, por lo que se cumple la aleatoriedad.



III Gráficos de análisis varianzas de residuos, software Minitab

Como se puede ver en los gráficos de distribución de residuales no se puede encontrar un patrón definido (ninguna especie de embudo), se puede concluir que los datos son independientes entre sí y están idénticamente distribuidos.

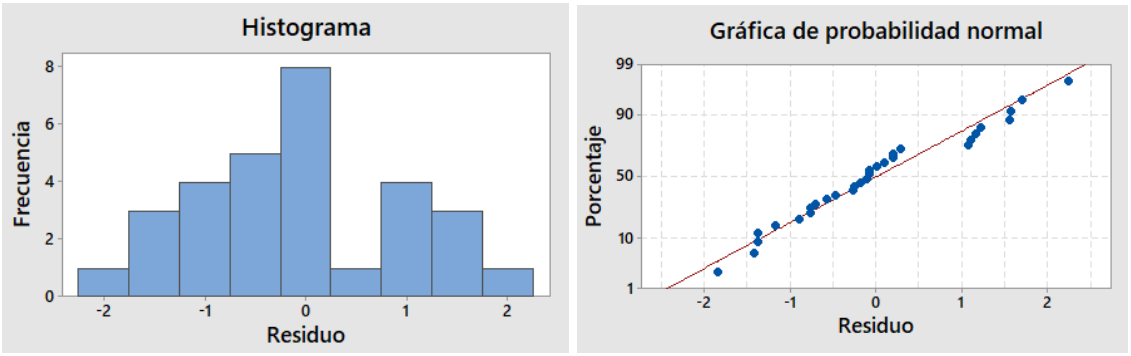
2) Todos los residuales deben tener igual varianza. Esto quiere decir, que la varianza debe ser constante en todo el rango de concentración dinámica de los residuales. Esta propiedad se conoce como homocedasticidad.



IV Prueba de igualdad de varianzas, software Minitab

Como el valor p>0,05 entonces se cumple la Ho: existe homogeneidad entre varianzas.

3) Todos los residuos son variables aleatorias con distribución (aproximadamente) normal con media 0.



V Gráficos de análisis de residuos, software Minitab

Como se ve en el histograma de la izquierda, se puede concluir que los residuos siguen una distribución normal, aunque en casos donde las muestras son pocas los histogramas no son muy confiables. En cuanto al gráfico de la derecha representa una regresión lineal de los residuos respecto a una distribución normal, como ajusta muy bien entonces se cumple la condición de normalidad.

CONCLUSIÓN

Todas las pruebas para el análisis de residuos dieron el visto bueno para el uso del diseño unifactorial con bloques, también la prueba de aditividad, por lo que modelo está bien planteado.

Los resultados obtenidos por el análisis de varianzas indican que en efecto hay diferencias entre los tiempos de demora según la empresa escogida, y que se hizo bien en bloquizar según la compañía telefónica de los números cargados, esto brindó mayor sensibilidad para tener un veredicto. Por último, la prueba de contraste establece que la empresa que brinda

mayor velocidad para las recargas es la llamada “Tcargo” y la de mayor tiempo de demora la llamada “ReVirtual”.

Otra observación interesante es la gran diferencia que hay entre las medias de los tiempos de demora según la compañía, teniendo a la más rápida casi de forma inmediata “Claro” y a la más lenta “Personal”, “Movistar” por su parte fue la que más problemas presentó a la hora de tomar mediciones, en ocasiones la recarga se rechazaba aun cumpliendo con el tiempo exigido por la máquina para realizar dos recargas seguidas a un mismo número, Personal en este aspecto fue la menos problemática. Una hipótesis de estas características podría deberse a que la compañía con mayor cantidad de usuarios por lo menos aquí en el norte es personal, luego claro y por último Movistar, además de que Claro es la compañía que más servicios de internet ofrece, tal vez por eso brinda mayor velocidad.

BIBLIOGRAFÍA

Rodríguez, H. I., Mautino, G. y Mercedes L. Mendez. (2021). Tema 9: Diseños en bloques. https://moodleing.unsa.edu.ar/course/view.php?id=180

Rodríguez, H. I. y Mautino, G. (2021). Tema 2: Estimación - Otros parámetros. https://moodleing.unsa.edu.ar/course/view.php?id=180

Robert O. Kuehl. Principios para el diseño de investigaciones. Comenzando con diseños totalmente aleatorizados. Diseño de experimentos. (2da ed). (2000). (pp. 2-71). THOMSON - LEARN ING.

Wackerly, Dennis D., William Mendenhall III y Richard L. Scheaffer. Estadística matemática con aplicaciones. (7ma ed). (2008). CENGAGE Learning.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Algunos casos aplicados de Test de Hipótesis para una y dos Poblaciones

Gisella Carla Mautino, Iván Rodríguez

Institución: Facultad de Ingeniería, Universidad Nacional de Salta.

Gisella Carla Mautino +5493884440566 gmautino@ing.unsa.edu.ar / gisemautino@gmail.com – *Héctor Iván Rodriguez +5493874129731* ivan@ing.unsa.edu.ar

RESUMEN

Con el objeto de desarrollar casos reales de aplicación de test de hipótesis, los docentes de la Cátedra Probabilidad y Estadística de la Universidad Nacional de Salta realizaron una investigación en una Planta Embotelladora de refrescos de la Provincia de Salta, pudiendo desarrollar modelos de aplicación de test de hipótesis, que contribuyen a la toma de decisiones en las siguientes situaciones:

1. Comprobar el ajuste y puesta a punto de una nueva llenadora para refrescos tamaños 375 ml. Es decir, si el equipo luego de la instalación se encuentra ajustado de manera que llena cada envase con un promedio de 375 ml.
2. Determinar si la proporción de fallas en los repuestos para la cadena de turnelas, del proveedor "X", es la declarada por el fabricante.
3. Analizar el rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor A, con respecto a la producida por el Proveedor B.

Se utilizarán nombres genéricos de los proveedores por resguardo de la identidad de estos.

Para facilitar su lectura, cada modelo desarrollado cuenta con un nombre, objetivos, desarrollo y conclusión.

Palabras Clave: Test de hipótesis. Media poblacional. Varianza poblacional. Poblaciones infinitas. Proporción poblacional.

INTRODUCCIÓN

Con el objeto de desarrollar casos reales de aplicación de test de hipótesis, los docentes de la Cátedra Probabilidad y Estadística de la Universidad Nacional de Salta realizaron una investigación en una Planta Embotelladora de refrescos de la Provincia de Salta, pudiendo desarrollar modelos de aplicación de test de hipótesis, que contribuyen a la toma de decisiones en las siguientes situaciones:

Situación de estudio 1: Al instalarse una nueva máquina llenadora, especializada en la producción de un nuevo tamaño 375 ml. Se desea comprobar que en realidad el equipo no llena los envases exactamente con 375 mililitros, sino que requiere ajustes, con un nivel de significación del 1%.

Situación de estudio 2: Luego de trabajar durante un año con un proveedor local de piezas para la sopladora de bidones "X", nos interesa probar que NO es cierta la aseveración del fabricante, que sostiene que como máximo el 2% de sus repuestos para la cadena de turnelas de la sopladora, presenta fallas antes de las 7000 horas de uso, luego de las cuales se recomienda su reemplazo. Con un nivel de confianza del 1%.

Situación de estudio 3: Se requiere analizar el rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor A, versus la producida por el Proveedor B: se pretende verificar con un nivel de significancia del 5%, si existe diferencia significativa, en los rendimientos de ambos proveedores.

Para facilitar su lectura, cada modelo desarrollado en base a cada situación de estudio cuenta con un nombre, objetivos, desarrollo y conclusión.

METODOLOGÍA

La metodología utilizada consiste en la colección de datos, planteo de las hipótesis, aplicación de la metodología de test de hipótesis apropiada a cada caso, y finalmente el aporte de recomendaciones en función de las conclusiones a las que se arriban con los tests.

DESARROLLO

CASO 1 Test de Hipótesis para la MEDIA POBLACIONAL de POBLACIONES NORMALES, con VARIANZA POBLACIONAL CONOCIDA y Poblaciones Infinitas.

Nombre del Modelo: Ajuste y puesta a punto de nueva llenadora, para refrescos tamaños 375 ml.

Imagen 1 Llenadora de botellas

Objetivos de Modelo

En una Planta Embotelladora de refrescos de la Provincia de Salta, se instala una nueva máquina llenadora, especializada en la producción de un nuevo tamaño 375 ml, en la que se utiliza una preforma de 20.6 grs.

Imagen 2 Preformas de 20.6 grs.

Se conoce por datos históricos de la empresa, que la población es Normal y que la desviación estándar es de 5 mililitros.

Se desea comprobar que en realidad el equipo no llena los envases exactamente con 375 mililitros, sino que requiere ajustes, con un nivel de significación del 1%.

Para ello se toma una muestra aleatoria de 30 refrescos, de una corrida de producción en régimen, y se obtiene que tienen un contenido promedio de 283,25 mililitros. [1]

[1] Aplicación del Teorema de Límite Central. Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2009). Estadística matemática con aplicaciones. *7a*. Col. Cruz Manca, Santa Fe, México, D.F.: Cengage Learning Editores S.A. de C.V.

Imagen 3 Botella de tamaño 375 ml a producir

Desarrollo:

X: Variable Aleatoria Continua, Cantidad de mililitros de refrescos vertidos en los envases de refrescos de tamaño 375 ml.

N → ∞ (tamaño de la población)

n = 30 (tamaño de la muestra)

$\bar{X}$ = 376.68 ml. (contenido promedio en la muestra)

σ = 5 ml. (desviación estándar de la población)

α = 0.01 (nivel de significación)

Tabla 1 datos del contenido promedio de la muestra aleatoria de 30 refrescos

Muestra	Valor obtenido
1	374.1
2	376.2
3	375.1
4	376.2
5	374.1
6	374.1
7	378.2
8	379.1
9	377.2
10	376.1
11	378.2
12	375.1
13	375.2
14	375.2
15	378.1
16	373.2
17	374.1
18	378.2
19	379.1
20	374.2
21	374.2
22	378.1
23	379.2
24	379.5
25	377.1
26	378.1
27	379.2
28	379.5
29	377.1
30	377.5
Promedio	376.68

Aplicamos los 10 pasos para la realización del Test de Hipótesis correspondiente [2]

1. Parámetro para probar: Media Poblacional μ
2. Los cursos de acción son:
   - Si μ = 375 ml → El equipo está listo para producir y no necesita ajustes
   - Si μ >375 m → El equipo necesita ajustes
   - Si μ < 375 m → El equipo necesita ajustes
3. En este caso no hay desigualdad EQUIVALENTE a la igualdad, pues ninguna conduce al mismo curso de acción
4. Se proponen las hipótesis:

$$H_0: \mu = 375$$

$$H_1: \mu \neq 375$$

5. Como la población es infinita, tiene distribución normal y se conoce la varianza poblacional, el estadígrafo de prueba será:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

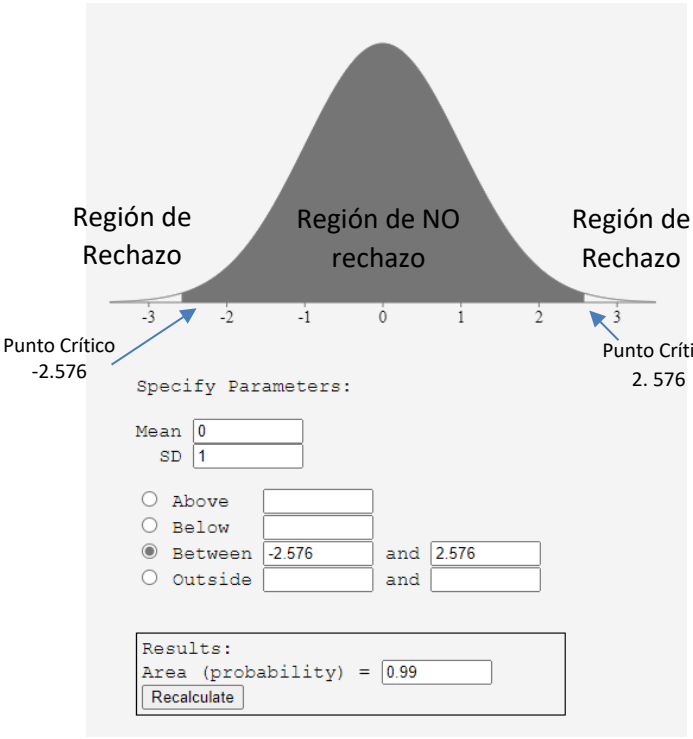
6. Como el estadígrafo de prueba sigue la Ley Normal, la Región Crítica pertenece al dominio de dicha Distribución. Como la desigualdad no equivalente es la desigualdad ≠ la Región Crítica está ubicada a la IZQUIERDA Y A LA DERECHA del Punto Crítico que es el fractil α = 0.01 → $\frac{\alpha}{2}$ = 0.005 → Z_c = 2.576

   (1 - α) = 0.99
7. La Regla de decisión estadística para rechazar la hipótesis nula es:

   Si $\left|Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| \geq 2.576 \Rightarrow$ Se rechaza H_0

[2] Metodología y notación de Walpole, R. E. (2012). Probabilidad y estadística para ingeniería y ciencias. México: Pearson Education.

Imagen 4 comprobación con calculadora on line Statbook
http://onlinestatbook.com/2/calculators/normal_dist.html



8. El Valor numérico del estadígrafo de Prueba es:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{376.68 - 375}{\frac{5}{\sqrt{30}}}$$

$$Z = 1.84$$

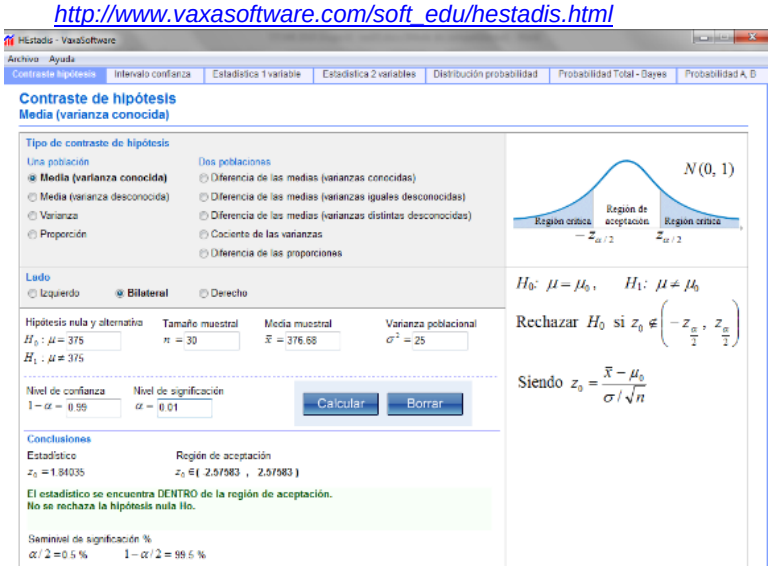
$$Z = 1.84 < 2.576$$

9. Como el valor del estadígrafo de prueba, no pertenece a la región crítica, la decisión estadística es de **no rechazar** la hipótesis nula.

   Es decir, no existe evidencia suficiente para rechazar H_0, con un nivel de riesgo del 0.01

10. La acción derivada es que el equipo no requiere ajustes, estaría listo para producir.

Imagen 5 Comprobación con la calculadora Vaxa Software
http://www.vaxasoftware.com/soft_edu/hestadis.html



Conclusión:

Al no existir evidencia suficiente para rechazar H_0, que plantea que el equipo luego de la instalación se encuentra ajustado de manera que llena cada envase con un promedio de 375 ml.

CASO 2 - Test de Hipótesis para la PROPORCIÓN POBLACIONAL - Poblaciones Infinitas.

Nombre del Modelo: proporción de fallas en los repuestos para la cadena de turnelas, del proveedor “X”

Imagen 6 Sopladora de Bidones PET

Objetivos de Modelo

Luego de trabajar durante un año con un proveedor local de piezas para la sopladora de bidones “X”, nos interesa probar que NO es cierta la aseveración del fabricante que sostiene

que como máximo el 2% de sus repuestos para la cadena de turnelas de la sopladora, presenta fallas antes de las 7000 horas de uso, luego de las cuales se recomienda su reemplazo. Con un nivel de confianza del 1%.

Para ello solicitamos al sector Mantenimiento el historial de fallas en los repuestos para la cadena de turnelas del proveedor "X", en una muestra aleatoria de 500 piezas, tomadas del listado histórico de fallas del equipo, obtenemos que 20 piezas tuvieron que ser reemplazadas por fallas antes de las 7000 horas de funcionamiento.

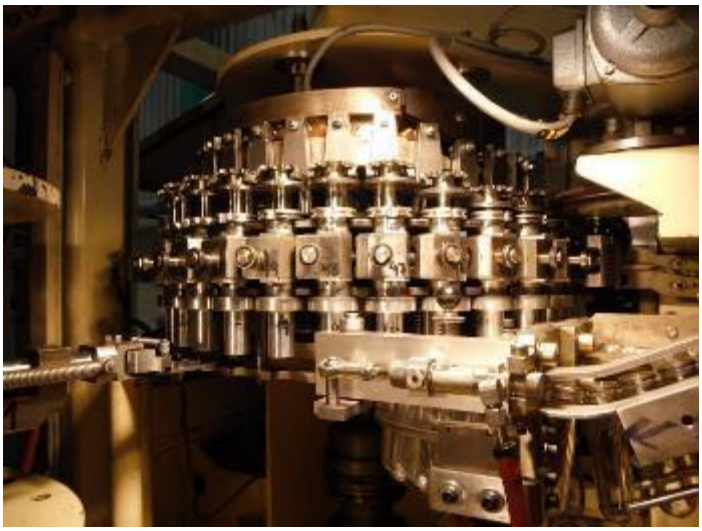


Imagen 7 Cadena de turnelas

Desarrollo

X: Variable aleatoria continua. Proporción de cadenas de turnelas que tuvieron que ser reemplazadas por fallas antes de las 7000 horas de funcionamiento.

N → ∞ (tamaño de la población)

n = 500 (tamaño de la muestra)

r = 20 piezas con fallas detectadas en la muestra

α = 0.01

$\bar{p} = \frac{r}{n} = \frac{20}{500} = 0.04$ proporción de fallas en la muestra

$\bar{q} = 1 - \bar{p} = 0.96$ proporción de piezas que no fallan en la muestra

Aplicamos los 10 pasos para la realización del Test de Hipótesis correspondiente [3]

1. Parámetro para probar: Proporción Poblacional **Π**
2. Los cursos de acción son:
   - Si Π = 0.02 → Se comprueba la aseveración del fabricante

[3] Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). Estadística para administración y Economía. *10a.* Col. Cruz Manca, Santa Fe, D.F., México: Cencage Learning.

- Si Π < 0.02 → Se comprueba la aseveración del fabricante
- Si Π > 0.02 → No se comprueba la aseveración del fabricante

3. En este caso la desigualdad < es EQUIVALENTE a la igualdad, pues ambas conducen al mismo curso de acción
4. Se proponen las hipótesis: $\mathbf{H_0: \Pi \leq 0.02}$

$$\mathbf{H_1: \Pi > 0.02}$$

5. El estadígrafo de prueba será:

$$Z = \frac{\bar{p} - \Pi}{\sqrt{\frac{\Pi * (1 - \Pi)}{n}}}$$

6. Como el estadígrafo de prueba sigue la Ley Normal, la Región Crítica pertenece al dominio de dicha Distribución. Como la desigualdad no equivalente es de ">" la Región Crítica está ubicada a la DERECHA del Punto Crítico que es el fractil (1- α) = 0.99 → α = 0.01 → Z_c = 2.326

7. La Regla de decisión estadística para rechazar la hipótesis nula es:

Si $Z = \frac{\bar{p} - \Pi}{\sqrt{\frac{\Pi * (1 - \Pi)}{n}}} \geq 2.326 \Rightarrow$ Se rechaza H_0

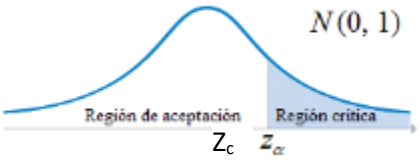


Imagen 8 Ilustración de las regiones de aceptación y rechazo para un test de hipótesis Normal

8. El Valor numérico del estadígrafo de Prueba es:

$$Z = \frac{\bar{p} - \Pi}{\sqrt{\frac{\Pi * (1 - \Pi)}{n}}}$$

$$Z = \frac{0.04 - 0.02}{\sqrt{\frac{0.02 * 0.98}{500}}}$$

$$Z = 3.19$$

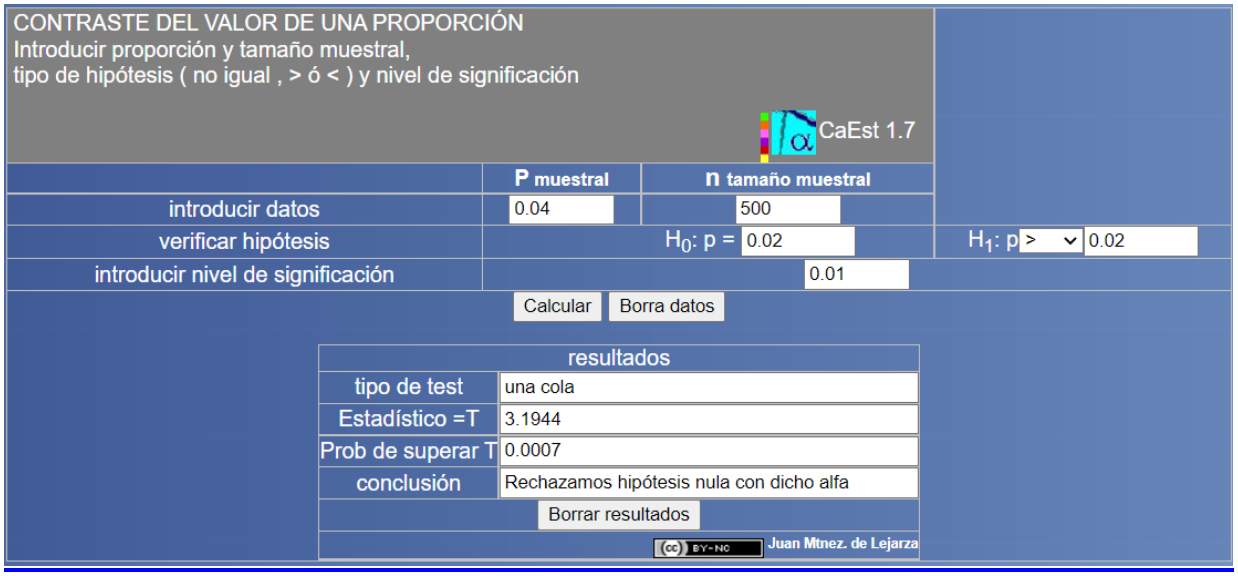
$$Z = 3.19 > 2.326$$

9. Como el valor del estadígrafo de prueba, pertenece a la región crítica, la decisión estadística es **rechazar** la hipótesis nula.

Es decir, existe evidencia suficiente para RECHAZAR H_0, en favor de H_1 con un nivel de riesgo del 1%.

10. La acción derivada resulta en que no se comprueba la aseveración del fabricante, que como máximo el 2% de sus piezas presentan fallas antes de las 7000 horas. Por ello, sería conveniente evaluar el desempeño de otros proveedores más confiables.

Imagen 9 Comprobación con la calculadora CaEst 1.7
https://www.uv.es/ceaces/scrips/tablas/c1prop.htm



Conclusión

Al rechazar H_0, efectivamente no se comprueba la aseveración del fabricante de que como máximo el 2% de sus piezas presentan fallas antes de las 7000 horas. Por ello, sería conveniente evaluar el desempeño de otros proveedores más confiables, cuyas piezas efectivamente duren las 7000 horas de funcionamiento.

CASO 3: Test de Hipótesis para COMPARAR las MEDIAS POBLACIONALES de DOS POBLACIONES NORMALES - Varianzas Poblacionales Desconocidas pero IGUALES.

Nombre del Modelo: Análisis el rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Ingenio A, versus la producida por el Ingenio B

Imagen 10 Big Bags de azúcar Proveedor A y Proveedor B

Imagen 11 Equipos de mezclado de azúcar y agua para obtener jarabe simple

Objetivos de Modelo

Se desea analizar el rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor A, versus la producida por el Proveedor B.
Según un estudio realizado hace 5 años ambos rendimientos eran similares. Se desea comprobar si hay cambios en la actualidad, ya que el precio de adquisición de ambos, es diferente.

El muestreo se realiza al azar, teniendo en cuenta los procedimientos de la Empresa, basados en la Norma IRAM 15 (Sistemas de muestreo para la inspección por atributos)[4], y los AQL (niveles de calidad aceptables), definidos para la aceptabilidad de un lote. En base a los

[4] Instituto Nacional de Racionalización de Materiales. (2010). Normas IRAM 15. Inspección por atributos, planes de muestra única, doble y múltiple, con rechazo. Argentina: Instituto Nacional de Racionalización de Materiales.

mismos, y al historial de los dos proveedores, se determina tomar al azar 25 muestras del proveedor A y 30 muestras del proveedor B.

Luego de analizar los 25 bolsones del proveedor A, se obtuvo un rendimiento promedio de 2400 litros de jarabe, con un desvío estándar de 14,5 litros. Para el proveedor B, se analizaron los 30 bolsones, obteniendo un promedio de 2340 litros, con un desvío estándar de 15,9 litros. Se sabe en base al historial de la empresa, que el proceso de elaboración de jarabe sigue una Ley Normal, y queremos verificar con un nivel de significancia del 5%, si existe diferencia significativa en los rendimientos de ambos proveedores.

Desarrollo

X_1: rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor A
X_2: rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor B
$X_1 \backsim N(\mu_1 ; \sigma_1)$; $X_2 \backsim N(\mu_2 ; \sigma_2)$; ambas variables normales
$n_1 = 25$ tamaño de la muestra del proveedor A
$n_2 = 30$ tamaño de la muestra del proveedor B

$\overline{X}_1 = 2400$ litros. Rendimiento promedio del proveedor A
$\overline{X}_2 = 2340$ litros. Rendimiento promedio del proveedor B
$S_1 = 14,5$ desviación estándar del proveedor A
$S_2 = 15,9$ desviación estándar del proveedor B
$\alpha = 0.05$ nivel de significancia

Aplicamos los 10 pasos para la realización del TH correspondiente [5]

1. Parámetro por probar: diferencia de 2 medias poblacionales ($\mu_1 - \mu_2$)
2. Los cursos de acción son:
   - Si ($\mu_1 - \mu_2$) = 0 → Los rendimientos promedios de ambos proveedores son iguales
   - Si ($\mu_1 - \mu_2$) > 0 → Los rendimientos promedios de ambos proveedores NO son iguales
   - Si ($\mu_1 - \mu_2$) < 0 → Los rendimientos promedios de ambos proveedores NO son iguales
3. En este caso No Hay DESIGUALDAD EQUIVALENTE a la igualdad, pues ninguna de las 2 desigualdades conduce al mismo curso de acción
4. Se proponen las hipótesis:

$$\mathbf{H_0: Si\ (\mu_1 - \mu_2) = 0}$$

[5] Walpole, R. E. (2012). Probabilidad y estadística para ingeniería y ciencias. México: Pearson Education.

H_1: Si ($\mu_1 - \mu_2$) ≠ O

5. Como las poblaciones se suponen infinitas, tienen una distribución normal y no se conocen las VAR Poblacionales, se verifica si las mismas son iguales o no, mediante el caso 9:

Aplicando los 10 pasos para la realización de TH del CASO 9:

9.1 Parámetro por probar: COCIENTE entre las 2 VARIANZAS POBLACIONALES $\frac{\sigma_1^2}{\sigma_2^2}$

9.2 Los cursos de acción son:

- Si $\frac{\sigma_1^2}{\sigma_2^2}$ = 1 Ambas varianzas son iguales
- Si $\frac{\sigma_1^2}{\sigma_2^2}$ > 1 No se puede considerar que ambas varianzas son iguales
- Si $\frac{\sigma_1^2}{\sigma_2^2}$ < 1 No se puede considerar que ambas varianzas son iguales

9.3 No existe desigualdad EQUIVALENTE a la igualdad, pues ninguna de ellas conduce al mismo curso de acción que la igualdad. Luego se trata de una Prueba BILATERAL.

9.4 Se proponen las HIPÓTESIS: **H_0**: $\frac{\sigma_1^2}{\sigma_2^2}$ = **1**

H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq$ **1**

9.5 Como las poblaciones se suponen infinitas y tienen distribución Normal, luego el Estadígrafo de Prueba es:

$$F_{(v1 = n1 - 1;\ v2 = n2 - 1)} \sim \frac{S_1^2}{S_2^2} * \frac{\sigma_1^2}{\sigma_2^2}$$

9.6 Como el estadígrafo de Prueba sigue la distribución F de Fisher/Snedecor, con (n1 - 1) = 24 grados de libertad en el numerador y (n2 – 1) = 29 grados de libertad en el denominador, la Región Crítica pertenece al Dominio de dicha distribución, y como NO hay desigualdad equivalente, la región crítica se subdivide en 2 áreas de igual tamaño. El punto crítico de la IZQUIERDA es el fractil: $\left(\frac{\alpha}{2}\right)$ = 0.025, y el punto crítico de la DERECHA, es el fractil: 1 – $\left(\frac{\alpha}{2}\right)$ = 0.975 → F_{C1} = 0.451 y F_{C2} = 2.154

Imagen 12 Comprobación mediante calculadora CaEst 1.7
https://www.uv.es/ceaces/scrips/tablas/taf.htm

9.7 La regla de Decisión Estadística para RECHAZAR la hipótesis nula es:

Si $\frac{S_1^2}{S_2^2} * \frac{\sigma_1^2}{\sigma_2^2} < 0.451$ o si $\frac{S_1^2}{S_2^2} * \frac{\sigma_1^2}{\sigma_2^2} > 2.154$ → **Se Rechaza H_0**

9.8 El valor numérico del estadígrafo de prueba es:

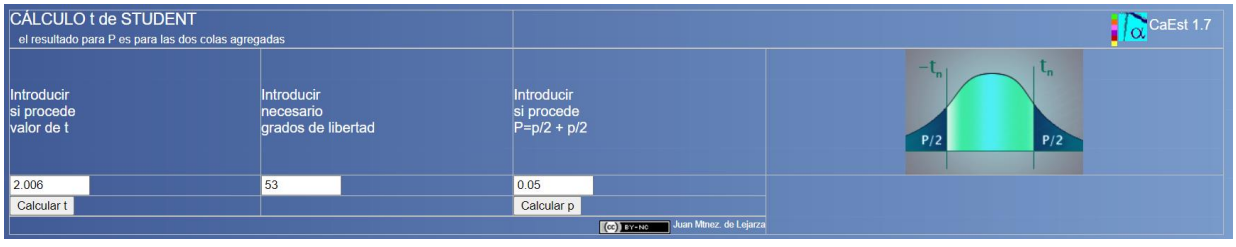
$\frac{14.5^{\,2}}{15.9^{\,2}} * 1 = 0.8317$ Como 0.451 < 0.8317< 2.154 → no pertenece a R_C

9.9 Como el valor del Estadígrafo de prueba, no pertenece a la Región Crítica, la Decisión Estadística es NO rechazar la Hipótesis Nula, con un nivel de confianza del 2%.

9.10 Acción derivada de la decisión estadística: se puede considerar que las varianzas de ambos proveedores A y B, son prácticamente iguales.

6. Como el estadígrafo de prueba sigue la Distribución t de Student, con ($n_1 + n_2 - 2$) = 53 grados de libertad, la Región Crítica pertenece al dominio de dicha Distribución. Como NO hay desigualdad, la Región Crítica se compone dos áreas de igual tamaño, una hacia la IZQUIERDA, cuyo punto crítico es el fractil $\left(\frac{\alpha}{2}\right) = 0.025$ → t_{c1} = -2.006; y la otra hacia la DERECHA del Punto Crítico que es el fractil $\left(1 - \frac{\alpha}{2}\right)$ = 0.975 → t_{c2} = 2.006

Imagen 13 Comprobación mediante calculadora CaEst 1.7
https://www.uv.es/ceaces/scrips/tablas/tastud.htm



7. La Regla de decisión estadística para rechazar la hipótesis nula es:

Si $\left|\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_a^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right| \geq 2.006 \Rightarrow$ Se rechaza H_0

8. El estadígrafo de Prueba es:

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_a^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \backsim t_{(n1+n2-2)\ gl}$$

Calculamos el valor de la Varianza Amalgamada:

$$S_a^2 = \frac{(n_1 - 1) * S_1^2 + (n_2 - 1) * S_2^2}{n_1 + n_2 - 2}$$

$$S_a^2 = \frac{24 * 14.5^2 + 29 * 15.9^2}{25 + 20 - 2} = 233.54$$

Reemplazamos en la expresión de t:

$$t = \frac{(2400 - 2340) - 0}{\sqrt{233.54\left(\frac{1}{25} + \frac{1}{30}\right)}} = 14.4984$$

$$t = 14.4984 > 2.006 \Rightarrow \in R_c$$

11. Como el valor del estadígrafo de prueba, pertenece a la región crítica, la decisión estadística es **rechazar** la hipótesis nula. Es decir, existe evidencia suficiente para RECHAZAR H_0, en favor de H_1 con un nivel de riesgo del 5%.
12. La acción derivada resulta en considerar que SI hubo cambios en el rendimiento promedio de jarabe por Big-Bag de azúcar producida por el Proveedor A, versus la producida por el Proveedor B.

Conclusión del modelo:

Al rechazar H_0, se comprueba que los rendimientos promedios de jarabe por Big-Bag de azúcar producida por el Proveedor A, versus la producida por el Proveedor B, son diferentes. Por el valor observado en el estadístico, esto sugiere un rendimiento promedio mayor del Proveedor A. Por ello, se deberá estudiar en mayor detalle con cuál de los dos proveedores conviene trabajar, teniendo en cuenta factores como la calidad y precio del azúcar.

CONCLUSIONES

Mediante la aplicación de test de hipótesis en los 3 modelos generados, hemos logrado en cada caso, enunciar conclusiones y sugerencias a la empresa, enunciadas en cada modelo. Resumiendo:

1. Ajuste y puesta a punto de nueva llenadora, para refrescos tamaños 375 ml: se instala una nueva máquina llenadora, especializada en la producción de un nuevo tamaño 375 ml. Se desea comprobar que en realidad el equipo no llena los envases exactamente con 375 mililitros, sino que requiere ajustes, con un nivel de significación del 1%.

Luego de aplicar el test de hipótesis, al no rechazar **H_0: μ = 375,** no existe evidencia suficiente para concluir que el equipo no está ajustado, de manera que llena cada envase con un promedio de 375 ml.

2. Al rechazar **H_0: $\Pi \leq 0.02$**, efectivamente no se comprueba la aseveración del fabricante de que como máximo el 2% de sus piezas presentan fallas antes de las 7000 horas. Por ello, sería conveniente evaluar el desempeño de otros proveedores más confiables, cuyas piezas efectivamente duren las 7000 horas de funcionamiento.
3. Al rechazar **H_0: $\mu_1 - \mu_2 = 0$**, se comprueba que los rendimientos promedios de jarabe por Big-Bag de azúcar producida por el Proveedor A, versus la producida por el Proveedor B, son diferentes. El valor observado en el estadístico sugiere un rendimiento promedio mayor para el Proveedor A. Por ello, si bien el rendimiento del proveedor A es mayor, se podría complementar el estudio analizando otros factores como calidad, tiempos de entrega y precio del azúcar.

BIBLIOGRAFIA

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). Estadística para administración y Economía. *10a.* Col. Cruz Manca, Santa Fe, D.F., México: Cencage Learning.

Instituto Nacional de Racionalización de Materiales. (1973). Normas IRAM 15. Inspección por atributos, planes de muestra única, doble y múltiple, con rechazo. Argentina: Instituto Nacional de Racionalización de Materiales.

Mendenhall, W. B. (2010). *Introducción a la probabilidad y estadística.* Santa Fe, D.F., México: Cencage Learning.

Software, V. (s.f.). Versión 1.9.8. Recuperado el 2021, de HEST - Software de Matemáticas, Herramientas de Estadística y Probabilidad: http://www.vaxasoftware.com/soft_edu/hestadis.html

Valencia, U. d. (s.f.). *Proyectos CEACES* , 1.7. (J. MARTÍNEZ DE LEJARZA ESPARDUCER, Productor, & Universidad de Valencia) Recuperado el 2021, de Contenedor Hipermedia de Estadística Aplicada a las Ciencias Económicas y Sociales: https://www.uv.es/ceaces/index.htm

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2009). Estadística matemática con aplicaciones. *7a.* Col. Cruz Manca, Santa Fe, México, D.F.: Cengage Learning Editores S.A. de C.V.

Walpole, R. E. (2012). *Probabilidad y estadística para ingeniería y ciencias.* México: Pearson Education.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Análisis de la producción vitivinícola de variedad Malbec en Cafayate

Daruich Aguilar, Santiago Agustín, Coronado, Jasmín Anabel.

Universidad Nacional de Salta- Facultad de Ingeniería

santi.daruich27@gmail.com - +543868412851
jaaz.coronado@gmail.com - +543875864464

RESUMEN

En Cafayate la vinificación es una de las actividades económicas más importantes. Para las bodegas es fundamental poder adaptarse a un mercado con gran competencia. Por lo que deben predecir la cantidad de producción en función a la materia prima.

El objetivo de este trabajo es la realización de un modelo lineal que relacione los kilogramos de uva con los litros de vinos producidos. Se trabajó con datos obtenidos de una bodega ubicada en la localidad de Cafayate. En primer lugar, se comprobaron los supuestos necesarios para realizar una regresión lineal, los cuales se cumplieron. Luego, se calculó el coeficiente de correlación muestral y se realizó un diagrama de dispersión, obteniendo una marcada relación lineal entre los litros de vinos obtenidos y los kilogramos de uva. Posteriormente, se adoptaron dos modelos lineales, se validó y eligió el más adecuado.

Palabras Clave: Malbec, Vinificación, Regresión Lineal.

INTRODUCCIÓN

En la provincia de Salta, Argentina, principalmente en el departamento de Cafayate, la elaboración del vino es una de las actividades económicas más importantes. Esta es una región que naturalmente ofrece las condiciones ideales para el crecimiento de viñedos excepcionales debido a su clima único de días soleados, baja humedad y amplitud térmica entre el día y la noche por la altura. El proceso de producción del vino varía enormemente de acuerdo a la bodega, la calidad y la variedad del vino. Nuestra bodega se dedica a la elaboración de pequeñas cantidades de vinos y, por lo tanto, limita su producción a 300.000 botellas. Como todos los grandes productores de vino, la familia cree que la calidad comienza en los viñedos, por lo que mantiene un metódico control y cuidado en todo el proceso.

El Malbec es su variedad insignia y la que mejor representa el paladar local: desde el 2011 es la cepa más cultivada en el país, y se ha posicionado como líder en volumen, calidad y exportaciones a nivel mundial.

El modelo lineal, que se quiere plantear en este trabajo, tiene una gran importancia en pequeñas y grandes bodegas. Será de gran utilidad para que un enólogo tenga una idea sobre el nivel de producción. También, le permitirá saber al gerente cuánto producto puede llegar a ofrecer. En base a eso, poder modificar los precios de acuerdo la oferta y demanda, manteniendo o ganando clientes.

METODOLOGÍA

Al momento de realizar el estudio, nos dirigimos a una bodega en donde entrevistamos a una Enóloga sobre la producción. Nos brindó los kilogramos de uva y litros de vino obtenidos en la vendimia del año 2021, con los cuales pudimos realizar el trabajo.

El proceso de la elaboración del vino comienza con la recolección de la uva. Es muy importante el estado sanitario de la uva para evitar fermentaciones, que originan aromas y gustos no deseados.
El transporte a la bodega se hace en remolques o en cajas en el menor tiempo posible. El remolque debe ser de acero inoxidable, estar protegido con una lona y provisto de un doble fondo para evitar la maceración del mosto. Las cajas se transportan en los remolques y se descargan manualmente en las bodegas. Antes de descargar, se pesa el vehículo en una báscula para saber el peso de la uva que entra.
Las uvas se descargan en una tolva de acero inoxidable provisto en el fondo un tornillo sinfín, que las conduce a la despalilladora.
Posteriormente, se realiza el prensado que se debe realizar en el menor tiempo posible para reducir la incorporación de aire. Por último, el mosto obtenido, se vuelca en grandes depósitos de madera o acero inoxidable, donde fermentará.

DESARROLLO

Análisis de los datos

Para llevar a cabo el desarrollo de este trabajo, compilamos datos con respecto a los kilogramos de uva y litros de vino Malbec, los cuales fueron obtenidos de una bodega ubicada en el departamento de Cafayate, (su nombre no se revela por razones de confidencialidad). Estos provienen del periodo de febrero a marzo cuando transcurre la vendimia, esta consiste en la recogida de la uva para la posterior producción del vino. Se trata de un momento clave dentro del proceso de elaboración de cualquier tipo de vino, ya que durante este tiempo se toman una serie de decisiones que son cruciales para determinar las características, calidad y cantidades del vino que se va a producir, debido a esto consideramos la importancia del análisis de dichos datos.

Tabla 1. Datos proporcionados por la bodega

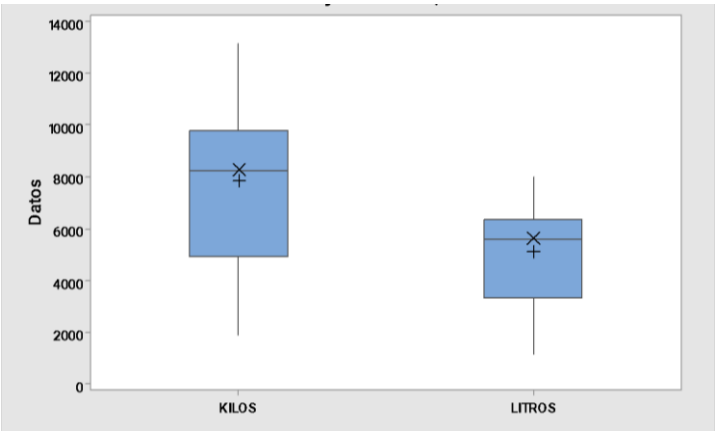
FECHA	Cantidad de uva (Kg)	Cantidad de vino obtenido (l)
02/02/2021	8105	6300
03/02/2021	4376	3200
03/02/2021	4075	3400
04/02/2021	7545	5250
05/02/2021	8418	5250
12/02/2021	9869	6500
17/02/2021	8970	6350
22/02/2021	5121	3000
01/03/2021	9735	6000
02/03/2021	9630	6000
03/03/2021	13155	8000
04/03/2021	11025	6500
06/03/2021	1900	1200
08/03/2021	7670	4500

Antes de empezar con el procesamiento de datos y el respectivo estudio, se realizó un diagrama de caja y bigotes. Este nos permite representar de forma gráfica la distribución de los valores de la variable. Además, nos permite visualizar los valores atípicos o casos extremos de la variable, en el caso de existir alguno será eliminado. De esta manera, se podrá evitar la distorsión de las respuestas obtenidas.

La caja representa la mitad central de los datos. La barra vertical que la fracciona en dos partes corresponde a la mediana, y el signo + señala el promedio de los datos (media muestral).

Los bigotes abarcan los datos que apartan como máximo 1,5 veces la diferencia entre el tercer cuartil y el primero (Rango Intercuartílico). Si algún dato supera estos límites se considera sospechoso.

Figura 1. Diagrama de caja y bigotes



Información del diagrama

- El rango intercuartílico, en el caso de los kilos de uva: Q_3-Q_1= 4833,75; es decir, el 50% de los datos se encuentra comprendido en ese valor.
- El rango intercuartílico, en el caso de los litros de vino Malbec: Q_3-Q_1= 3037,5; es decir, el 50% de los datos está comprendido en ese valor.
- Como la caja correspondiente a los kilos es más ancha que la correspondiente a los litros, indica que los datos de los kilos están más dispersos que los de litros. Como los bigotes del primer diagrama son más largos, tienen mayor variabilidad.

- Como la media está próxima a la mediana, ambas partes de la caja son aproximadamente iguales.
- No se encontró ningún valor atípico. Por lo tanto, no se elimina ningún dato.

Consecuentemente, se prosiguió con el análisis estadístico de los datos. Para ello se optó por utilizar el software Minitab, el cual arrojo una tabla con los siguientes resultados:

Tabla 2. Estadísticos descriptivos

Variable	Media	Desv.Est.	Varianza	Mínimo	Q1	Mediana	Q3	Máximo	Asimetría	Curtosis
KILOS	7828	3038	9227051	1900	4935	8262	9769	13155	-0,37	-0,13
LITROS	5104	1829	3344409	1200	3350	5625	6388	8000	-0,68	0,07

Los valores de asimetría y de curtosis nos informan sobre la forma de la distribución de nuestras variables. La asimetría es la medida que nos permite conocer la distribución de los datos con respecto a la media, mientras que la curtosis es una medida de forma que mide la mayor o menor concentración de los datos alrededor de la media.

Figura 2. Histograma de kilos de uva

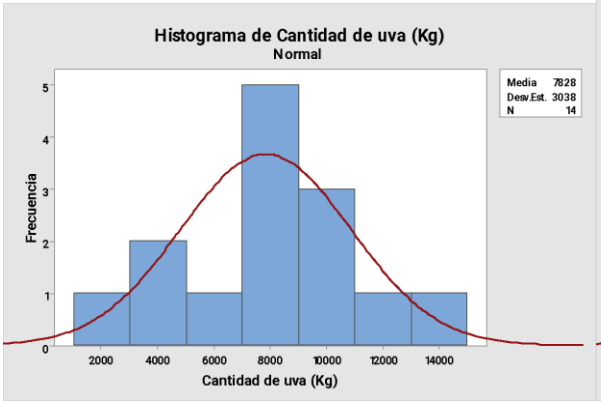
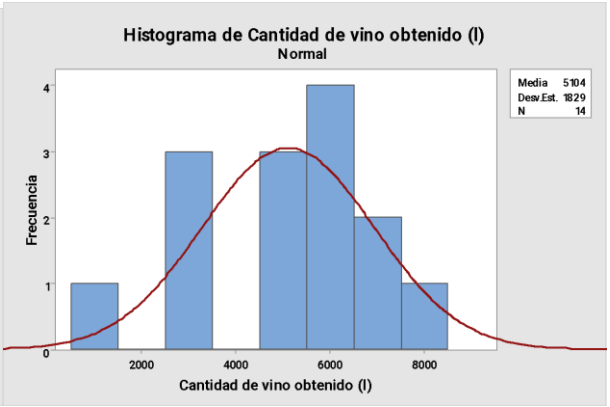


Figura 3. Histograma de litros de vino



En nuestro caso, para las dos variables, el valor de asimetría es negativo. Por lo tanto, las distribuciones tienen un sesgo a la izquierda, es decir que contiene pocas mediciones anormalmente pequeñas.

Para el caso del valor de curtosis de la variable kilos, al ser un valor negativo, nos indica una distribución platicúrtica; es decir, que en las colas de la distribución hay más variables acumuladas.

El valor de curtosis de la variable litros, al ser un valor positivo, nos indica una distribución leptocúrtica; es decir, que existen más datos acumulados en torno al valor de la media.

Como los resultados de asimetría y curtosis no son significativamente grandes, se aproximan mucho al valor cero, se podría realizar una suposición de normalidad para las dos variables.

Análisis de normalidad

Para poder confirmar el supuesto de normalidad se realizó el contraste de Shapiro y Wilks. Plantea la hipótesis nula que una muestra proviene de una distribución normal. Elegimos un nivel de significancia del 0,05 y tenemos una hipótesis alternativa que sostiene que la distribución no es normal. Se rechazará la normalidad cuando el valor calculado sea menor que el valor crítico dado en las tablas.

En el cuadro siguiente se muestran los valores obtenidos:

Tabla 3. Contraste de Shapiro y Wilk

VARIABLES	ESTADÍSTICO CALCULADO (w_0)	VALOR DE TABLA (w_α)
Kilos	0,968527354	0,874
Litros	0,93699786	0,874

Como el valor de los estadísticos calculados son mayores que el valor de tabla, se dice que no tenemos las pruebas suficientes para rechazar la hipótesis nula, por lo que aceptamos la hipótesis de normalidad. También se lo puede apreciar gráficamente:

Figura 4. Gráfica de probabilidad normal de KILOS

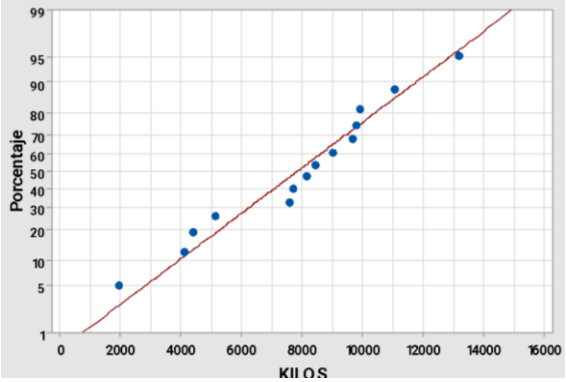
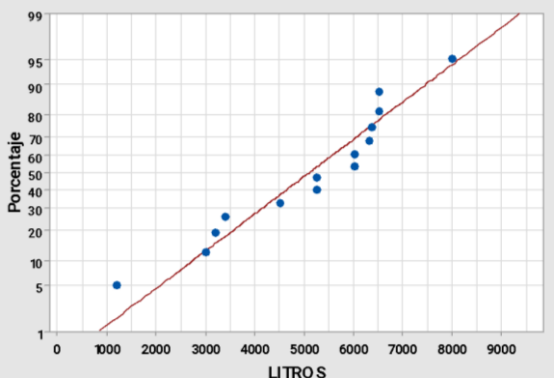


Figura 5. Gráfica de probabilidad normal de LITROS



Análisis de independencia

Uno de los supuestos que se deben cumplir al momento de realizar una regresión lineal es que las observaciones de la muestra sean independientes.

Cuando las observaciones son dependientes, las varianzas de los estimadores son erróneas y los resultados obtenidos no tienen validez.

Utilizamos un contraste de rachas, el cual nos permite verificar que la muestra es aleatoria, es decir, que las sucesivas observaciones son independientes. La hipótesis nula afirma que las observaciones son independientes, mientras que la hipótesis alternativa afirma que las observaciones están relacionadas. Con un nivel de significación del 0,05; los resultados obtenidos se muestran en el siguiente cuadro:

Tabla 4. Contraste de rachas

VARIABLES	k	r	Percentil 2,5	Percentil 97,5
Kilos	7	5	3	12
Litros	7	6	3	12

Como el valor observado r para las dos variables está contenido en el intervalo de aceptación, no hay evidencia suficiente en los datos para rechazar la hipótesis de independencia. En consecuencia, no se encontró evidencia para suponer que las observaciones están relacionadas.

Coeficiente de correlación

Se comienza el análisis de los datos calculando el coeficiente de correlación, el cual es una medida que permite conocer la fuerza y el sentido de la relación lineal entre 2 variables cuantitativas. Entre más cercano es a 1 es más fuerte, entre más cercano a 0 débil, hasta

llegar a hacerse nula. Si los valores del coeficiente de relación son -1 es una asociación lineal perfecta negativa, si es 0 no existe relación y si es 1 es una asociación lineal perfecta positiva.

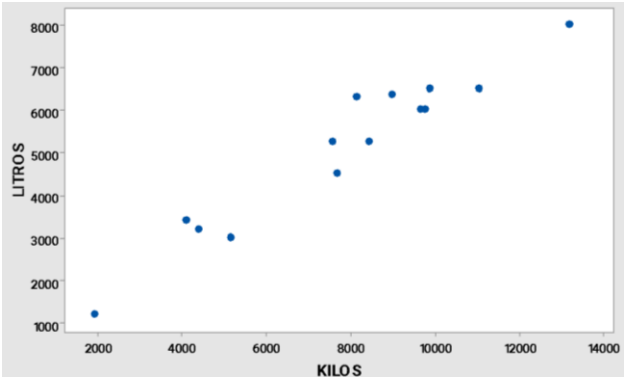
En la tabla siguiente se muestra los resultados entregados por Minitab, en donde se puede observar el coeficiente de correlación. Debajo del coeficiente figura el Valor P de la prueba, que establece que la relación encontrada es estadísticamente significativa.

Tabla 5. Correlación entre las variables

Correlación de Pearson	0,966
Valor p	0,000

Con los datos originales se realizó un diagrama de dispersión, el cual indica que existe una relación entre los kilogramos de uva y los litros obtenidos, y parece ser razonable la consideración tentativa del modelo lineal.

Figura 6. Gráfica de dispersión LITROS vs KILOS



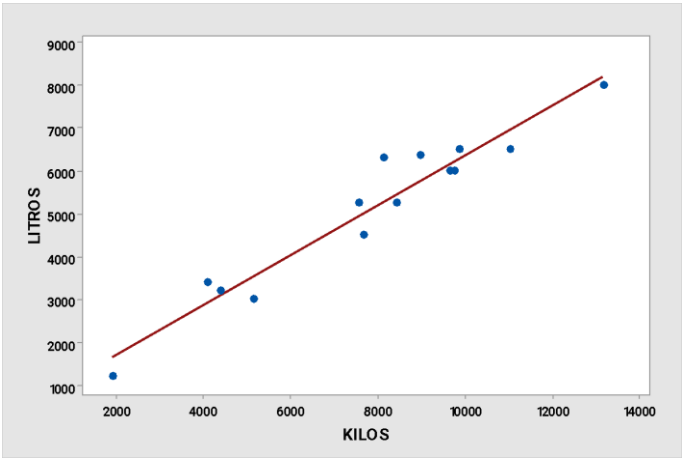
Elección del modelo

Se ajusta el modelo de regresión lineal.

$$y = \hat{\beta}_o + \hat{\beta}_1 x$$

Por medio de la plataforma Excel y Minitab se desarrollaron los estudios:

Figura 7. Gráfica del modelo de regresión lineal



Concluimos que $y = 0{,}5816x + 550{,}77$ con $R^2 = 0{,}9332$. Sin embargo, a pesar de que el modelo en general es adecuado ya que el valor P de la prueba es 0 (Tabla 5), y por lo tanto

es menor que el nivel de significancia 0,05, el coeficiente de correlación es 0,966, cercano a 1, las pruebas de hipótesis:

$$H_o: \beta_O = 0 \qquad H_o: \beta_1 = 0$$
$$H_1: \beta_0 \neq 0 \qquad H_1: \beta_1 \neq 0$$

se rechazan, es decir, a nivel poblacional la ordenada al origen y la pendiente de la recta de regresión son distintas de cero; el error estándar para la ordenada al origen no es del mismo orden que el error estándar de la pendiente, por lo tanto, se decide ajustar a un modelo sin constante.

Tabla 6. Análisis de la varianza del modelo lineal

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	40573811	40573811	167,69	0,000
KILOS	1	40573811	40573811	167,69	0,000
Error	12	2903511	241959		
Total	13	43477321			

El modelo sin constante tiene la forma:

$$y = \beta x + \varepsilon$$

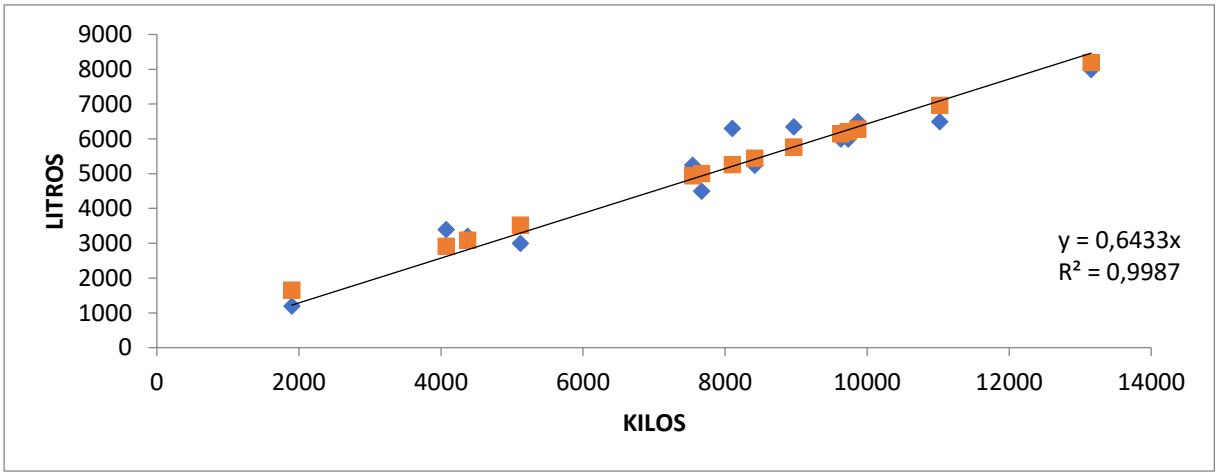
La salida computacional que muestra el modelo ajustado es la siguiente:

$$y = 0{,}6433\, x$$

El coeficiente de determinación obtenido es igual a 0,9872, mucho más cercano a 1 que la ecuación obtenida anteriormente.

Se puede apreciar el ajuste del modelo a partir del siguiente gráfico:

Figura 8. Gráfica de curva de regresión ajustada



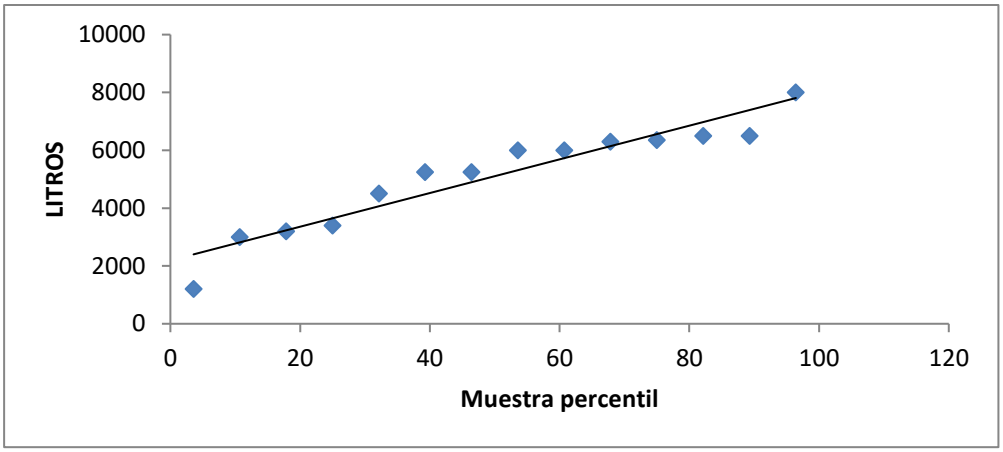
Validación del modelo

Con el objetivo de validar el modelo adoptado, se obtuvo el coeficiente de determinación y la gráfica del modelo lineal sin constante (figura 8).

Al ser el valor del coeficiente de determinación 0,9872 (un número muy cercano a 1), se concluye que el modelo se ajusta perfectamente. Lo que quiere decir que es explicativo.

Para terminar de validar el modelo, se debe cumplir el supuesto de normalidad del error estadístico. Graficando los residuos se puede confirmar dicho supuesto.

Figura 9. Gráfica de distribución normal de los residuos.



CONCLUSIONES

Se concluye que el modelo estimado es:

$$lt\ obtenidos = 0{,}6433 * kg\ uva$$

Representando apropiadamente los datos.

El modelo lineal logrado, aunque sencillo permite observar la dependencia de los litros obtenidos en un proceso de vinificación con respecto a los kilogramos de uva que ingresan a una bodega.

Es útil para representar los datos y es una alternativa válida para modelar matemáticamente este tipo de procesos, en los cuales todavía el arte de elaboración del vino y el criterio del enólogo, son primordiales a la hora de la obtención del producto.

El plan de negocios de la empresa es expandir la producción un 15%, por ello a través del modelo podemos recomendar a la Enóloga a incrementar la producción de uva en 18134 kg.

$$71450lt + 10717lt = 0{,}6433 * kg\ uva$$

$$127728 = kg\ uva$$

BIBLIOGRAFÍA

George E. P. Box, J. S. (2008). Estadística para Investigadores : diseño, innovación y descubrimiento . Barcelona: Reverté.

Hines, W. W. (1996). *Probabilidad y estadística para ingeniería y administración.* México: Continental, S.A. de C.V.

RONALD E. WALPOLE, R. H. (2012). *Probabilidad y estadística para ingeniería y ciencias, Novena edición.* México: PEARSON EDUCACIÓN.

Rosa Ana Rodriguez, M. R. (2010). *Modelado estadístico de la producción de vinos tintos finos.* Colombia: Ingeniería e Investigación.

WINES OF ARGENTINA. (s.f.). Obtenido de WINES OF ARGENTINA: https://www.winesofargentina.org/es/about-us/wofa/

Minitab, LLC. (2018). *Minitab*. Retrieved from https://www.minitab.com

IV Jornadas Internacionales de Estadística Aplicada

Análisis de temperaturas durante la cocción de ladrillos macizos

Farias Marcelo Alejandro, Colodro Adrián, Flores Federico Alberto, Soliz Chesa Sergio Gabriel, Rojas Marcela Abigail

Institución: Universidad Nacional de Salta, Facultad de Ingeniería.

INTRODUCCIÓN:

En la actualidad aún se presenta la producción de ladrillos cerámicos de forma artesanal, esto está originado por la alta demanda de este producto para la construcción. El proceso de fabricación comienza mezclando agua, arcilla, aserrín entre otros, hasta tener una pasta uniforme y manejable, que es llevada a un lugar abierto donde se coloca en moldes de madera; una vez retirado el molde se obtiene el ladrillo en crudo que se deja al sol por un tiempo de secado de 48 horas en promedio. Finalmente, los ladrillos son llevados a hornos de cocción artesanales durante un periodo que oscila entre 2 y 3 días a una temperatura máxima que varía entre 800 y 1300 °C, hornos que principalmente funcionan con gas natural, carbón, y leña entre otros.

Los bloques para mampostería fabricados de arcilla son utilizados en la construcción de todo tipo de edificios, por su disponibilidad y relación costo/beneficio debido a sus propiedades tanto estructurales como térmicas, la importancia de la cerámica además de lo económico radica en que el proceso de cocción de la pasta permite la obtención de un producto duro que no es modificable en su forma, que es frágil, pero presenta resistencia mecánica y dureza, capaz de soportar agua y en algunos casos con baja conductividad térmica.

Las propiedades mecánicas de los materiales utilizados en la construcción son consideradas para el diseño de las estructuras, en cualquier edificación. Los bloques de mampostería son sometidos a cargas que pueden llegar a fracturarse si no poseen las propiedades adecuadas, por lo que es indispensable para cualquier comercializador de este producto garantizar a los constructores el cumplimiento de las exigencias de acuerdo con las normas establecidas en cada país. Sin duda, la cocción es clave, dado que en esta etapa se manifiesta si el proceso de fabricación se ha realizado correctamente y de esta depende que el ladrillo cocido alcance las propiedades mecánicas esperadas, según las exigencias de las normas.

La cocción es un proceso fisicoquímico donde se genera movilidad atómica que une las partículas de arcilla y disminuye la porosidad; es necesario mantener la rapidez con la que se varía la temperatura en el horno, dado que un descenso súbito en la temperatura puede generar una rápida contracción y provocar tensiones que llegan a fracturar el material; igualmente se debe conocer y mantener el rango de temperatura adecuado durante la cocción, el cual depende del material utilizado en cada quema.

UBICACIÓN

El estudio se realizó en la ladrillera El Recreo, ubicada en el municipio de Ocaña, en Colombia, a una altitud de 1754 metros sobre el nivel del mar, con las coordenadas N 08°13'52,14" y W 73°20'20,39" con una temperatura promedio de 21 °C; el horno de sección transversal circular tiene un diámetro interior de 2,12 metros y una altura de 4,52 metros con espesor de pared de 0,24 metros, es cargado con 4300 ladrillos en cada quema.

METODOLOGÍA

Se organizaron los bloques en el horno por niveles; para tomar las muestras se utilizó el muestreo por estratos, se dividió el área del horno en 3 niveles de manera que cada uno sea claramente diferenciable con el objetivo de tomar las muestras y que estas representan el conjunto o nivel del cual provienen; de cada nivel se seleccionaron 20 muestras y se utilizaron cinco para cada ensayo. Para registrar la temperatura interna presentada en cada nivel del horno se instalaron termopares de bulbo de aleación de cromo aluminio tipo K con aislamiento cerámico y para las temperaturas externas termopares de alambre tipo K con recubrimiento fiberglass a 900 °F.

Modelo utilizado: regresión lineal múltiple

DESARROLLO

Las propiedades que se decidieron analizar para estimar el módulo de rotura o flexión y verificar si estas influyen en él fueron las siguientes:

Tasa de absorción inicial

La tasa de absorción inicial determina la cantidad de agua que absorbe el bloque por medio de sus poros en un minuto. Esta cantidad de agua que absorbe un producto cerámico afecta a sus propiedades finales y por ende, puede llegar a causar deficiencias estructurales y de calidad.

En nuestro experimento se realizó la medición de 5 muestras seleccionadas en cada nivel del horno y consistió en determinar el área de la superficie, la masa del bloque seco y la masa del bloque húmedo; para medir estas muestras se utilizó un pie de rey digital con rango de 350 mm y para pesarlos se utilizó una balanza electrónica digital.

La tasa de absorción se determinó con la siguiente ecuación:

$$TIA = \frac{G}{A}$$

Donde G es la diferencia entre la masa inicial seca y la masa final (húmeda) (g/min) y A es el área neta en contacto con el agua (cm^3)

Los resultados obtenidos se presentan a continuación:

Muestras	Area de C.	Masa inicial	Masa final	diferencia	TIA	TIA .Prom	Desvio Stan
1	284,99	3,41	3,51	0,09	0,33		
2	278,31	3,41	3,50	0,09	0,32		
3	283,66	3,35	3,42	0,07	0,25	**0,34**	**0,10**
4	280,08	3,55	3,63	0,08	0,28		
5	277,12	3,41	3,56	0,14	0,52		
Promedio	**280,83**	**3,43**	**3,52**	**0,10**	**0,34**		
6	270,84	3,08	3,14	0,06	0,24		
7	277,98	3,19	3,27	0,08	0,30		
8	275,61	3,03	3,13	0,11	0,38	**0,35**	**0,12**
9	274,94	2,94	3,02	0,08	0,29		
10	262,61	2,98	3,12	0,14	0,54		
Promedio	**272,40**	**3,04**	**3,14**	**0,10**	**0,35**		
11	284,61	3,54	3,67	0,12	0,43		
12	282,47	3,36	3,55	0,19	0,66		
13	284,80	3,43	3,54	0,11	0,39	**0,38**	**0,19**
14	285,39	3,59	3,64	0,05	0,18		
15	290,17	3,59	3,66	0,07	0,25		
Promedio	**285,49**	**3,50**	**3,61**	**0,11**	**0,38**		

Se encontró que la absorción inicial de agua presenta promedios de 0,34; 0,35 y 0,38 para los niveles 1, 2 y 3 respectivamente.

Porcentaje de absorción durante 24 horas

Se seleccionaron 5 muestras por cada nivel del horno, luego se secaron y enfriaron, se encontró la cantidad de masa seca, para sumergir en agua en un rango de 15°C a 30°C por 24 horas.

La cantidad de agua que absorbe un ladrillo afecta principalmente en su consistencia física y su durabilidad. Si la unidad tiene absorción alta, puede presentar cambios volumétricos significativos o permeabilidad alta a la penetración de agua, y puede causar decoloración.

La cantidad de absorción de agua durante estas 24 horas se determinó mediante la siguiente ecuación:

$$\%absorción = \frac{Ws - Wss}{Ws} * 100$$

Donde Wss es la masa sumergida de la muestra (g) y Ws la masa seca de la muestra previa a la inmersión (g).

Los datos obtenidos de esta experiencia son:

Muestras	Masa seca Ws (g)	Masa sumergida Wss (g)	Diferencia Ws y Wss	Absorción	%Absorción	Abs.24 hs Pro	Desvío Stan.
1	3,25	3,75	-0,50	0,15	15,47		
2	3,20	3,67	-0,47	0,15	14,74		
3	3,25	3,73	-0,48	0,15	14,64	**15,04**	**0,87**
4	3,21	3,73	-0,52	0,16	16,30		
5	3,20	3,64	-0,45	0,14	14,05		
Promedio	**3,22**	**3,70**	**-0,48**	**0,15**	**15,04**		
6	3,21	3,73	-0,52	0,16	16,22		
7	2,95	3,49	-0,54	0,18	18,36		
8	2,85	3,36	-0,51	0,18	17,78	**17,25**	**1,11**
9	2,86	3,32	-0,46	0,16	15,90		

10	2,83	3,34	-0,51	0,18	18,00		
Promedio	**2,94**	**3,44**	**-0,51**	**0,17**	**17,25**		
11	3,47	4,02	-0,55	0,16	15,92		
12	3,42	3,96	-0,54	0,16	15,80		
13	3,56	4,08	-0,53	0,15	14,77	**15,39**	**0,47**
14	3,43	3,95	-0,52	0,15	15,21		
15	3,56	4,11	-0,54	0,15	15,26		
Promedio	**3,49**	**4,02**	**-0,54**	**0,15**	**15,39**		

La técnica usada nos permite determinar la absorción de agua final al término de 24 horas, las muestras ubicadas en el nivel 2 del horno son las que presentan mayor tasa de absorción con un porcentaje promedio del 17,25 %, entre el nivel 1 y el nivel 3 los valores son similares 15,04 % y 15,39 %, respectivamente.

Resistencia a la compresión

La resistencia a la compresión es una de las propiedades más importantes para definir la calidad del bloque cerámico, es la característica mecánica principal de materiales para la construcción como el concreto o los ladrillos. Se define como la capacidad para soportar una carga por unidad de área, y se expresa en términos de esfuerzo, generalmente en kg/cm^2, MPa y con alguna frecuencia en libras por pulgada cuadrada (psi).

El ensayo de compresión se realizó a las 5 muestras seleccionadas de cada nivel del horno, las cuales se secaron durante 24 horas en un horno marca Pinzuar, a temperatura entre 110 °C y 115 °C, luego se enfriaron a 24 °C ± 8 °C, con valores de humedad del 30 al 70 % por 4 horas, posteriormente se enfrentaron con yeso aplicado en cada cara con un espesor máximo de 3 mm, finalmente se llevaron las muestras a una universal marca Pinzuar de 1000 KN, y se aplicó carga en el área de contacto hasta llevarlas a la falla.

La resistencia a la compresión para cada muestra se obtuvo por medio de la siguiente ecuación:

$$fcp = \frac{W}{A}$$

Donde fcp es la resistencia de la muestra a la compresión (MPa), W carga máxima de rotura (N) y A el promedio de áreas brutas superior e inferior de la muestra (mm2).

Los datos obtenidos fueron los siguientes:

Muestras	Area A (mm2)	Carga máxima W (N)	Resistencia a la compresión fcp (Mpa)	Resistencia a la compresión promedio fcp (Mpa)	Desviación estándar (Mpa)
1	27266,97	77200	2,831264347		
2	26872,43	81500	3,032848164		
3	27266,4	82960	3,042572543	2,93	0,10
4	27233,04	78500	2,882527988		
5	26806,41	76850	2,866851622		
Promedio	**27089,05**	**79402,00**	**2,93**		
6	23383	94340	4,034555019		
7	26920	155330	5,770059435		
8	27025	102120	3,778723404	4,05	1,07

9	28084	79080	2,8158382		
10	27116	104879	3,867790235		
Promedio	**26505,60**	**107149,80**	**4,05**		
11	28353	182640	6,441646387		
12	28134	197610	7,02388569		
13	28238	218070	7,72257242	7,25	0,59
14	28919	228850	7,913482486		
15	28188	201325	7,142223641		
Promedio	**28366,40**	**205699,00**	**7,25**		

Según los datos obtenidos se puede observar que las muestras ubicadas en el nivel 3 presentan mayor resistencia a la compresión (7,25 MPa.) mientras que las del nivel 1 fueron los bloques con la peor resistencia a la compresión con un 2,93 MPa.

Creación del modelo de regresión

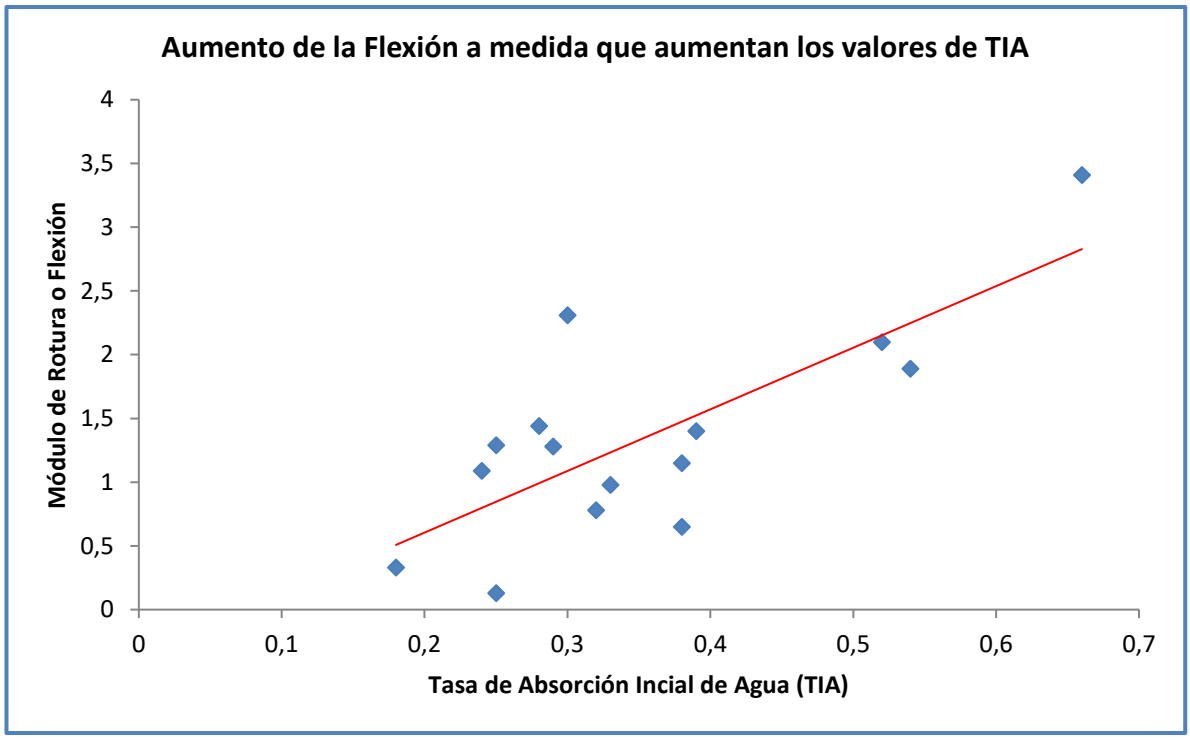
Objetivo: establecer un Modelo de Regresión Lineal Simple (MRLS) que permita estimar el Módulo de Rotura o Flexión, con la Tasa de Absorción Inicial de Agua (TIA). No se encontró relación entre la Resistencia a la Compresión y las demás variables, así como tampoco se observó una influencia significativa de la Inmersión durante 24 horas o la Velocidad de Aumento de la Temperatura, con el Módulo de Rotura.

Para la resolución del problema se diseñó un MRLS, utilizando la ecuación (1), donde la variable de respuesta Y es el Módulo de Rotura o Flexión, la variable regresora X_1 es la Tasa de Absorción Inicial de Agua (TIA).

$$Y = \beta_0 + \beta_1 * X_1$$

Ecuación (1): Modelo de Regresión Lineal Simple

Para poder estimar estos coeficientes β_0 y β_1 del modelo se analizaron 15 pares ordenados (X, Y) que constituyen la muestra cómo se observa en la Gráfica (1).



Gráfica (1): Diagrama de dispersión de datos muestrales del Módulo de Rotura o Flexión en función de la Tasa de Absorción Inicial de Agua (TIA).

Para finalizar el trabajo se verificaron los modelos obtenidos a través del Análisis de Varianzas y las pruebas de hipótesis de los coeficientes β_0, β_1 para poder realizar predicciones válidas que tengan un fundamento objetivo para establecer un punto óptimo de trabajo.

Modelo de Regresión Lineal Simple ajustado.

Mediante el software Excel® se obtuvo la función lineal que mejor ajusta los valores muestrales como se observa a continuación.

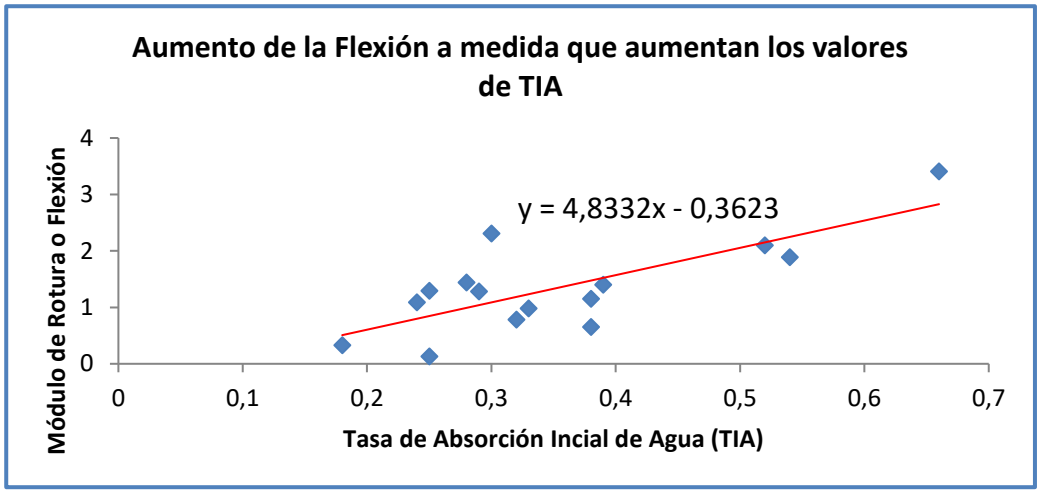
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,759254645
Coeficiente de determinación R^2	0,576467616
R^2 ajustado	0,543888202
Error típico	0,560025106
Observaciones	15

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	5,549407781	5,549407781	17,69422905	0,00102714
Residuos	13	4,077165553	0,313628119		
Total	14	9,626573333			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*
Intercepción	-0,362293596	0,431684453	-0,839255603	0,416492635	-1,294891158
X1	4,833221081	1,149002143	4,206450885	0,00102714	2,350952865

Tabla (1) Análisis de regresión del modelo con Excel®



Gráfica (2): Recta del Modelo de Regresión Lineal Simple ajustado utilizando Excel®

Se obtuvo la siguiente información con estos gráficos:

Para el primer caso.

· **Pendiente β_1**: Representa la variación de la variable de respuesta Y (Módulo de Rotura o Flexión), cuando la variable regresora X (TIA) varía en una unidad de aumento de Tasa de Absorción Inicial de Agua. La pendiente de la recta es positiva como se puede observar en el gráfico, por lo tanto, a medida que aumenta la Tasa de Absorción Inicial de Agua aumenta el Módulo de Rotura o Flexión.

· **Ordenada β_0**: La intersección con el eje Y representa el valor de la variable de respuesta Y (Módulo de Rotura o Flexión), cuando la variable de regresión X (TIA) es igual a cero. Representa el Módulo de Rotura o Flexión inicial de los ladrillos recién cocidos.

· **Coeficiente de determinación R^2**: Es una medida de qué tan bien explica el modelo matemático la variabilidad del Módulo de Rotura o Flexión de los ladrillos. En este caso la variabilidad del Módulo de Rotura o Flexión, está explicada en un 57,64% por humedad que poseen los ladrillos luego del moldeado, se sabe que la cantidad de agua que absorbe un producto cerámico afecta sus propiedades finales cuando es utilizado en mampostería.

Análisis de Varianza

Utilizando el software Excel® se procedió a realizar el estudio de varianza del modelo Ecuación (1), para conocer los aportes a la variación de los valores de la variable de respuesta Y por parte de la Regresión y del Error Experimental, obteniendo los resultados que se muestran en la siguiente tabla:

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	5,549407781	5,549407781	17,69422905	0,00102714
Residuos	13	4,077165553	0,313628119		
Total	14	9,626573333			

Tabla (2) Cuadro del análisis de varianza para el modelo de regresión

Con este estudio se puede realizar la prueba de hipótesis para verificar si la constante estimada asociada a la variable regresora X es válida

Hipótesis Nula: $\beta_1 = 0$

Hipótesis Alterna: $\beta_1 \neq O$

Se calculó un estadístico observado F = 17,69 con grados de libertad $v_1 = 1$ y v2 = 13 y se comparó con un valor crítico $F_{crítico}$ = 0,001 con grados de libertad v1 = 1 y v2 = 13 considerando un nivel de significancia α = 0,05. Como resultado existe suficiente evidencia para rechazar la hipótesis nula, indicando que esta variable X (Tasa de Absorción Inicial de Agua) es significativa en el modelo.

Intervalos de confianza para los coeficientes β_1 y β_0

Empleando el software Excel® se procedió a calcular los intervalos de confianza del 95% y del 99% para estimar los verdaderos valores de los coeficientes β_1 y β_0, como se observa en la Tabla (3).

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	-0,362293596	0,431684453	-0,839255603	0,416492635	-1,294891158	0,570303966
X1	4,833221081	1,149002143	4,206450885	0,00102714	2,350952865	7,315489297

Tabla (3) Intervalos de confianza

Interpretación

El IC para la pendiente poblacional β_1 está comprendido entre el valor 2,350953 y el valor 7,315489, con una confianza del 95%. Puede decirse entonces que en 95 de cada 100 intervalos de confianza construidos, la pendiente poblacional β_1 tendrá un valor entre 2,350953 y 7,315489.

Podemos decir que en el 95% de los casos en que calculemos los IC, los valores del intercepto β_0 o Módulo de Rotura estarán entre -1,29489 y 0,5703 Pa.

Análisis de correlación lineal simple ρ

Para asegurar la veracidad del modelo se calculó el coeficiente de correlación lineal ρ como se indica en la ecuación (2), teniendo como valor promedio de X = 0,354 g/cm^2/min (que son los valores promedio de la Absorción Inicial de Agua luego del moldeado de los ladrillos). Entonces el valor promedio de Y = 1,349 Pa.

$$\rho = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Ecuación (2) Coeficiente de correlación lineal ρ

Como resultado se obtuvo un valor de ρ = 0,76. Esto permite decir que, con los datos muestrales considerados se observa una significativa correlación entre variables, del intervalo [±0,85 , ±0,95] (FUERTE).

Podemos informar que ambas variables consideradas: X = Absorción Inicial de Agua de los ladrillos macizos recién moldeados; Y = Módulo de Rotura o Flexión de los ladrillos luego de la cocción en los hornos, se encuentran fuertemente correlacionadas.

CONCLUSIONES

La sencillez matemática para la regresión lineal, hacen de esta metodología una herramienta sencilla que permita estimar el Módulo de Rotura o Flexión, con la Tasa de Absorción Inicial de Agua, con una buena aproximación.

Pudiendo utilizar el software Excel, para modelar y calcular exitosamente la regresión lineal, aplicando los conocimientos adquiridos en clases.

Al comparar la curva ideal de cocción para las arcillas, con la curva real de los tres niveles en el horno, se observa un gran desfase entre ellas durante las primeras horas del precalentamiento, aunque el mayor desfase se presenta con el tercer nivel que corresponde a la parte superior del horno y es que la combustión se inicia en la parte inferior del horno y avanza en la dirección de abajo hacia arriba y la velocidad de calentamiento es lenta.

Dicho desfase da lugar a una auténtica lluvia ácida sobre el material seco, ahuecando toda su estructura y manchando las superficies expuestas a los gases. Debido a que durante el proceso de cocción no se controló la temperatura ni el tiempo, esto incidió en las propiedades finales del ladrillo. Por lo tanto no se encontró relación entre la Resistencia a la Compresión y las demás variables, así como tampoco se observó una influencia significativa de la Inmersión durante 24 horas o la Velocidad de Aumento de la Temperatura, con el Módulo de Rotura.

Se encontró una incidencia de la tasa de absorción inicial (TIA) y el módulo de rotura (MR), se obtuvo un coeficiente de correlación múltiple r = 0.76 y un coeficiente de determinación r2 = 0.58 que explica la proporción de la varianza del módulo de rotura explicado por el modelo, el coeficiente de regresión α = -0.36, expresa el valor del MR cuando la TIA y la velocidad son iguales a 0.

BIBLIOGRAFÍA

Afanador, N. y Guerrero, M.R. (2012). Propiedades físico-mecánicas en ladrillos macizos para mampostería. Ciencia e Ingeniería Neogranadina, 22(1), 43-58.

Alfaro, M. (2002). Introducción al muestreo minero. Santiago de Chile: Instituto de Ingenieros de Minas de Chile.

Apuntes de la Cátedra de la facultad de ingeniería – Estadística Experimental.

Shackelford, J. (2010). Introducción a la Ciencia de Materiales para Ingenieros. México: Pearson Educación.

García, C.; García, M. y Vaca, M. (2013). Resistencia mecánica de ladrillos preparados con mezclas de arcilla y lodos provenientes del tratamiento de aguas residuales. Revista Tecnura, 17(38), 68-81.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Árbol de decisión. Diseño de procesos bajo condiciones de Incertidumbre

Orlando José Domínguez y Julieta Martínez

Facultad de Ingeniería, CIUNSa, Universidad Nacional de Salta. Salta

odominguez@ing.unsa.edu.ar; jmartinez@ing.unsa.edu.ar

RESUMEN
En Ingeniería química se trabaja para resolver problemas determinísticos. En la cátedra de Diseño de Procesos se pretende incorporar herramientas para el manejo de la información bajo condiciones de incertidumbre. Brindar soluciones racionales para los problemas donde se maneja información incierta. Se propone, como novedad, resolver estos inconvenientes, a través de la integración de conceptos que, si bien son a esta altura conocidos, los mismos no fueron utilizados para la resolución de árbol de decisión. La aplicación de la incertidumbre no tiene otro sentido sino para la toma de decisión, considerando que la función objetivo, los parámetros, y los atributos están sujetos a variaciones. Se formula un problema global donde además de incorporar valores de probabilidad, estos también pueden ser distribuciones de probabilidad, considerando de esta manera la aleatoriedad de los resultados en la vida cotidiana. Este tipo de problemas se resuelve mediante herramientas disponibles actualmente, y software accesible que permite obtener un resultado más ajustado a la realidad, presentando los resultados semejantes a un pronóstico de ocurrencias en el futuro. Se muestran ejemplos donde se aplican los métodos y técnicas que se enseñan y aplican en la Cátedra de Diseño de Procesos. Finalmente, se resuelve un problema específico aplicando la herramienta al análisis de la sensibilidad de los parámetros a fin de observar la variación en los resultados de un problema de tipo árbol de decisiones.

Palabras clave: diseño de proceso, diseño bajo incertidumbre, análisis sensibilidad, árbol de decisión, diseño integrado.

INTRODUCCIÓN

En Ingeniería Química, la característica principal de los problemas y las operaciones en planta es que están sujetas a variaciones en los parámetros, en otras palabras a incertidumbre. Esta variación se manifiesta de múltiples formas, la misma está presente en el entorno del proceso, debido a la ausencia de cierta información para definir completamente el análisis, como precios, demandas, etc. Algunos de ellos se mencionan a continuación:
Incertidumbre en los parámetros: acerca del valor verdadero de los parámetros de tipo tecnológico, usados en el análisis.
Incertidumbre en el modelo, estos no son apropiados para representar la realidad.
Incertidumbre en el tiempo: los resultados de las determinaciones se materializan en el futuro, el mismo a este momento es totalmente desconocido.
Incertidumbre en los componentes climáticos y temporales: ya que un fenómeno climático puede influir y cambiar un proyecto.
Incertidumbres asignadas a datos de referencia tomados de manuales: La incertidumbre expandida de medida, cuyo símbolo sugerido es U, se determina como medida de incertidumbre que define un intervalo sobre el resultado de medida "y" dentro del cual se cree con una alta probabilidad, que su valor de medida Y real estará en el intervalo, $y-U \leq Y \leq y+U$, el cual se escribe comúnmente como Y=y±U, (Ruiz Armenteros et al., 2010).
En tanto que el enfoque conocido como diseño en condiciones de incertidumbre, se ha estudiado por más de cuatro décadas. Su estudio comenzó, antes de que se desarrollarán las herramientas necesarias, originalmente el enfoque consistía en utilizar los valores nominales para el diseño básico, y posteriormente aplicar factores de sobre dimensionamientos de carácter netamente empírico, tal como se aplica en el dimensionamiento de equipos, o como coeficientes de seguridad o de sobredimensionamiento, para considerar las incertidumbres involucradas, Maroto, et al., (2001)
La utilización excesiva de factores de diseño y el uso de valores nominales, ignora otros valores posibles de las incertidumbres, además de que el empleo de factores de sobre diseño no garantiza el funcionamiento viable en toda la incertidumbre y puede variar si no se tiene conocimiento sobre el grado de flexibilidad del diseño en cuestión, lo que puede resultar en costos adicionales innecesarios.
Sobre la base de los objetivos de diseño, González Cortes et al., (2012), mencionan que los enfoques sistemáticos han sido por lo general agrupados en dos categorías. El primero, referido al diseño óptimo para un grado fijo de flexibilidad en las que el diseño debe ser factible en todos los valores inciertos en un conjunto discreto de escenarios factibles que varían con el tiempo (problema de diseño multiperíodo), y el segundo donde el diseño debe ser factible en rangos especificados de un conjunto semi infinito de escenarios (problema general de diseño bajo incertidumbre).
La característica, que debe considerarse en el modelado matemático de estos sistemas, es el hecho de que las variaciones de las variables tienen un comportamiento aleatorio, Scenna, (2000).
Esto es, debido a características inherentes al proceso, factores climáticos o de mercado, etc., las variables de operación no tienen valores únicos y/o fijos, sino que pueden fluctuar en torno a un valor estable, normal o nominal, admitiendo cualquier valor comprendido en un determinado rango de incertidumbre. Un ejemplo de esta problemática es la formulación de los problemas de síntesis de redes de intercambio calórico flexibles y de trenes de destilación integrados flexibles, tal como se cita en Scenna y Benz (2000).
En Ingeniería Química, la función objetivo, desde un punto de vista económico, son normalmente muy sensibles a los precios utilizados, a las variaciones de las variables de entradas, tales como la materia prima, energía, y también a las estimaciones del costo del capital del proyecto. Estos costos y precios se pronostican o estiman, para situaciones futuras, por lo que normalmente están sujetos a un error considerable. La estimación de costos y la predicción de los precios son inciertos, desconocidos al instante de usar las mismas en la determinación de la función objetivo. También existe incertidumbre en las variables de decisión, ya sea por variación en las condiciones de las entradas de la planta, por variaciones climáticas, por variaciones introducidas por operación inestable de la planta, o por la imprecisión en los datos del diseño y las ecuaciones de restricción (Sinnott y Towler, 2012).

METODOLOGÍA

Aplicaciones en Ingeniería Química

El concepto de Diseño en condiciones de incertidumbre figura como contenido mínimo del programa de la materia Diseño de Procesos de la carrera de Ingeniería Química de la Facultad de Ingeniería de la Universidad Nacional de Salta. Se presentan, en esta sección, los contenidos teóricos para abordar el tema de Árbol de decisión, integrando el mismo con el Análisis de sensibilidad.

En esta sección se presentan y explican las diferentes herramientas utilizadas para cada uno de los ejemplos desarrollados en la cátedra, que permiten obtener mejoras en los cálculos y análisis de estos.

Árbol de decisión

Se consideran, desde el punto de vista práctico problemas típicos, uno es resolver el árbol de toma de decisión, mediante su optimización a través de programación lineal y programación lineal mixta, con búsqueda de extremo.

Esta técnica permite analizar decisiones de tipo secuencial, basada en el uso de resultados y probabilidades asociadas. Los árboles de decisión se pueden utilizar para generar sistemas expertos, búsquedas binarias y árboles de juegos. Mediante este tipo de estructuras se permite visualizar todas las diferentes alternativas que pueden ocurrir con su correspondiente valoración económica o valor esperado de cada alternativa.

Para este tipo de problemas, se utiliza para resolverlo una aplicación que trabaja sobre planilla de cálculo, con el cual se realiza un análisis similar de sensibilidad. Símil en el sentido que los cambios se realizan manualmente, de forma muy rudimentaria. Para incorporar y facilitar este tipo de análisis, se puede aplicar una herramienta, un complemento que se incorpora sobre Excel, denominado en inglés como *Simple Decision Tree*, que permite construir de forma progresiva, automatizada y muy simple, árboles de decisión complejos y elaborados como (Slashdot Media, 2012) (Tree Plan, 2016), y también (Precisión Tree, 2021).

Este tipo de complemento facilita la construcción del árbol de decisión, e incorpora en las celdas las fórmulas automáticamente, por lo que se debe ingresar de forma manual solo los valores de probabilidad, los demás son calculados inmediatamente. Este tipo de automatización permite disponer de los resultados de forma más rápida, y destinar ese ahorro de tiempo, en proponer algunos cambios en los parámetros, como también realizar un informe de los resultados aún más detallado, que para los casos realizados en papel.

Diseño bajo condiciones de Incertidumbre

Es otro aspecto que se estudia dentro del marco de Diseño de Procesos bajo condiciones de incertidumbre. En este tópico se presentan cuatro conceptos nuevos que son: criterios de diseños, los que también se denominan atributos del diseño, alternativas de diseño, escenarios y los resultados.

Una vez definidos los atributos del diseño, y generadas las alternativas e identificados los escenarios, queda planteada una matriz de resultados, para cada una de las alternativas. Dando un problema de decisión multicriterio, por lo que se debe reducir cada matriz a un vector, ya sea por algún método, tales como el uso de la teoría de juegos con algunas de las estrategias como el de Maximin, denominado criterio de Wald con una visión pesimista, otro el Maximax con una visión optimista, o bien el criterio de Minimax del costo de oportunidad o de Savage, o quizás el criterio Hurwicz intermedio entre la visión pesimista y la optimista o alguna otra como programación lineal, PL, (Taha, 2012). La elección de cada uno de estos métodos dependerá del tomador de decisión y sus consideraciones.

Posteriormente para cada alternativa se debe reducir los vectores a un escalar, mediante el uso de la función de utilidad mediante la cual se realiza la transformación del vector a un solo valor escalar. Este escalar involucra la contribución de todos los atributos para los diferentes escenarios para cada alternativa, Gallardo Ku, (2018). De tal manera que al elegir la mejor de todas las alternativas, también se está eligiendo la mejor combinación de los atributos de diseños involucrados en la función de bondad (Varian, 2012).

De los cuatro conceptos mencionados anteriormente, el conjunto de escenarios son los que están sujetos a la incertidumbre, a través de la asignación de probabilidad de ocurrencia, de materialización de dicho escenario. Los escenarios están sujetos al siguiente conjunto de propiedades:

a. Influyen significativamente sobre el resultado del sistema.
b. No dependen de la voluntad de quien toma la decisión.
c. Su valor es incierto en el momento de tomar la decisión.

Este subconjunto puede subdividirse en aquellos que:

i. No dependen de ninguna voluntad. En estos casos se estudia mediante la teoría de las decisiones individuales. Ejemplo de estos escenarios, son las situaciones climáticas.
ii. dependen de otras voluntades, con intereses distintos. Estos se estudian mediante la teoría de juegos y del planeamiento estratégico. Un ejemplo de escenario de este tipo es cuando en el proyecto influye la posible incorporación de una nueva competencia.

La incertidumbre en este tipo de problemas está incorporada en la probabilidad de ocurrencia del escenario propuesto.

El mayor desafío e inconveniente se presenta en el método de reducción de la matriz de resultados de los criterios de diseños para cada alternativa a un valor escalar por alternativa. La tabla 1 representa los resultados de los criterios de diseños y_i para cada escenario s_k, para una determinada alternativa a_j.

$\{a_j\}$	s_1	s_2	s_3	...	....	s_k
y_1						
y_2						
y_3			$y_i(s_k, a_j)$			
:						
:						
:						
y_l						

Tabla 1. Matriz de resultados para la alternativa *j*.

Donde y_i (s_k, a_j) es la probabilidad de que el criterio de diseño tome el valor y_i dada la ocurrencia del escenario s_k, la alternativa de diseño a_j, y las demás hipótesis (H) impuestas al diseño, esto se expresa como p (y_i | s_k, a_j, H).

En la Fig. 1, se puede observar la construcción del árbol de decisión para la producción de un producto determinado, que se pueden producir por j alternativas con l criterios de diseño y dónde está sujeto a k escenarios inciertos con sus respectivas probabilidades. La toma de decisión corresponde a la elección de aquella alternativa que tenga el mejor valor de todos los Uj valores de utilidad esperados.

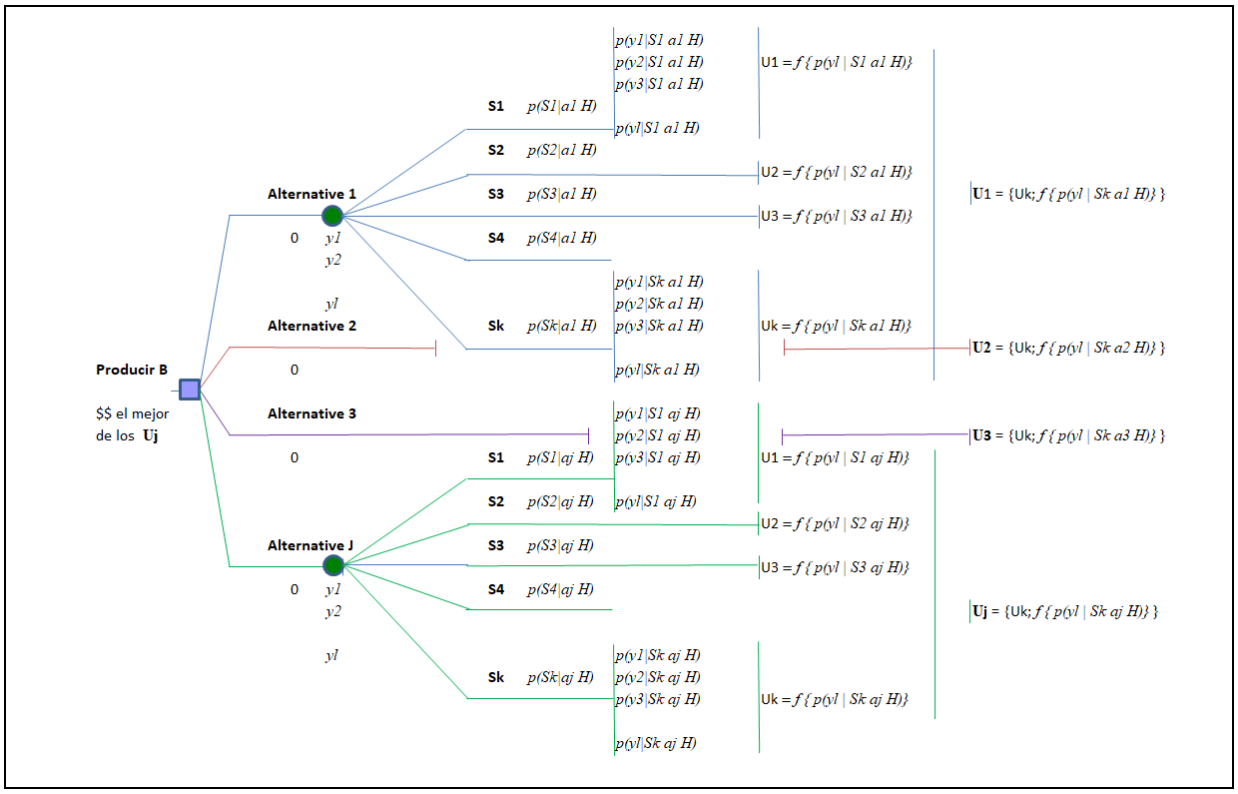


Figura 1. Árbol de decisión para j alternativas diferentes de producción del producto B

Análisis de sensibilidad

El análisis de sensibilidad es una herramienta utilizada en el modelado que permite analizar cómo los diferentes valores de un conjunto de variables independiente afectan los resultados de un proyecto, bajo ciertas condiciones que permite comprender el efecto de la incertidumbre.

Desde el punto de vista formal el análisis de sensibilidad es una técnica que estudia el impacto que tiene sobre una variable dependiente de un modelo de valor (función objetivo, $U=\{uj\}$) las variaciones en una de las variables independiente que lo conforman.

Este análisis observa las variaciones del proyecto ante el aumento o disminución en alguna de sus variables o parámetros claves, manteniendo el valor de las demás constante. Es decir, este análisis se ejecuta de a una variable a la vez y se supone independencia entre las distintas variables que sí o sí influyen en el valor del proyecto.

Una mejora en la resolución de este tipo de problemas es la incorporación de herramientas de análisis de riesgo y manejo de la incertidumbre, como los complementos denominados Cristal Ball, @Risk o bien Risk Simulator, todos ellos operan sobre planillas de cálculos, como Excel.

Ambas empresas cuentan con este tipo de complementos similares, presentan versiones académicas y/o trial de prueba que permite acceder a ellos por un corto periodo de tiempo. Se basan en el modelado predictivo, previsión, simulación y optimización del sistema abordado. En general los problemas de análisis de sensibilidad trabajan sobre la simulación Monte Carlo, que consiste en generar un cuadro de pronósticos que muestra rango entero de posibles valores y la posibilidad de alcanzar cualquiera de ellos, León Sánchez et al. (2004), Del Carpio Gallegos, (2007).

Una resolución de un determinado problema en Excel da una solución determinística, esto significa que se tiene un solo resultado de la variable de salida, para una sola combinación de valores fijos de las variables.

La aplicación de estos complementos permite definir una distribución de probabilidades de valores, dentro de un rango, para cada una de las variables. El complemento permite generar todas las combinaciones de todas las variables en 1.000, 10.000, 100.000, corridas diferentes (simulación de Monte Carlo), generando una distribución de probabilidades de la variable dependiente, que contempla todos los posibles resultados que se pueden dar. Distribución de

la variable de salida que nos permite analizar el efecto de la incertidumbre en el problema, proporcionando un análisis más profundo, por ende, un comportamiento más real del sistema estudiado.
Para cada variable de entrada, se puede proponer diferentes tipos de distribuciones, dependiendo de la información disponible por el tomador de decisión.
Las aplicaciones pueden disponer de diferentes distribuciones para asignar, a la variable de entrada o independiente, se encuentran entre ellas la distribución Normal, Triangular, Uniforme, Logarítmico, Uniforme discreta, Binomial, Exponencial, Pert, Poisson, Gamma, Weibull, además permite incorporar una tabla de valores.

DESARROLLO

Introducción al análisis de sensibilidad
Se aplicó el análisis de sensibilidad, al estudio de los índices de rentabilidad de un proyecto de inversión, variando los diferentes factores que influyen en el mismo. Los indicadores de rentabilidad más utilizados en el estudio de factibilidad económica de un proyecto son: el valor actual neto (VAN), conocido también como valor presente neto (VPN), la tasa interna de retorno (TIR) es otro, también la relación beneficio costo (B/C), o bien el periodo de recuperación de la inversión e inclusive el índice de rentabilidad, entre los mencionados por (Towler, Sinnott, 2012), también Peters (2003).
Los indicadores de rentabilidad, en general, dependen de las variables independientes, tales como el costo de la materia prima (MP), costo de la mano de obra (MO), inversión fija (IF), capital de trabajo (CW), capacidad de producción (CP), precio de ventas (PV), etc.
La elección de qué variables independientes considerar para el análisis, depende de aquellas que contribuyen con mayor porcentaje a la variable dependiente, en nuestro estudio al valor del costo total de inversión (CT).
Otra de las cuestiones a considerar es el rango mínimo y máximo de variación de las variables independientes, en definitiva el porcentaje de aumento y disminución respecto del valor nominal. Muchos autores toman un mismo porcentaje de aumento y disminución e inclusive igual para totas las variables independientes. Este porcentaje de variación mínimo/máximo no necesariamente deben ser iguales para todas las variables de entrada y menos simétrica. En este sentido Towler y Sinnott (2012), presentan una tabla con diferentes parámetros característicos y con rangos de variaciones típicos. En general estas variaciones son simétricas e iguales, variando levemente entre variable y variable. Estos deltas de variación pueden estar entre 5%, 10% o 20% tanto para su incremento como para su disminución. Esta simetría es formal, desde el punto de vista práctico y a través de la experiencia en proyectos de inversión en nuestro país, se comprueba que es prácticamente imposibles que disminuyan en un porcentaje igual que los hacen los aumentos de los valores por ejemplo de los costos de materia prima o los costos de mano de obra, la experiencia indica que estos valores siempre aumentan y que jamás disminuyen, aunque probabilísticamente pueden ocurrir.

Aplicación al ejemplo de sensibilidad
Desde un punto de vista práctico, el análisis de sensibilidad se aplica al estudio de la factibilidad económica, del proyecto de producción de pentaborato de sodio mediante lixiviación con gas CO_2, desarrollado en la tesis de Domínguez, (2019).
Los valores característicos de los parámetros del proyecto considerado como caso base se presentan a continuación:

Valor de la Inversión Total: 8,95 Millones U$S
Tasa de oportunidad: 13%

Las variables independientes o, de entrada:

Costo Materia Prima: 366,74 U$S/t
Precio Producto: 1.200 U$S/t
Inversión en Capital Fijo: 7,44 millones U$S/t
Capacidad de producción: 10.000 t/a

Mientras que las variables de Salida a analizar:

Valor Actual Neto (VAN): 6,87 millones U$S

Tasa Interna de Retorno (TIR): 12,87% a

Se considera una variación de ±20% simétrica al valor base, y para que coincida con el estudio realizado por Domínguez, (2019).

A partir de la planilla de cálculo donde se determina el costo de Inversión Total (CT), se identifican las celdas donde se ingresan los valores de las variables independientes, variables de entrada, se las define como variables inciertas. A estas celdas variables se les asigna una distribución de probabilidades, para este caso se asigna una distribución triangular, debido a que se conoce de antemano el valor nominal, un valor inferior y uno superior.

De igual forma, a las demás variables independientes, a estudiar, se asigna una distribución de probabilidades, dependiendo su tipo a considerar de acuerdo al conocimiento o bien a la información disponible de las mismas. Por lo general cuando se desconoce o no se tiene información se asigna una distribución del tipo Gaussiana, también conocida como distribución Normal. Todas las variables de entradas, que se hayan definido su distribución, mediante el botón *definir distribución*, cambian de color la información de las celdas, para poder identificarlas rápidamente.

Seguidamente, se identifican las celdas donde se obtiene la información de las variables de salidas o dependientes, a las que se definen como variable dependiente, haciendo click en el botón *añadir salida*, en la aplicación.

Para las variables de salidas, como ser el VAN, depende de cada uno de los complementos utilizados, sea este Cristral Ball, @Risk o Risk Simulator, se realiza seleccionando el botón *definir prevención, Añadir salida* o bien *propiedades de pronóstico*, respectivamente. Con dicha acción se define a la celda que presenta un comportamiento incierto a la salida, la misma se representa por una distribución de probabilidad, en este caso del VAN, de igual manera se procede para la TIR, o cualquier otro indicador que se quiera estudiar.

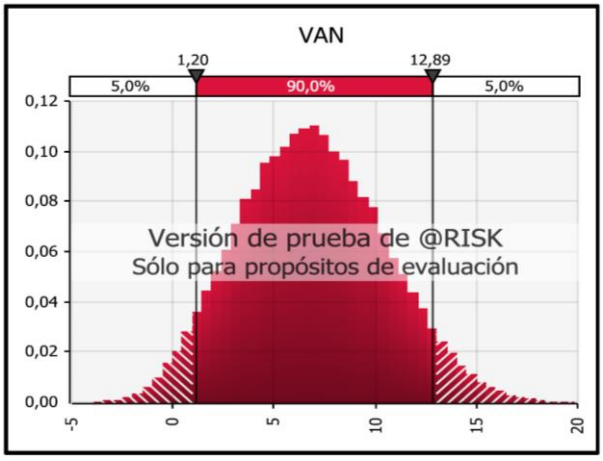


Figura 2. Distribución de probabilidad de la Respuesta VAN

Finalmente, definidas las celdas de entradas y de salidas, se inicia la simulación tipo Monte Carlo, presionando el botón *Iniciar*, *Iniciar simulación* o *Correr*, inicia la simulación. Este proceso realiza tantas simulaciones como se quiera y hasta 100.000 iteraciones, de todas las combinaciones de las variables de entrada posibles para generar una distribución de la variable de salida como se observa en la Fig. 2.

En la Fig. 2 se observa que la probabilidad de que el VAN sea mayor a cero es del 96%. Hay un 4% en las condiciones más extremas inferior donde el proyecto no es factible.

De igual manera, se obtienen distribuciones de probabilidad para las demás salidas requeridas.

Estos complementos, disponen de diferentes tipos de gráficos, para manifestar el comportamiento de las variables de salidas, tales como la curva acumulada o el gráfico de tornado que muestra en qué cantidad y en orden jerárquico la contribución de las variables de entrada a la variación de la salida. En la Fig. 3 se visualiza el grafico de tornado de la TIR, en él se muestra la importancia relativa de cada variable de entrada.

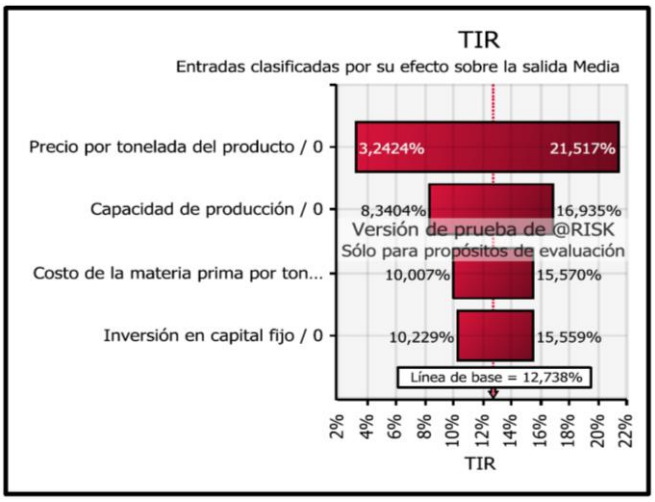


Figura 3. Curva de tornado de la respuesta TIR

En la Fig. 3, se observa la importante contribución del precio del producto a la variación de la TIR. También se observa la relevancia del precio del producto sobre las demás variables, mostrando una jerarquía sobre la capacidad de producción y a su vez de está sobre el costo de materia prima. En tanto que la inversión en capital fijo es el que menos influye en la variación de la TIR.

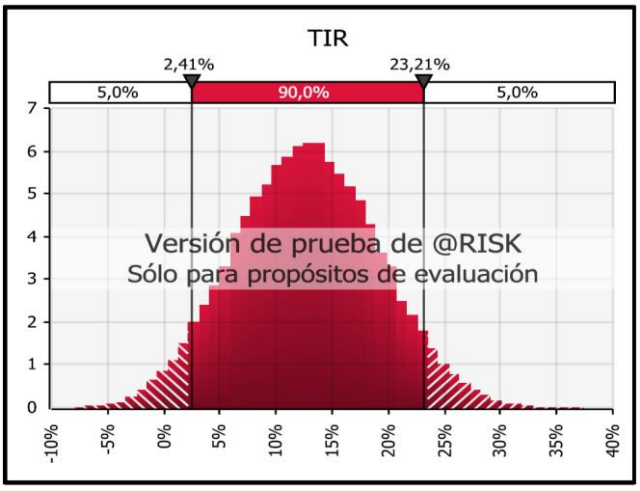


Figura 4. Distribución de probabilidad de la TIR.

El análisis de las gráficas de la salida de la tasa interna de retorno proporciona quizás más información. Se percibe en la Fig. 4 la curva de distribución de probabilidad de la TIR, que representa la contribución de las 100.000 simulaciones de todas las combinaciones posibles de las variables de entradas.
El complemento, por defecto se representa en la Fig. 4 la situación que se produce para el rango de probabilidad de ocurrencia del 90%, que supone que la TIR se encuentra entre un valor mínimo de 2,41% y un valor máximo de 23,21%. También muestra que con un 96% de probabilidades el proyecto presenta valores de la TIR positivos y mayores a 0% (cero).
Con probabilidades cercanas al 50% de las veces el proyecto comienza a ser rentable igualando y supera a la tasa de oportunidad.

CONCLUSIONES

La utilización de softwares y complementos para el análisis del efecto de la incertidumbre, permiten realizar mejores conclusiones y de mayor profundidad, obteniendo simulaciones más

cercanas a la realidad, con conclusiones más amplias sobre los resultados, con métodos rigurosos, ganando mayor certeza y disminuyendo de forma considerable la incertidumbre inicial.

Al emplear estos nuevos conceptos, ya sea de forma independiente, o mediante la incorporación de fórmulas en las planillas de cálculos, o la utilización de los complementos vistos en este trabajo, representan una gran ventaja en el análisis de la incertidumbre de problemas ingenieriles.

Brindan mayor información que los problemas determinísticos. Como se ha visto es posible transformar un problema de diseño de procesos bajo condiciones de incertidumbre, representado por la matriz de resultado en un diagrama de árbol de decisión, por lo que también será factible el camino inverso.

Tanto para los problemas del tipo de árbol de decisión como para los problemas de análisis de sensibilidad de proyecto, se puede incorporar variaciones mediante distribuciones de probabilidad de las variables de entrada porque nos permite disponer de distribuciones de probabilidad de las variables de salida, mejorando de esta forma el abordaje clásico, determinístico, obteniendo un solo valor del resultado.

Así también al incorporar el concepto de escenario, evento de posible ocurrencia en el futuro incierto, en forma de probabilidad de ocurrencia, mejora los pronósticos de análisis.

Se favoreció el análisis de los resultados, en especial en los problemas abordados en la cátedra de Diseño de Procesos de la carrera de Ingeniería Química, de la Universidad Nacional de Salta.

BIBLIOGRAFÍA

Analytic Add-ins, by TreePlan Software For Mac Excel and Windows Excel, Complement TreePlan for Excel. 2016.
https://treeplan.com/

Del Carpio Gallegos, J., Análisis de riesgo en la evaluación de alternativas de inversión utilizando Cristal Ball, *Gestión y Producción, Ind. Data* 10(1), p55-58, 2007.

Domínguez, O. J., *Desarrollo de Tecnologías para la obtención de boratos refinados*, 1er Edición. O.J. Domínguez, Buenos Aires, Argentina. 2019.

Gallardo Ku, J. D., *Notas en teoría de la incertidumbre*. 1a ed., Pontificia Universidad Católica del Perú, Fondo Editorial, Lima, Perú, 2018. ISBN 978-612-317-433-0

González Cortés, M., Pedraza Gárciga, J., Clavelo Sierra, D., González Suárez, E., Incertidumbre en la Integración de Procesos para el desarrollo de Biorefinerías. *Rev. Centro Azúcar* Vol 42, No. 3, Julio-septiembre 2015 (pp. 30-38).

León Sánchez, D. P., Quintero Rodríguez, I. M., Zuñiga Muñoz, W., *Crystal Ball. Software de Análisis y Simulación de Riesgo*. Unidad de informática y Comunicaciones, Facultad de Ciencias Económicas, Universidad Nacional de Colombia, Bogotá, Colombia, PDF, 2004. http://www.fce.unal.edu.co/media/files/UIFCE/Finanzas/Crystal_Ball_1.pdf.

Maroto, A., Boqué, R., Riu, J., y F. Xavier Rius *Incertidumbre y precisión*. 2001. Técnicas de Laboratorio, 266 (2001) 834-837. http://www.quimica.urv.es/quimio/general/incert.pdf

Peters, M., Timmerhaus, K. y West, R. E., *Plant Design and Economics for Chemical Engineers*, 5th Edition, s.l.: McGraw-Hill, 2003. ISBN 0-07-119872-5.

Ruiz Armenteros Antonio M.; García Balbao José L.; Mesa Mingorance José L., *Error, Incertidumbre, Precisión y Exactitud, términos asociados a la calidad Espacial del dato Geográfico*. 1st International Congress on Unified and Multipurpose Cadastre. Universidad de Jaén, España. (2010). ISBN 978-84-8439-519-5

Scenna, N., *Modelado, simulación y optimización de procesos químicos*. Ciudad Autónoma de Buenos Aires, Argentina: Universidad Tecnológica Nacional, 2000.

Scenna, N. J., Benz, S. J., Introducción al diseño de procesos químicos, Breves nociones, En Nicolás J. Scenna. (Ed.), *Modelado, simulación y optimización de procesos químicos*. Universidad Tecnológica Nacional, (pp 29-82), CABA, Argentina, 2000.

Sinnott, R., Towler, G. *Diseño en Ingeniería química*, 5ta edición, Ed. Reverte, Barcelona, España, 2012. ISBN: 978-84-291-7199-0.

Slashdot Media, Source Force: Simple Decision Tree, Sacramento, California, EEUU, (2012). https://sourceforge.net/projects/decisiontree/files/decisiontree/1.4/

Taha, H. A., *Investigación de Operaciones*, Person Education, 9na edición, Naucalpan de Juárez, México, 2012.

Towler G. P., Sinnott, R., *Chemical engineering design: principles, practice, and economics of plant and process design*, Ed. Reverte, 2da Ed., United States of America, 2012.

Varian, Hal R., Microeconomía Intermedia, octava edición, Antoni Bosch editor, Barcelona, España, 1996.

IV Jornadas Internacionales de Estadística Aplicada
9 y 10 de diciembre de 2021

Percepción del riesgo generado por el tránsito vial de alumnos de una escuela semiurbana

Angélica Noemí Arenas[1,3], Héctor Iván Rodríguez[1,3], Heriberto Eduardo Esperón[1], María Josefina Méndez[2], Matías Ezequiel Cardozo[1].

[1]Facultad de Ingeniería, [2]Facultad de Humanidades, [3]Instituto IIDISA, de la Universidad Nacional de Salta. Salta.

angelica@ing.unsa.edu.ar, teléfono celular 387154460604.

RESUMEN

Se presenta un estudio basado en encuestas para obtener información sobre la percepción del riesgo vial de estudiantes de la Escuela Primaria Submarino ARA y de los vecinos de la localidad de San Luis (Salta). Este trabajo se planteó por objetivos el estudio de herramientas para identificar las brechas en la protección de la niñez en el espacio vial, proponer soluciones que generen conductas más seguras de los actores viales y la participación de estudiantes en una práctica de investigación en el medio social. Los resultados obtenidos muestran que los niños poseen una percepción de bajo a mediano riesgo al cruzar la ruta nacional de alto tránsito; a diferencia de los vecinos, que ponderan a este acto cotidiano como una exposición de alto riesgo. Se esperan lograr cambios hacia conductas más seguras de los actores-usuarios de la vía mediante el aporte de medidas logradas entre la escuela, el municipio, la policía provincial y este proyecto, a partir de la información conseguida por las encuestas.

Palabras Clave: seguridad vial, percepción, riesgo, siniestro.

INTRODUCCIÓN

La seguridad vial es un aspecto de importancia en la calidad de vida de las comunidades e impacta de manera significativa en la vida de las personas. Los siniestros viales constituyen un efecto negativo del tránsito vehicular con consecuencias leves, graves y fatales sobre las personas, así como puede producir daños en la infraestructura pública y la propiedad privada.

Entre los actores del espacio vial se encuentran los usuarios viales vulnerables[6] (UVV) constituido por los peatones, los ciclistas y motociclistas (Organización Mundial de la Salud, 2013) y los usuarios de vehículos autopropulsados destinados al transporte de personas y mercancías.

Las políticas de preponderancia de los vehículos y otros factores como la socio cultura, la infraestructura y la existencia de vías con alto tráfico vehicular, entre otros, puede afectar a la movilidad de los UV porque se presentan riesgos hacia los UVV derivado del modo de conducción de los conductores en el espacio vial y la infraestructura, entre otros aspectos de importancia.

En forma particular, la seguridad vial (SV) de la niñez es un aspecto de política comunitaria de desarrollo incipiente en nuestra sociedad, especialmente alrededor de los entornos educativos. Desde distintas organizaciones[7] y administraciones se destaca la necesidad de educar tempranamente a los niños (Pérez, 2008, págs. 144-147) como estrategia de prevención frente a los riesgos de este colectivo de UV en la vía.

En relación con los UVV, un sector de especial interés para este estudio es el de los niños, y se concentra la atención en su movilidad hacia las escuelas, tanto urbanas como rurales.

En este trabajo, el análisis de los factores de influencia sobre la seguridad vial se considera a través del modelo integrado para la evaluación de factores de influencia en los accidentes (Aparicio & Arenas, 2014) y que ha sido aplicado al análisis de la accidentalidad de furgonetas en España en el año 2012.

[6] En el Informe sobre la Situación Mundial de la Seguridad Vial 2013, la Organización Mundial de la Salud define a los peatones, ciclistas y motociclistas como los usuarios vulnerables de la vía pública.

[7] La convención de Naciones Unidas para los Derechos de la Infancia destaca la necesidad de aplicar la responsabilidad social de proteger a los niños y prestar los servicios y ayuda necesaria para abordar esta problemática.

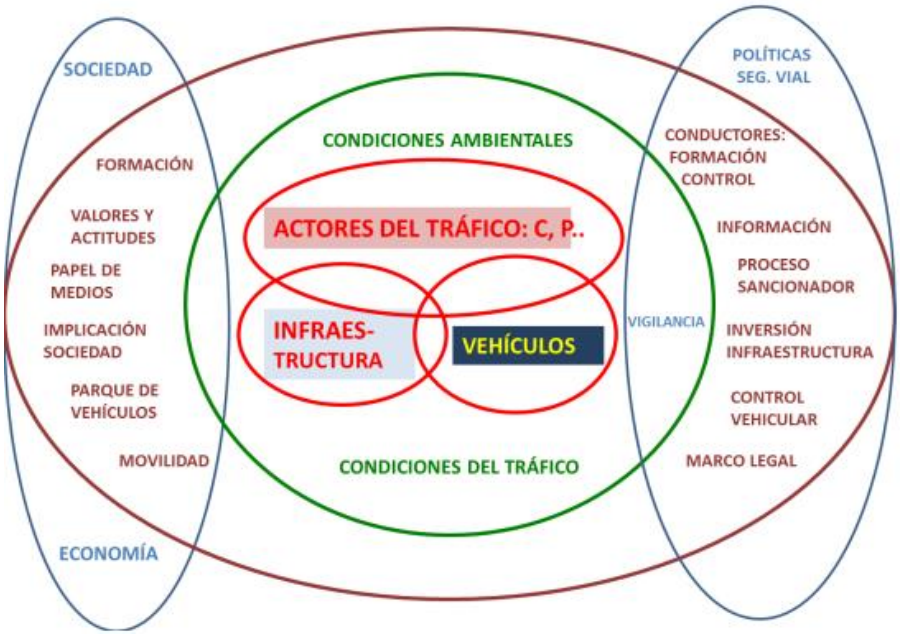


Figura 1: Factores de influencia en accidentes de tráfico. Fuente: Aparicio et al (2011).

Para hacer frente a la complejidad de la investigación de accidentes de tránsito, por el gran número de variables de influencia, se han estructurado en capas en un modelo denominado Modelo Integrado de Investigación Científica de Accidentes (MIICA). La estructura del modelo se muestra en la Figura 2, en la parte central se ubican los factores más directamente relacionados con la ocurrencia de accidentes y en las capas más externas, los demás, en función de la mayor o menor "lejanía" con la que pudieran ejercer influencia sobre los anteriores.

El modelo MIICA (Aparicio & Arenas, 2014) es fruto de la conceptualización llevada a la práctica del equipo en la investigación de la accidentalidad de furgonetas. Existen otras aplicaciones del equipo investigador de UPM (Accidentes de peatones, accidentes con camiones de mercancías, con motocicletas ocurridos en las carreteras españolas) cuyos enfoques pueden inscribirse dentro del modelo MIICA, ya que su estructura incluye varios métodos y herramientas de análisis, tipos de modelos; ensayos y experimentación, simulación y cálculo.

Este modelo define en la parte central los factores más directamente relacionados con la ocurrencia del accidente y en las más externas, los demás, en función de la mayor o menor lejanía que a su vez ejercen influencia sobre las internas.

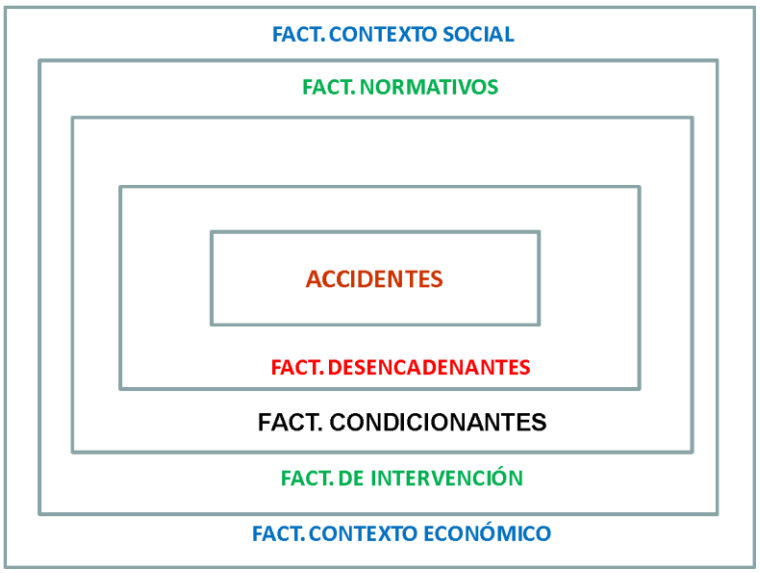


Figura 2: Factores y capas de análisis del accidente de tráfico. Fuente: F. Aparicio y B. Arenas, Proyecto FURGOSEG – España (2011).

Las capas del modelo MIICA se definen en torno a Factores condicionantes, Factores normativos y de intervención de los poderes públicos, Factores de contexto social y Factores de contexto económico. Una aproximación a cada uno de ellos se ilustra en la Figura 3.

Para el desarrollo de los modelos disponibles a través de la estadística y el desarrollo de herramientas a través de la ingeniería matemática, son fundamentales las bases de datos, no solo las de accidentes sino también las del sistema de información del país o la región de indicadores económicos, sociales, de condiciones meteorológicas, de tráfico, de exposición, de censo de conductores, etc., siendo fundamentales la información del sistema de recolección de accidentes de tránsito.

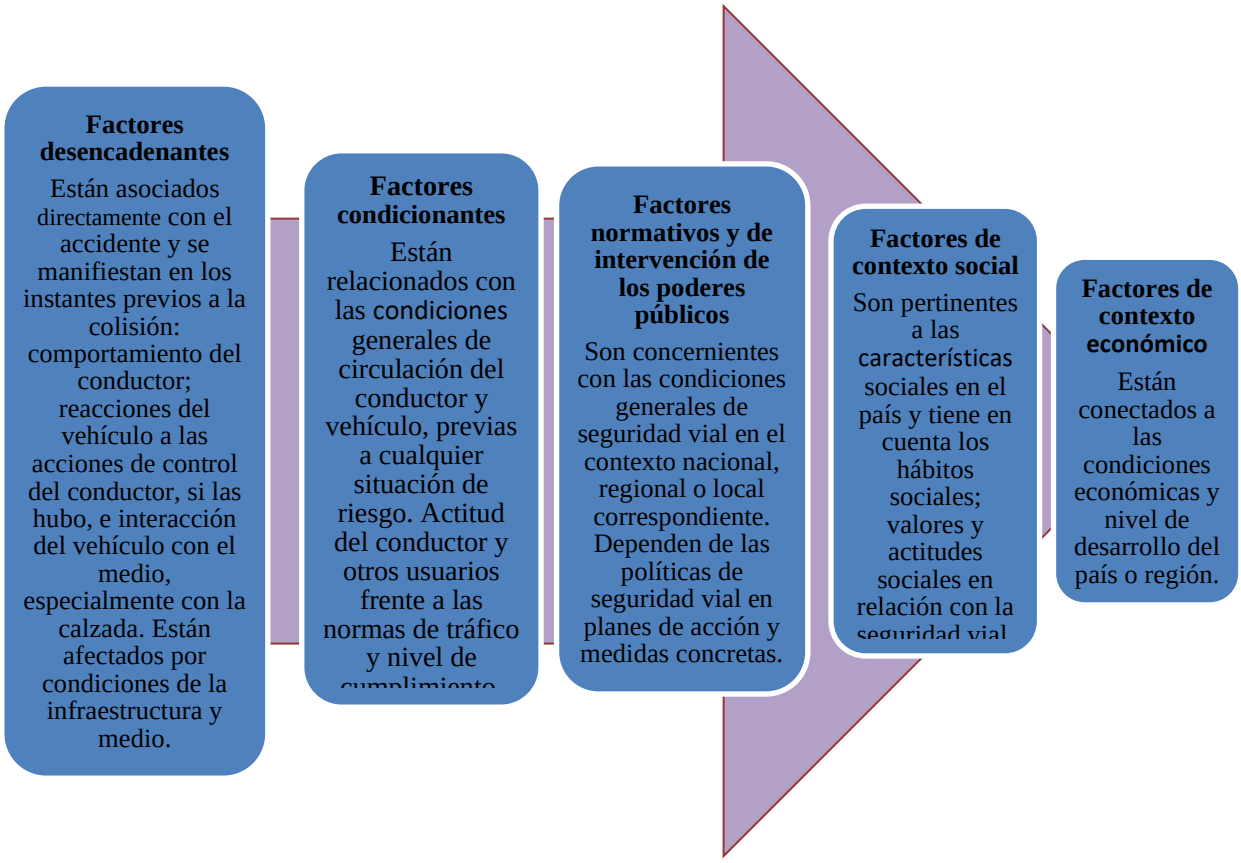


Fig. 3: Factores y análisis de las capas adyacentes del modelo que influencian en los accidentes de tráfico. F. Aparicio, B. Arenas. Proyecto FURGOSEG (2012). 1.2 Proyecto COOPERAXVII-12 (TCP).

El estudio de los factores directamente relacionados con la ocurrencia del accidente se centra en los factores condicionantes y desencadenantes cuya información se puede extraer de las bases de datos de accidentes y de la investigación en profundidad de muestras de accidentes que se pueden definir dentro de la investigación de tipo microscópica.

Varios países han propuesto la limitación normada de la velocidad, en zonas urbanas y rurales; y su cumplimiento es objeto de discusión puesto que no se cumple en un ciento por ciento la ley en los diferentes territorios de nuestro país, básicamente por el modo de conducción. El progreso en aplicar la ley y su efectivo cumplimiento en la reducción de la velocidad en la vía, en especial para este proyecto referido a entornos escolares, tiene camino por recorrer. Aun siendo que constituye uno de los cinco principales factores de riesgo de lesiones y muerte en un siniestro vial como indica la Organización para la Cooperación y el Desarrollo Económico (OECD) y el Banco Mundial (OECD-The World Bank, 2020).

La reducción de velocidad a 20 km/h y 30 km/h en zonas de tránsito de peatones y ciclistas es una de las políticas para definir zonas calmadas con el objetivo de disminuir los riesgos derivados del tránsito.

Con el objeto de evaluar la percepción del riesgo del tránsito vial se realizaron encuestas a la población infantil[8] de la escuela semiurbana de la localidad de San Luis (Salta, Argentina) y a vecinos próximos al entorno escolar.

[8] Se entrevistaron alumnos del nivel primario del turno mañana de la escuela Submarino ARA y vecinos de la localidad de San Luis.

METODOLOGÍA

Las encuestas constituyen una herramienta de investigación relevante para procesos de recolección de datos, los que posteriormente tendrán un fin determinado; uno de ellos es construir información valiosa que permita la toma de decisiones y recomendaciones, basadas en la información obtenida del público objetivo. Son aplicadas tanto a nivel social como empresarial (Quispe Limaylla, 2013).

La encuesta, para este trabajo, constituye un instrumento de investigación social realizada de manera científica y se elaboró con preguntas en un formulario estructurado de respuesta cerrada.

El proceso de muestreo contempló dos segmentos del universo[9] estudiado, el primero lo constituyen estudiantes de la escuela primaria Submarino ARA, mientras que el segundo, se integra por vecinos colindantes al establecimiento y mayores de 18 años. El tamaño de la muestra fue de 125 casos efectivos, de los cuales 78 fueron estudiantes de la escuela y 47 vecinos (Peña Sánchez de Rivera, 2001; Montgomery, 1991).

A cada encuestado se les consultó sobre temas referidos a la seguridad vial enfocando en la percepción del riesgo que el tránsito vehicular ocasiona sobre las personas, y en especial, en los niños que acuden a la escuela.

Cabe mencionar que la ruta nacional 51 tiene su traza enfrente a la entrada principal del establecimiento y posee un intenso tráfico de motos, vehículos de transporte de personas y de cargas.

Fotografía 1. Vista de la ruta nacional 51 y parada de colectivo frente a la Escuela ARA.
Fuente: Material fotográfico propio.

DESARROLLO

Procedimiento del relevamiento de datos

El muestreo para relevar los datos se realizó por medio de una encuesta presencial y mediante un cuestionario estructurado e impreso. El equipo de encuestadores estuvo formado por una estudiante de la carrera de licenciatura en Ciencias de la Comunicación, para documentar la experiencia y por estudiantes de las carreras de Ingeniería Electromecánica e Industrial, con conocimientos de estadística y entrenamiento para entrevistar.

[9] El universo constituye la población de la localidad de San Luis, sobre esa población se identificaron dos segmentos de interés: los niños de la Escuela ARA y vecinos que residen en el entorno inmediato a la escuela.

Fotografía 2. Estudiante universitario con uno de los niños encuestados. Fuente: Material fotográfico propio.

Previamente a la experiencia se realizó una prueba piloto para evacuar dudas, eliminar errores y eventuales correcciones al cuestionario. En el proceso los encuestadores tomaron al azar grupos de 7 estudiantes en todos los niveles y que asisten diariamente a la escuela. Cada encuestado fue entrevistado en forma individual.

Fotografía 3. Parte del equipo explicando la metodología para la toma de muestra en la experiencia. Fuente: Material fotográfico propio.

En la Figura 1 se presenta la ficha técnica del segmento correspondiente a los estudiantes, con datos de la muestra, el tipo y tamaño, así como el error estadístico.

FICHA TECNICA	
Universo	Alumnos de primaria turno mañana de la escuela Submarino ARA de San Luis
Muestra	Presencial con cuestionario impreso.
Tipo de Muestreo	Aleatorio, sobre población finita con universo de 275 estudiantes.
Tamaño de Muestra	78 casos efectivos
Fecha de toma de muestra	Viernes 3 de diciembre de 2021
Cobertura	Sobre el total de 11 grados de nivel primaria
Errores estadísticos	Menores a 0.10 con una confiabilidad del 95% de acuerdo a la distribución hipergeométrica para población finita

Figura 1. Ficha técnica de alumnos.

La ficha técnica del universo de vecinos de la Escuela, con datos del tipo de muestra y tamaño de esta, se consigna en la Figura 2.

FICHA TECNICA	
Universo	Vecinos colindantes de la escuela Submarino ARA de San Luis
Muestra	Presencial con cuestionario impreso.
Tipo de Muestreo	Aleatorio, sobre población finita con universo de 275 estudiantes.
Tamaño de Muestra	47 casos efectivos
Fecha de toma de muestra	Viernes 3 de diciembre de 2021
Cobertura	Sobre el total de 80 viviendas que son las más próximas a la escuela.
Errores estadísticos	Menores a 0.10 con una confiabilidad del 95% de acuerdo a la distribución hipergeométrica para población finita

Figura 2. Ficha técnica de vecinos.

RESULTADOS

La muestra de estudiantes primarios se organiza por 53,8% del género femenino y 46,2%, del masculino; en cuanto a las edades el 34,6% tenían entre 6 y 9 años y el 65,4% entre 10 y 13 años, lo que constituye valores consistentes con los datos poblacionales según el censo 2010 (INDEC, 2022). Con esta información de composición de género y edad se representa el gráfico circular por género y edad en la Figura 3.

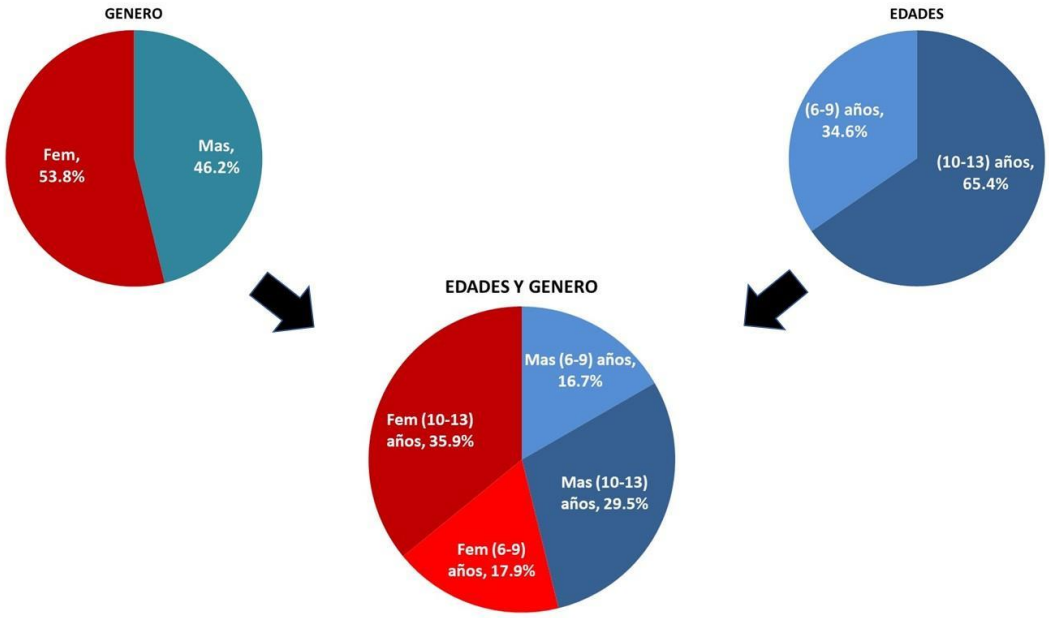


Figura 3. Composición de la muestra por género y edad.

En cuanto a los modos de transporte, el 39,7% llega por el servicio urbano de pasajeros y 35,9% en automóvil. Es de destacar que 17,9% llega caminando, lo cual no es una cifra menor, teniendo en cuenta que se está analizando el riesgo que tiene el peatón. Este valor indica a prima facie el valor del riesgo potencial de la población bajo estudio.

Es visible la particularidad que presenta el segmento de las niñas más pequeñas (entre 6 y 9 años) que son las que mayormente llegan caminando (35,7%) como se aprecia en la Figura 4.

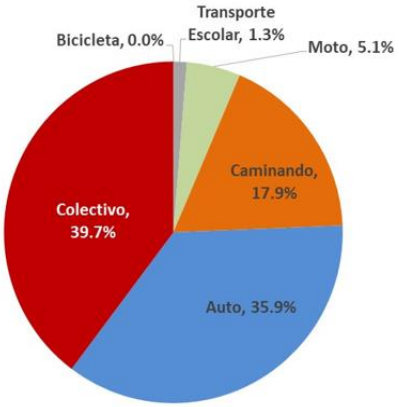


Figura 4: Distribución de los medios de transportes utilizados para llevar los niños a la escuela.

En la Tabla 5 se consigna el tipo de transporte utilizado para el transporte de los alumnos y la composición etaria y género.

Tabla 5: Tipo de transporte y composición etaria.

		EDAD Y GENERO			
		Mas (6-9) años	Mas (10-13) años	Fem (6-9) años	Fem (10-13) años
		% del N de columna	% del N de columna	% del N de columna	% del N de columna
¿En que te traen a la escuela?	Auto	53,8%	39,1%	42,9%	21,4%
	Transporte Escolar	0,0%	0,0%	0,0%	3,6%
	Colectivo	30,8%	47,8%	14,3%	50,0%
	Moto	0,0%	4,3%	7,1%	7,1%
	Bicicleta	0,0%	0,0%	0,0%	0,0%
	Caminando	15,4%	8,7%	35,7%	17,9%
	Total	100,0%	100,0%	100,0%	100,0%

Se destaca que, aproximadamente, la mitad (44,9%) de la población infantil considera que es difícil cruzar la ruta, sobre todo en la franja de las niñas de mayor edad (10 a 13 años) como se indica en la Figura 5. Esta dificultad percibida se condice con la significativa mayoría (65,4%) que afirma que el tránsito vehicular es intenso (ver Figura 6) y de alta velocidad, que se muestra en la Figura 7.

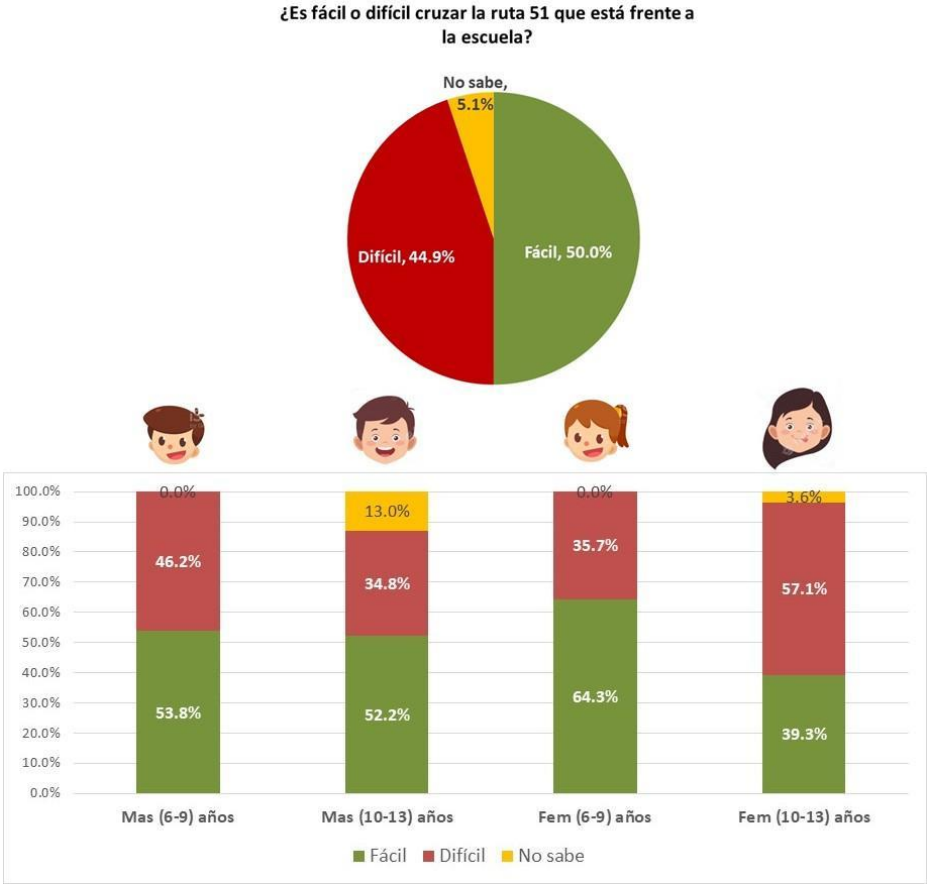


Figura 5.Dificultad de cruce en la vía nacional frente a la escuela.

Figura 6. Densidad de tránsito vehicular sobre ruta nacional.

Figura 7. Percepción de la velocidad vehicular sobre ruta nacional.

Sin embargo, es bajo el temor al riesgo, el 59% de los niños afirma no tener *nada de miedo*, como se muestra en la Figura 8, lo que revela hay conciencia de denso tránsito vehicular que dificulta la movilidad de alumnos de a pie, peligro al que no se le teme y esto es un indicador que acrecienta el riesgo potencial observado.

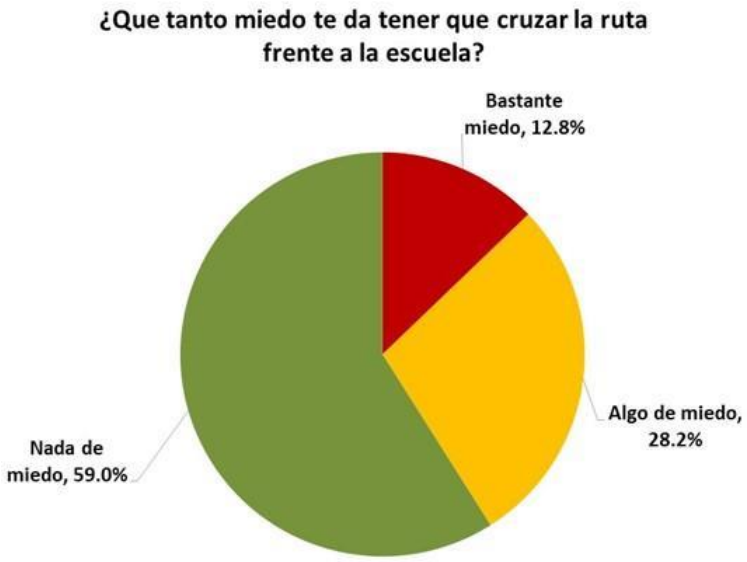


Figura 8. Percepción de temor por la población de alumnos de nivel primario durante el cruce por la ruta nacional.

Cabe resaltar que el 65,4% de los encuestados describieron que, al momento de cruzar la ruta para la entrada hacia la escuela, deben esperar que los automovilistas pasen y luego efectuar el cruce debido al modo de conducción, en el que no se respeta la prioridad del peatón. Esta situación se refleja en la **Figura 9.** Esta conducta se verifica tanto en la ruta nacional como en las calles aledañas a la escuela.

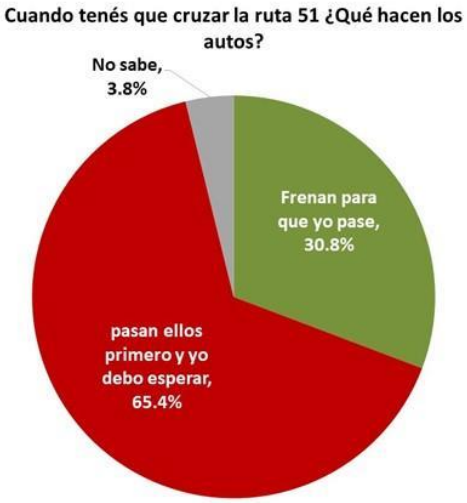


Figura 9. Práctica de los conductores de vehículos frente al paso del peatón sobre ruta nacional.

Así también, el 56,4% de los encuestados manifestaron que la velocidad practicada por los vehículos es elevada. La velocidad de vehículos frente a la escuela está restringida a un valor máximo de 20 km/h.

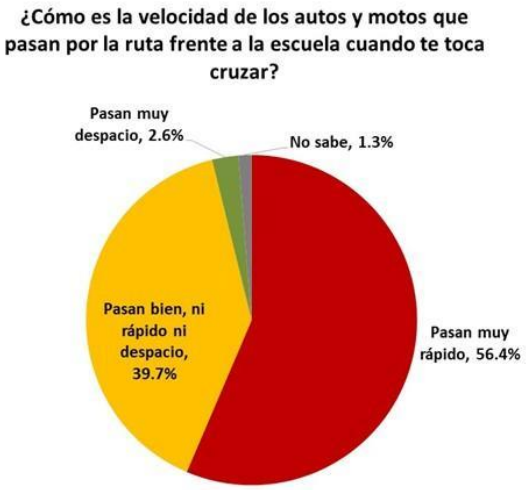


Figura 10. Percepción de la población infantil respecto de la velocidad de tránsito de vehículos.

La muestra de vecinos refleja que el 63,8% son de género femenino y el 36,2% masculino. En cuanto a las edades, el 51,1% son menores de 39 años y el 48,9% son mayores de 39 años.

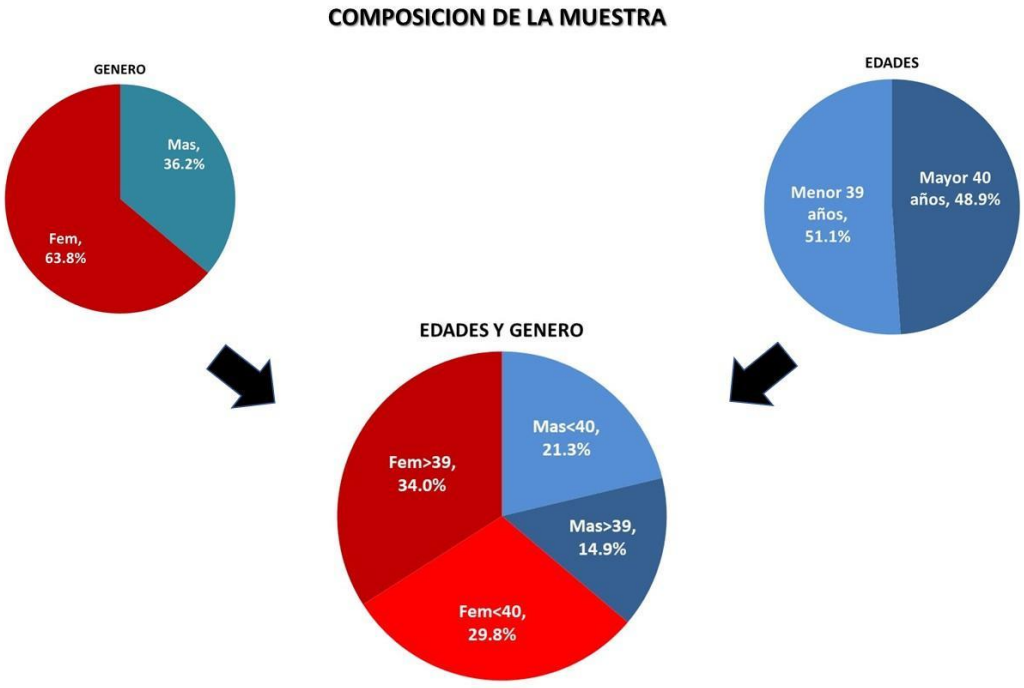


Figura 11: Composición en edad y género de la muestra de vecinos.

Por un lado, los vecinos manifiestan que el cruce de la ruta 51 es peligroso como se indica en la Figura 12.

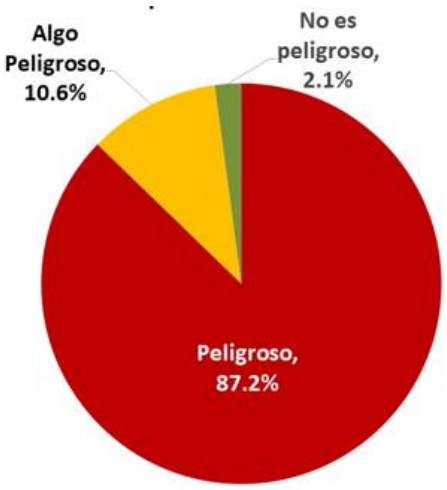


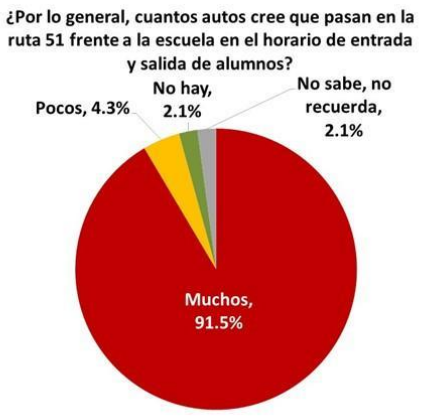
Figura 5. Distribución porcentual sobre el peligro de cruzar la ruta 51.

Por el otro lado, los vecinos señalan que se acrecienta la dificultad de cruce en los horarios de ingreso y egreso al establecimiento, percepción compartida, entre los que tienen y no tienen hijos que asisten a la escuela, información que se refleja en la Tabla 6.

Tabla 6. Horarios de mayor dificultad para el cruce de la ruta nacional.

		¿Tiene Hijos o algún familiar que asiste a la escuela ARA?	
		Sí	No
		% del N de columna	% del N de columna
¿En que horas cree ud que se pone más complicado cruzar la ruta 51 que esta frente a la escuela ARA?	A la madrugada	25,0%	3,2%
	A la entrada de alumnos a la escuela ARA	43,8%	48,4%
	A la salida de alumnos del ARA	31,3%	45,2%
	A la tarde	0,0%	0,0%
	A la noche	0,0%	0,0%
	No sabe	0,0%	3,2%
	Total	100,0%	100,0%

No obstante, hay conciencia de un intenso tránsito y alta velocidad que causa temor, que se muestra en la Figura 14, y este temor es mayor entre quienes tienen hijos que acuden a la escuela, como se presenta en la



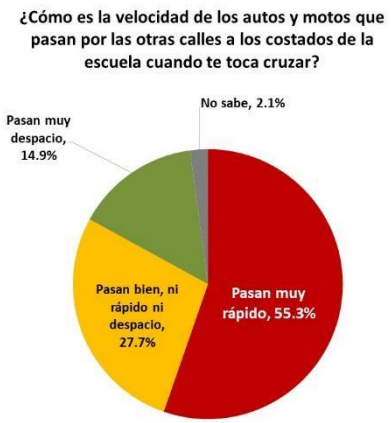


Tabla 7. Sensación de temor en el cruce de la ruta nacional.

		¿Tiene Hijos o algún familiar que asiste a la escuela ARA?	
		Sí	No
		% del N de columna	% del N de columna
¿Que tanto miedo le da tener que cruzar la ruta 51 frente a la escuela en horario de salida y entrada de alumnos?	Bastante miedo	50,0%	38,7%
	Algo de miedo	37,5%	32,3%
	No da miedo	12,5%	25,8%
	No sabe	0,0%	3,2%
	Total	100,0%	100,0%

No hay conciencia de la prioridad del peatón por parte de conductores, indican los vecinos en un elevado porcentaje (91,5%) como se indica la Figura 15.

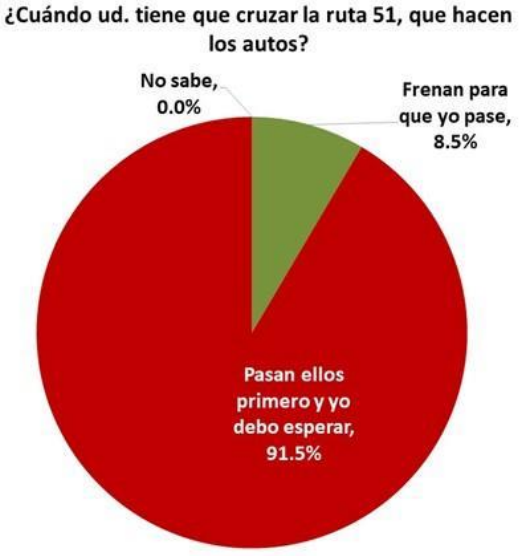


Figura 15: Práctica de los conductores de vehículos frente al paso del peatón sobre ruta nacional.

Por último, los vecinos creen, en un porcentaje que sumado da 78,7%, de regular a malo el control que ejerce la policía sobre los aspectos de la SV en el entorno escolar, situación que se refleja en la Figura 16.

Figura 16. Percepción de los vecinos sobre los controles de tránsito en la ruta nacional.

CONCLUSIONES

El estudio permitió la obtención de información a través de datos recolectados mediante la opinión de alumnos y vecinos colindantes a la escuela, por medio de una encuesta, que medía entre otras variables, la percepción del riesgo de cruce por una ruta nacional cuya traza pasa por enfrente de la escuela. La experiencia, además, permitió aplicar de manera práctica los conocimientos de estadística y de las ciencias sociales de los estudiantes universitarios, en un problema real que atañe al medio social, el cual es la SV en entornos escolares.

El presente estudio reflejó que los alumnos de la escuela perciben el peligro de manera diferenciada de acuerdo con el género. Se detectó que las niñas de la franja de menor edad consideran al cruce de la calle enfrente de la escuela como menos peligroso, con respecto al resto de los alumnos encuestados, en un porcentaje del 64,3%.

Las niñas en la franja de mayor edad (10 a 13 años) presentan una respuesta inversa a las de menor edad, mencionadas anteriormente, advirtiendo que existe dificultad y existencia de peligro en el paso de la ruta nacional. Es la franja que percibe el peligro en mayor porcentaje que el resto.

Se observa que, en la franja de mayor edad, para ambos géneros, desconocen si se trata de un peligro (13% en varones y 3,6% en niñas), lo que constituye un porcentaje nada despreciable y consideran no conocer la existencia de riesgo en la seguridad vial.

Los vecinos consideran, en un alto valor porcentual (más del 87%), que existe una condición que genera temor en el cruce de la ruta nacional durante los horarios de entrada y salida de las clases, y no depende de la condición de que el vecino tenga o no hijos asistiendo a la escuela. Esta conclusión está sustentada con la prueba estadística CHI^2, que arrojó un valor p de 0,57.

Respecto a la percepción de dificultad para cruzar la ruta frente a la escuela, los horarios de entrada y salida de clases son considerados los más complicados, tanto por los vecinos con o sin hijos asistiendo a dicha escuela. En el caso del horario de la madrugada, este es considerado peligroso sólo por aquellos vecinos con hijos que son alumnos de esa escuela (ver Tabla 6). Esto se respaldó con la prueba estadística CHI^2, que arrojó un valor p de 0,082.

Estos hallazgos revelan la importante y urgente necesidad de educación y concientización de manera continua en temas de seguridad vial, dirigida hacia los alumnos de la escuela.

Con el desarrollo de esta experiencia, se observa que existe una percepción variable y de menor conciencia de peligro en la población infantil escolar; a diferencia de los adultos vecinos que advierten el riesgo de la seguridad vial en la comunidad inmediata al entorno educativo.

El análisis realizado permite recomendar la profundización de la educación vial de los niños para lograr una mayor cognición sobre los riesgos de la movilidad, con un enfoque sobre los usuarios vulnerables viales y encomendar la realización de acciones para mejorar los aspectos de la seguridad vial mediante campañas impulsadas por los entes locales -escuela, municipio y policía-. El objetivo de éstas consiste en motorizar cambios positivos en la conducta de los actores motorizados de la vía y la sensibilización de la problemática de la seguridad vial, en especial de niños, en la comunidad toda.

Así también, se recomienda el control policial debido a las condiciones inseguras de esta vía semiurbana, hasta alcanzar mejoras factibles en la infraestructura, y que contribuye a su vez a concienciar a los conductores de vehículos.

REFERENCIAS

Aparicio, F., & Arenas, B. (2014). Factores y capas de análisis del accidente de tráfico. *Securitas Vialis, 18*, 125-149.

Devore, J. L. (2008). *Probabilidad y estadística para ingeniería y ciencia* (7ª Edición ed.). Brooks/Cole.

Montgomery, D. (1991). *Diseño y análisis de experimentos.* México: Grupo Editorial Latinoamericano.

OECD-The World Bank. (2020). *Panorama de la Salud: Latinoamérica y el Caribe 2020.* París: OECD Publishing, 2020.

Organización Mundial de la Salud. (2013). *Informe Sobre la Salud en el Mundo 2013: Investigaciones Para Una Cobertura Sanitaria Universal.* OMS.

Peña Sánchez de Rivera, D. (2001). *Fundamentos de Estadística.* Alianza Editorial.

Pérez, V. (2008). La aportación de la familia como agente educador básico. En V. Pérez, & M. Pardo, *Educación y Seguridad Vial. La aportación de los agentes sociales a la movilidad segura* (págs. 141-170). España: Etrasa.

Quispe Limaylla, A. (2013). *El uso de la encuesta en las ciencias sociales.* Madrid: Ediciones Díaz de Santos.

IV Jornadas Internacionales de Estadística Aplicada

Proceso integral de investigación estratégica, metodología de ajuste y eliminación del efecto de autocorrelación de los errores en los modelos de regresión lineal aplicados a los trackings. Caso de las elecciones de Nayarit 2021 en contexto de Pandemia COVID-19

Rodríguez, Héctor Iván; Montenegro Ibarra, Jorge Aníbal

Datos de contacto:

heivro@gmail.com - Cel +5493874129731

jamixtlan71@hotmail.com – Cel +5213112530658

RESUMEN

En este documento se describe un procedimiento para desarrollar un conjunto de técnicas y tipos de muestreo basados en la estadística aplicada, orientado a obtener información necesaria para el diseño de la Estrategia de campaña, su seguimiento y monitoreo de control, de tal manera que permita realizar, en tiempo y forma, las modificaciones necesarias de acuerdo con el desarrollo y dinámica de una contienda electoral. También se desarrolla la metodología para ajustar los trackings y cómo deben eliminarse los efectos de autocorrelación de los errores para una mejor precisión. Dicho procedimiento se aplicó en las elecciones estatales de Nayarit realizadas el 6 de junio del 2021 en contexto de Pandemia generada por el COVID-19 con significativa precisión en el logro de los objetivos planteados.

Palabras Clave: Muestreo, Estrategia, COVID 19, Autocorrelación, Tracking, Nayarit, confianza.

INTRODUCCIÓN

El 6 de junio de 2021 fueron las elecciones para elegir gobernador en el Estado de Nayarit – México. El presente trabajo describe **el proceso de investigación estratégica aplicada**, que refleja cómo se realizaron y aplicaron las investigaciones estadísticas orientativas, para el trazado de la estrategia y el seguimiento de la campaña electoral.

METODOLOGÍA

PROCESO DE INVESTIGACIÓN ESTRATÉGICA

Comprendió tres etapas, a saber:

1) Investigación Estratégica
2) Aplicación de la estrategia
3) Medición de seguimiento

Investigación estratégica: Comprendió el diseño investigativo orientado a obtener el conocimiento del escenario donde se desarrollaría la campaña electoral, y las variables relacionadas con la decisión de voto. Fue importante conocer a fondo el Humor Social, ya que este es el que determina el tono de la campaña. También fue importante determinar para cada candidato los atributos, que fueron tanto formadores de imagen, como motivadores del voto. Esta etapa es la que permitió el diseño estratégico de la campaña y definió el arranque de ésta.

Aplicación de la estrategia: De acuerdo con el Humor Social, el análisis coyuntural de la investigación, el conocimiento de la dinámica de la imagen de los candidatos y determinados cuales son los atributos motivadores del voto, se diseñó la estrategia y planificación de las acciones de comunicación para instalar "*El Mensaje"* en el electorado.

Medición de Seguimiento: Una vez implementada la estrategia en campo se procedió a realizar el seguimiento con el objetivo de tener el control de campaña, de tal manera de ir ajustando la estrategia y el mensaje, de acuerdo con las variaciones de intención de voto. El seguimiento se realizó mediante un Tracking, complementado con tres encuestas cara a cara, estas funcionan como reguladoras para el ajuste del tracking, y la medición de posibles causas observadas en las variaciones de la intención de voto y temas coyunturales.

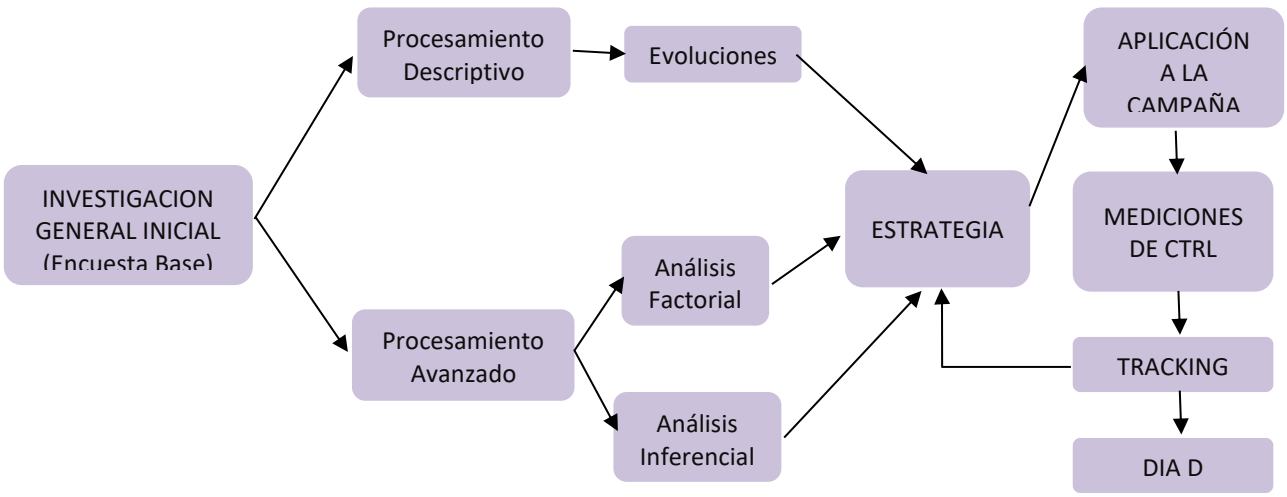


Ilustración 1

DESARROLLO

IMAGEN E INTENCIÓN DE VOTO AL INICIO DE LA CAMPAÑA

Si bien el candidato Navarro Quintero posee un buen punto de arranque, se observó gran desconocimiento de los otros candidatos lo que significa un potencial crecimiento positivo de ellos. La intención de voto de Navarro Quintero, está en el orden de la imagen positiva, y con un nivel importante de indecisos de 39.3%, lo que indicaba una elección abierta con posibilidades para cualquier candidato.

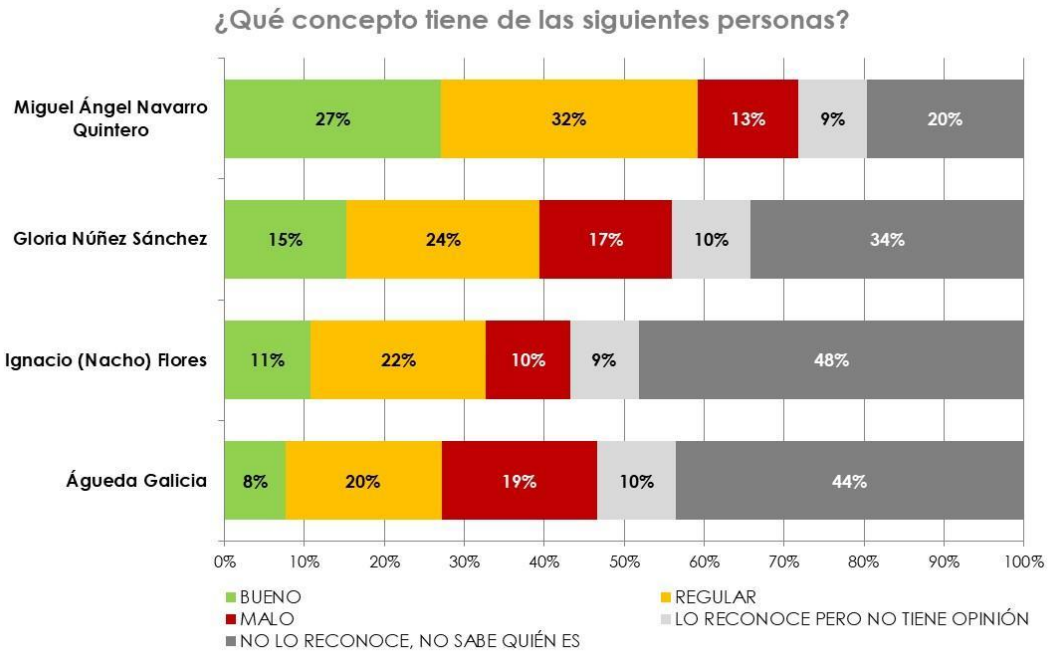


Gráfico 1

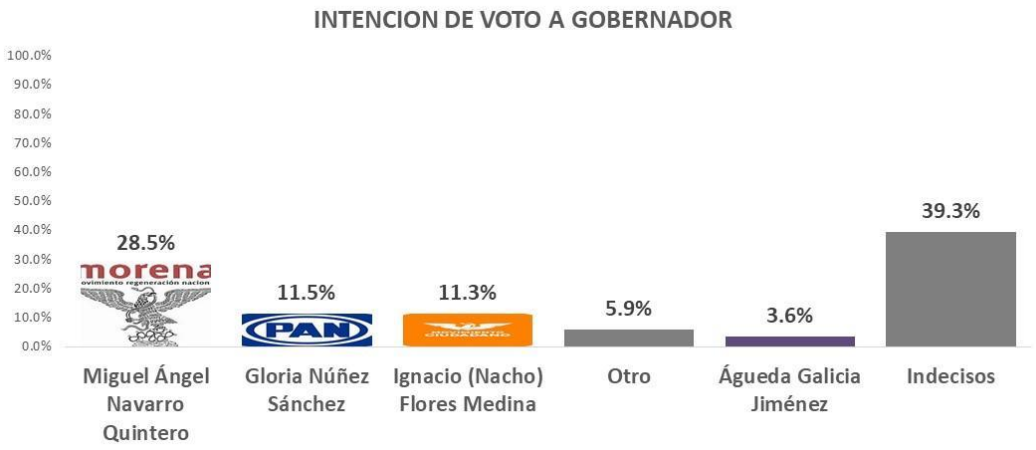


Gráfico 2

TRACKING DE SEGUIMIENTO Y CONTROL

Obedeciendo a razones impuestas por la pandemia generada por el COVID-19, que dificultó la realización del muestreo diario cara a cara, fue necesario diseñar una muestra robótica que necesitó ser calibrada y corregida.

El proceso fue complejo, ya que la muestra robótica no solo se alejaba de las bondades del contacto cara a cara, sino que no permitía el control exacto de las cantidades de respuestas por municipio que satisfagan las condiciones de aleatoriedad, por lo que se necesitó calcular factores de ponderación de acuerdo con la población del estado, para ajustar la representatividad y poder obtener la muestra corregida a nivel estado.

Esta muestra corregida por factores de ponderación tuvo además tres ajustes adicionales durante el mes de aplicación del tracking, mediante muestras cara a cara realizadas como mediciones de contraste, para ajustar sesgos de la muestra robótica. Entre estos ajustes, uno de los factores importantes fue determinar la baja intención de asistencia a votar por temor a la pandemia, lo que significó filtrar la muestra de abstencionistas. Por último, queda corregir los efectos de autocorrelación debido a que los trackings usan la media móvil. Eliminado este efecto se presenta el tracking a nivel estado, sobre los cuales se tomaron las decisiones diarias de campaña. (Ver ilustración 2).

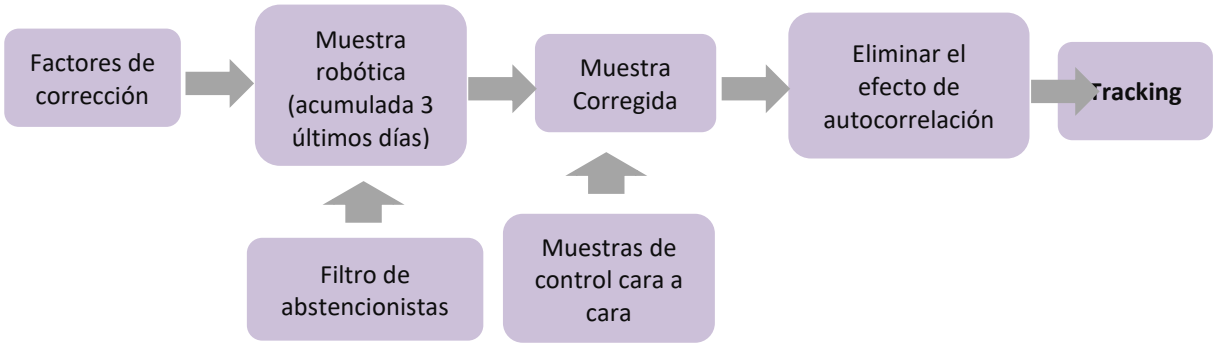
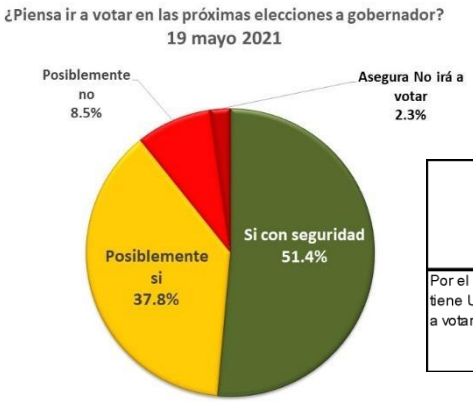


Ilustración 2



		¿Piensa ir a votar en las próximas elecciones a gobernador?				
		Si con seguridad	Posiblemente si	Posiblemente no	Asegura No irá a votar	Total
		% del N de tabla	% del N de tabla	% del N de tabla	% del N de tabla	% del N de tabla
Por el Covid, ¿Tiene o no tiene Ud. temor de salir a ir a votar?	Si	12.3%	9.9%	1.7%	1.1%	25.0%
	Un poco	13.3%	17.6%	4.4%	.2%	35.5%
	No	25.8%	10.3%	2.4%	.9%	39.5%
	Total	51.4%	37.8%	8.5%	2.3%	100.0%

Gráfico 3

Diseño Técnico del Tracking

El objetivo del Tracking no solo fue de medir la evolución del voto, sino también de poder determinar las posibles causas de las variaciones observadas en la intención de voto, en función de los atributos formadores establecidos en la muestra base inicial. Por lo tanto, se diseñó un modelo denominado **Tracking de dos pasos** para poder tener valores por región. En este caso se dividió al Estado de Nayarit en tres regiones, **Tepic; Región Norte y Región Sur**. El primer valor del Tracking contiene las tres regiones, pero es insuficiente el tamaño muestral para segmentar por cada región, por lo que en el segundo día de Tracking se reemplaza una región por otra, pero permite tener datos de dos regiones acumuladas, de tal forma que se obtiene como beneficio lo siguiente:

1) Tamaño muestral adecuado para lectura de tracking para la zona.
2) Anticipación del impacto de la zona reemplazada en los resultados con dos días de anticipación.

De esta manera a cada día de tracking general del Estado, acompaña un tracking regional espaciado con dos días. Ver gráficos más adelante.

Se muestra como ejemplo los primeros 15 días con 13 valores de tracking que grafica el diseño mencionado. (Ver ilustración 3)

DIA	FECHA	TRACK 1	TRACK 2	TRACK 3	TRACK 4	TRACK 5	TRACK 6	TRACK 7	TRACK 8	TRACK 9	TRACK 10	TRACK 11	TRACK 12	TRACK 13
1	30-abr.	NORTE 1												
2	1-may.	TEPIC 1	TEPIC 1											
3	2-may.	SUR 1	SUR 1	SUR 1										
4	3-may.		NORTE 2	NORTE 2	NORTE 2									
5	4-may.			TEPIC 2	TEPIC 2	TEPIC 2								
6	5-may.				SUR 2	SUR 2	SUR 2							
7	6-may.					NORTE 3	NORTE 3	NORTE 3						
8	7-may.						TEPIC 3	TEPIC 3	TEPIC 3					
9	8-may.							SUR 3	SUR 3	SUR 3				
10	9-may.								NORTE 4	NORTE 4	NORTE 4			
11	10-may.									TEPIC 4	TEPIC 4	TEPIC 4		
12	11-may.										SUR 4	SUR 4	SUR 4	
13	12-may.											NORTE 5	NORTE 5	NORTE 5
14	13-may.												TEPIC 5	TEPIC 5
15	14-may.													SUR 5

Ilustración 3

Factores de Ponderación para el tracking a nivel Estado y Regional

Se detalla en la ilustración 4 los factores de ponderación usados para obtener la muestra corregida a nivel estado y por región.

Código	Municipio	Población total 2020	PESO
1	Acaponeta	37,232	3
2	Ahuacatlán	15,393	1.2
3	Amatlán de Cañas	11,536	1
4	Compostela	77,436	6.3
5	Huajicori	12,230	1
6	Ixtlán del Río	29,299	2.4
7	Jala	19,321	1.6
8	Xalisco	65,229	5.3
9	Del Nayar	47,550	3.8
10	Rosamorada	33,567	2.7
11	Ruíz	24,096	2
12	San Blas	41,518	3.4
13	San Pedro Lagunillas	7,683	0.6
14	Santa María del Oro	24,911	2
15	Santiago Ixcuintla	93,981	7.6
16	Tecuala	37,135	3
17	Tepic	425,924	34.5
18	Tuxpan	30,064	2.4
19	La Yesca	13,719	1
20	Bahía de Banderas	187,632	15.2
		1,235,456	100

FACTORES DE PONDERACION POR REGION		
ZONA SUR	PESO	PESO REL
Bahía de Banderas	15.2	41.5%
Compostela	6.3	17.2%
Xalisco	5.3	14.5%
Resto Norte	9.8	26.8%
TOTAL NORTE	36.6	100.0%
	PESO	PESO REL
TEPIC	34.5	100.0%
ZONA NORTE	PESO	PESO REL
Santiago Ixcuintla	7.6	26.3%
Del Nayar	3.8	13.1%
Acaponeta	3	10.4%
Rosamorada	2.7	9.3%
San Blas	3.4	11.8%
Tecuala	3	10.4%
Resto sur	5.4	18.7%
TOTAL SUR	28.9	100.0%

Desarrollo del Tracking

FICHA TECNICA	
Universo	Residentes de Nayarit, mayores de 18 años.
Muestra	Robótica, aleatoria.
Tipo de Muestreo	Polietápica, por conglomerados, estratificada.
Tamaño de Muestra	1734 casos efectivos acumulados en los 3 últimos días
Responsable del Procesamiento e informe	Mag. Ing. HECTOR IVAN RODRIGUEZ
Fecha Inicio	30 de abril de 2021
Fecha toma última muestra	3 al 5 de Junio de 2021
Cobertura	Estado de Nayarit: 20 municipios
Errores estadísticos	Menores al 5% con una confiabilidad del 95% de acuerdo a la distribución Binomial aproximada a la normal

Ajustes Para Mejorar La precisión del Tracking

Los problemas que se presentaban ante un tracking alimentado por una muestra robótica eran los siguientes:

1) **Muestreo no aleatorio**: Si bien las llamadas automáticas se disparaban al azar, tenía la restricción que no todo el electorado disponía de celular o teléfono para recibir llamadas y poder contestar así la encuesta, además las bases de datos conseguidas no incluían al 100% de la población del estado.

2) **Forma de consulta robótica**: No estaba probada su eficacia.

3) **Autocorrelación de los errores**: Debido a que el tracking se conforma de una muestra con promedio móvil con paso 3 días, es decir si la muestra arranca el día 1, entonces el primer resultado se presenta el día 3 acumulando las muestras de los días 1, 2 y 3. El resultado del día 4 se compone por las muestras de los días 2, 3 y 4. Y así sucesivamente, es decir que se va reemplazando el primer tercio de la muestra por el muestreo nuevo de cada día. Esta técnica se realiza para suavizar las variaciones observadas por el error estadístico y no estadístico, **pero introduce correlación entre las observaciones, ya que cada valor no es independiente del anterior porque contiene valores comunes**. Es conocido que esta correlación produce distorsión en las tendencias calculadas, y por lo tanto deben ser corregidas. Se muestra a continuación este efecto y su corrección.

Introducción Técnica

Desde el punto de vista estadístico la autocorrelación de una variable, es la dependencia consigo misma a lo largo del tiempo. En los datos de series temporales, como lo son los trackings, un problema habitual es la presencia de autocorrelación de las perturbaciones.

Las consecuencias más importantes de la existencia de la autocorrelación son:

a) Las estimaciones de los parámetros en modelos de regresión lineal utilizando mínimos cuadrados ordinarios dejan de ser eficientes.

b) La inferencia estadística basada en la matriz de varianzas y covarianzas del estimador por mínimos cuadrados ordinarios será errónea.

Un modelo de frecuente aplicación en los trackings de las campañas políticas es el siguiente modelo general de regresión lineal**:**[10]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + u_i$$

El cual suele ser reducido a la siguiente expresión de aplicación:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + u$$

Que obedece al modelo de la recta cuando los β_2 … β_n valen cero y a una expresión polinómica en los otros casos.

En este modelo:

Y***: Es la intención de voto calculada por el modelo para un determinado día.***

X: Es el número correspondiente al día de tracking.

β_i***: Son parámetros determinados para el modelo propuesto.***

u: Son los errores aleatorios.

Los modelos clásicos de regresión lineal parten del importante supuesto de no existencia de autocorrelación entre los errores u, lo cual no se cumple en los modelos de tracking construidos con la acumulación de tres días de muestreo, y reemplazo sucesivo del último día por el primero de la serie. Este reemplazo se realiza para tener una curva suavizada, que permita detectar cambios debido a la intención de voto y no del propio error aleatorio, pero introduce autocorrelación.

Naturaleza de la Autocorrelación

La autocorrelación se da cuando se observa correlación entre los errores u_1, u_2, u_3, …, u_i. Frecuentemente se observan en series de tiempo. Un tracking no solo es una serie de tiempo, sino también un modelo donde los valores observados no son independiente del dato anterior, sino que dependen de los valores de las dos últimas mediciones para un día determinado.

$$Y_i = f(X_i\ X_{i-1}\ X_{i-2})$$

El modelo de nuestro tracking presenta **rezagos**, es decir que cada valor no es independiente del valor anterior, por lo tanto, el modelo a aplicar es el de **autorregresión**[11]

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + u_t$$

Esta ecuación contempla la dependencia del valor de cada medición dependiendo de la medición anterior.

[10] Mendenhall, W. B. (2010). *Introducción a la probabilidad y estadística.* Santa Fe, D.F., México: Cencage Learning.

[11] Walpole, R. E. (2012). *Probabilidad y estadística para ingeniería y ciencias.* México: Pearson Education.

Y_t: *Es la intención de voto medida en el día t.*

$\boldsymbol{X_t}$: *Es el número correspondiente al día t de tracking (fecha).*

β_i: *Son parámetros determinados para el modelo propuesto.*

Y_{t-1}: *Es la intención de voto medida el día anterior al día t.*

$\boldsymbol{u_t}$: *Es el error aleatorio en el día t.*

Verificación de la existencia de autocorrelación

Una de las formas prácticas y sencillas de verificar la existencia de autocorrelación entre las perturbaciones, consiste en graficar los errores en función del tiempo y observar la existencia de ciertos patrones.

Determinación del tipo de correlación

Para saber si la correlación es positiva o negativa, se grafica cada error en función de su valor anterior, es decir, u_t contra u_{t-1} . Si el gráfico resultante muestra tendencia con pendiente positiva, indica correlación positiva y viceversa. Recuérdese que una correlación positiva del error, significa que si aumenta el error anterior aumenta el error presente y viceversa. Una correlación negativa significa que si aumenta el error anterior, disminuye el error presente y viceversa.

Se presentan en la Tabla 1 los cálculos de los errores residuales, para los tres principales candidatos a gobernador y sus correspondientes gráficos.

DIA	Miguel Ángel Navarro Quintero			Ignacio (Nacho) Flores Medina			Gloria Núñez Sánchez		
	Int. Voto	Estimación Lineal del voto	RESIDUALES	Int. Voto	Estimación Lineal del voto	RESIDUALES	Int. Voto	Estimación Lineal del voto	RESIDUALES
2-May	48.9%	46.9%	-2.0%	14.8%	16.4%	-1.7%	15.9%	18.1%	-2.2%
3-May	48.5%	46.9%	-1.6%	14.9%	16.5%	-1.6%	17.0%	18.1%	-1.1%
4-May	50.0%	47.0%	-3.1%	14.6%	16.6%	-2.0%	17.8%	18.1%	-0.3%
5-May	48.0%	47.0%	-1.0%	15.0%	16.6%	-1.6%	19.0%	18.1%	0.9%
6-May	43.4%	47.0%	3.6%	18.2%	16.7%	1.5%	18.9%	18.1%	0.8%
7-May	44.0%	47.1%	3.0%	18.1%	16.8%	1.3%	18.7%	18.1%	0.6%
8-May	44.7%	47.1%	2.4%	18.9%	16.8%	2.1%	17.7%	18.1%	-0.4%
9-May	47.9%	47.1%	-0.7%	16.6%	16.9%	-0.3%	18.8%	18.1%	0.7%
10-May	45.7%	47.2%	1.4%	17.1%	17.0%	0.2%	19.2%	18.1%	1.1%
11-May	45.6%	47.2%	1.6%	17.3%	17.0%	0.3%	21.6%	18.1%	3.4%
12-May	46.7%	47.2%	0.6%	17.7%	17.1%	0.6%	20.6%	18.1%	2.5%
13-May	46.3%	47.3%	0.9%	18.0%	17.2%	0.9%	21.1%	18.1%	3.0%
14-May	47.0%	47.3%	0.3%	17.9%	17.2%	0.7%	17.6%	18.1%	-0.5%
15-May	45.5%	47.3%	1.9%	17.7%	17.3%	0.4%	16.3%	18.1%	-1.9%
16-May	47.2%	47.4%	0.2%	16.6%	17.4%	-0.8%	15.3%	18.1%	-2.8%
17-May	49.2%	47.4%	-1.8%	17.0%	17.4%	-0.4%	15.9%	18.1%	-2.2%
18-May	47.6%	47.4%	-0.1%	19.3%	17.5%	1.8%	17.4%	18.1%	-0.7%
19-May	46.3%	47.5%	1.2%	20.3%	17.6%	2.7%	18.0%	18.1%	-0.1%
20-May	48.3%	47.5%	-0.8%	18.9%	17.6%	1.3%	18.0%	18.1%	-0.2%
21-May	52.0%	47.5%	-4.5%	17.3%	17.7%	-0.4%	16.9%	18.1%	-1.2%
22-May	51.8%	47.6%	-4.2%	16.6%	17.8%	-1.2%	16.9%	18.1%	-1.2%
23-May	49.3%	47.6%	-1.7%	17.9%	17.8%	0.1%	16.6%	18.1%	-1.5%
24-May	46.7%	47.7%	0.9%	18.1%	17.9%	0.2%	18.4%	18.1%	0.3%
25-May	46.9%	47.7%	0.8%	18.4%	18.0%	0.4%	19.4%	18.1%	1.2%
26-May	47.6%	47.7%	0.1%	17.9%	18.0%	-0.2%	19.6%	18.1%	1.5%
27-May	48.3%	47.8%	-0.5%	17.2%	18.1%	-0.9%	18.8%	18.1%	0.7%
28-May	48.8%	47.8%	-1.0%	17.7%	18.2%	-0.5%	17.1%	18.1%	-1.0%
29-May	48.4%	47.8%	-0.6%	17.7%	18.2%	-0.5%	17.9%	18.1%	-0.3%
30-May	47.0%	47.9%	0.8%	17.0%	18.3%	-1.3%	16.4%	18.1%	-1.7%
31-May	47.7%	47.9%	0.2%	16.7%	18.4%	-1.7%	17.0%	18.1%	-1.1%
1-Jun	48.9%	47.9%	-1.0%	15.5%	18.4%	-2.9%	17.5%	18.1%	-0.6%
2-Jun	45.9%	48.0%	2.1%	18.1%	18.5%	-0.4%	20.4%	18.1%	2.3%
3-Jun	43.9%	48.0%	4.1%	19.0%	18.6%	0.4%	19.9%	18.1%	1.8%
4-Jun	47.9%	48.0%	0.1%	20.5%	18.6%	1.8%	18.9%	18.1%	0.8%
5-Jun	49.6%	48.1%	-1.6%	20.1%	18.7%	1.4%	17.4%	18.1%	-0.7%

Tabla 1. cálculos de los errores residuales.

Claramente el gráfico 4 de los errores residuales en función del tiempo muestra como patrón una sinusoide que evidencia falta de independencia, y el gráfico 5 de errores en función de su error anterior muestra tendencia positiva.

Gráfico 4

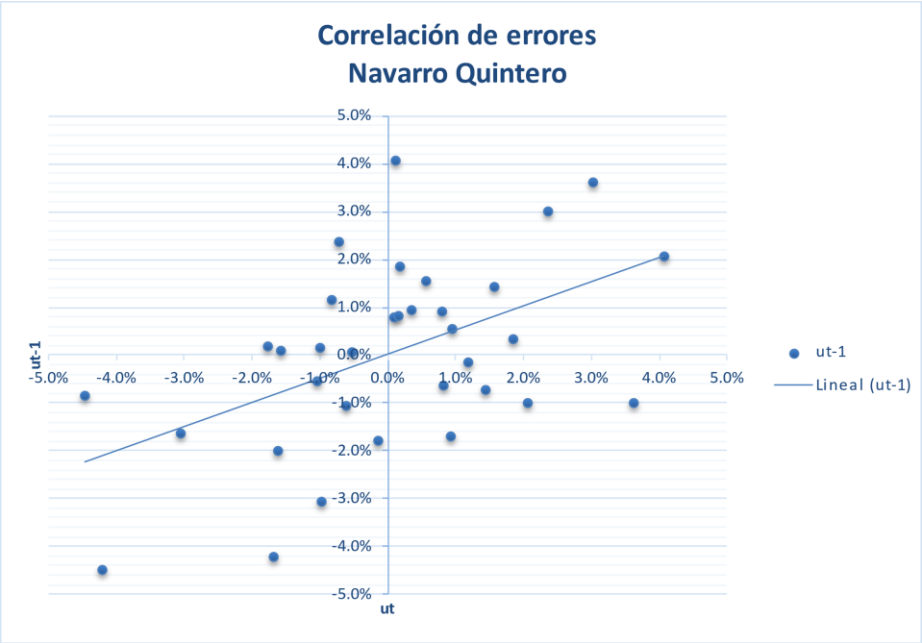


Gráfico 5

De manera similar se leen estos patrones en los gráficos 6 al 9 de los otros dos candidatos.

Gráfico 6

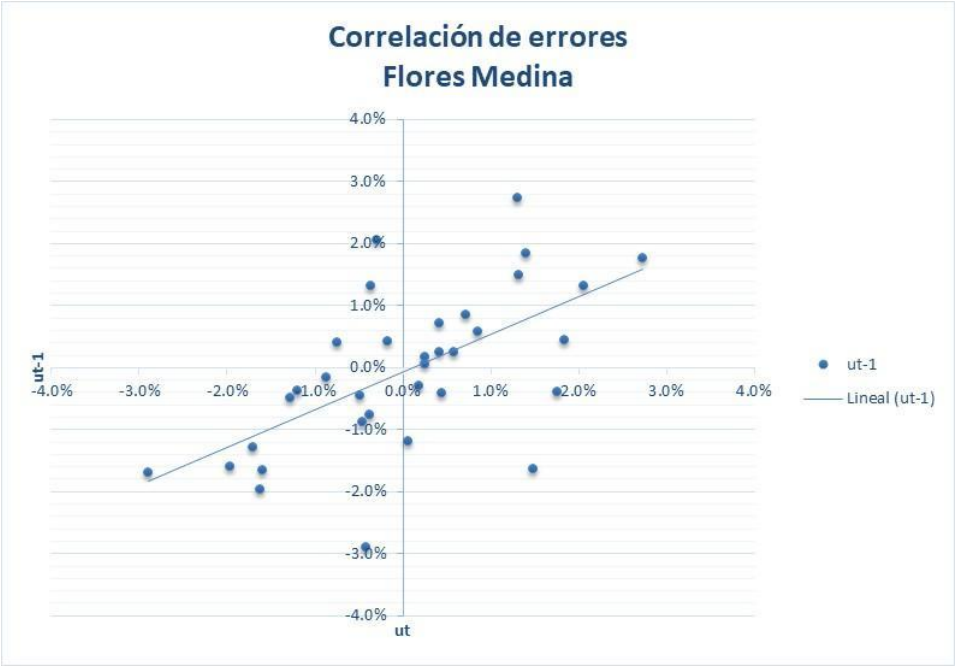


Gráfico 7

Gráfico 8

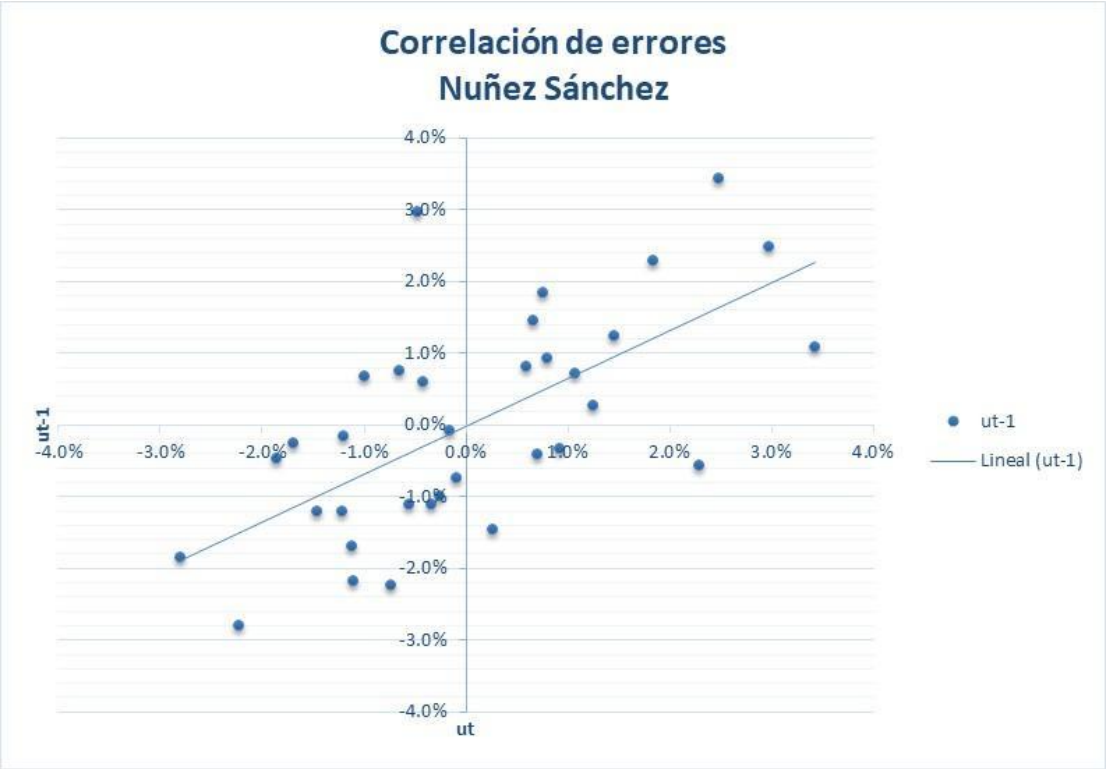


Gráfico 9

OBSERVACIONES: La evidencia de existencia de autocorrelación de los errores, impone la necesidad de realizar ajustes al modelo de regresión a usar en el tracking para estimar los intervalos del error donde estarán contenidos con mayor probabilidad los resultados electorales.

CORRECCIONES AL TRACKING: ELIMINACIÓN DE LA AUTOCORRELACIÓN

Una vez determinada la existencia de autocorrelación en el modelo, es de interés poder establecer el modelo correcto. Para ello lo que se hace es eliminar el problema de autocorrelación, de tal manera que nos permita obtener el modelo deseado.

Se usó el modelo de autocorrelación de primer orden.

$$u_t = \rho u_{t-1} + \varepsilon_t \quad (1)$$

Como modelo inicial se usó el de la línea recta siguiente:

$$Y_t = \beta_0 + \beta_1 X_t + u_t \quad (2)$$

Como planteamos que los errores ut siguen un modelo autorregresivo de primer orden, el objetivo es reformular al modelo de tal forma que los errores no estén correlacionados. Para ello se efectúa la siguiente operación:

Planteamos el modelo para el tiempo t-1

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + u_{t-1} \quad (3)$$

Multiplico ambos miembros por ρ

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho u_{t-1} \quad (4)$$

Restando (2) y (4)

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\beta_0 + \beta_1 X_t - \rho\beta_1 X_{t-1} + u_t - \rho u_{t-1}$$

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho\, X_{t-1}) + u_t - \rho u_{t-1} \quad (5)$$

Reemplazando en (5) por la ecuación (1)

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho\, X_{t-1}) + \varepsilon_t \quad (6)$$

De esta manera se han eliminado los errores correlacionados, quedando solamente la componente aleatoria del error independiente con valor esperado cero. Esta ecuación recibe el nombre de ***ecuación en diferencia generalizada****, y se puede escribir de la siguiente forma:*

$$\dot{Y}_t = \dot{\beta}_0 + \beta_1\, \dot{X}_t + \varepsilon_t \quad (7)$$

Donde:

$$\dot{Y}_t = Y_t - \rho Y_{t-1}$$

$$\dot{\beta}_0 = \beta_0(1-\rho)$$

$$\dot{X}_t = (X_t - \rho\, X_{t-1})$$

Debemos conocer **ρ** para poder resolver el problema, para ello estimamos el factor de autocorrelación con la siguiente ecuación:

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t{}^2} \quad (8)$$

PROCEDIMIENTO

1. Calcular mediante algún procedimiento $\hat{\rho}$

2. Construir los valores $\dot{Y}_t$; $\dot{X}_t$

3. Formular un modelo aceptable como $\dot{Y}_t = \dot{\beta}_0 + \beta_1\, \dot{X}_t + \varepsilon_t$

4. Realizar una regresión de las $\dot{X}_t$ *sobre las* $\dot{Y}_t$

5. obtener de la regresión los parámetros estimados $\dot{\beta}_0$ *y* β_1

6. Corroborar que los nuevos residuales para los valores $\dot{Y}_t$; $\dot{X}_t$ *son aleatorios.*

7. Obtener el modelo corregido $Y_t = \beta_0 + \beta_1\, X_t + \varepsilon_t$

Aplicación al tracking de Nayarit

La tabla 2 muestra los valores de la intención de voto acumulado de tres días para cada fecha y los errores de acuerdo con el modelo lineal planteado $Y_t = \beta_0 + \beta_1\, X_t + u_t$

Aplicando la fórmula (8) se obtiene un valor de $\hat{\rho}$*=0.5032 Lo que permite calcular los valores* $\dot{Y}_t$ $\dot{X}_t$ sobre las cual se realiza la regresión obteniéndose los siguientes coeficientes:

NAVARRO QUINTERO	
b0*=	0.229297651
b1=	0.000638154
b0=	0.461535574

$S_y = 0.017$ *(Desvío estándar de* Y_t *)*

Por lo tanto, el modelo corregido resulta**:** ***y = 0.0006x + 0.4615***

El Intervalo con una confianza del 95% aplicando la distribución T de Student para cada día resulta:

$$LS = Y_t + 1.96\ S_y$$

$$LI = Y_t - 1.96\ S_y$$

Puede verse en el gráfico 10 estos resultados y la línea de tendencia corregida para la intención de voto del candidato a gobernador del estado, el Dr. Navarro Quinteros. El modelo sugiere que el valor para el día 6 de julio en que se realizará el acto electoral tiende a 49% con un valor de medición del del día 5 de 49,63% lo que indica mayor probabilidad que el valor este comprendido en el intervalo superior del gráfico 10.

Miguel Ángel Navarro Quintero												
Fecha	Xt	Yt	Estimación Lineal del voto	ut	ut . ut-1	ut2	X*	Yt*	Estimación corregida del voto	Voto observado	LS	LI
2-May	2	48.89%	46.91%	1.98%					46.28%	48.89%	49.62%	42.95%
3-May	3	48.55%	46.94%	1.61%	0.03%	0.03%	1.99	23.95%	46.35%	48.55%	49.68%	43.01%
4-May	4	50.03%	46.97%	3.05%	0.05%	0.09%	2.49	25.60%	46.41%	50.03%	49.74%	43.07%
5-May	5	47.99%	47.01%	0.98%	0.03%	0.01%	2.99	22.82%	46.47%	47.99%	49.81%	43.14%
6-May	6	43.41%	47.04%	-3.63%	-0.04%	0.13%	3.48	19.26%	46.54%	43.41%	49.87%	43.20%
7-May	7	44.05%	47.07%	-3.03%	0.11%	0.09%	3.98	22.20%	46.60%	44.05%	49.94%	43.26%
8-May	8	44.74%	47.11%	-2.37%	0.07%	0.06%	4.48	22.57%	46.66%	44.74%	50.00%	43.33%
9-May	9	47.87%	47.14%	0.73%	-0.02%	0.01%	4.97	25.36%	46.73%	47.87%	50.06%	43.39%
10-May	10	45.74%	47.18%	-1.44%	-0.01%	0.02%	5.47	21.65%	46.79%	45.74%	50.13%	43.46%
11-May	11	45.63%	47.21%	-1.58%	0.02%	0.02%	5.97	22.62%	46.86%	45.63%	50.19%	43.52%
12-May	12	46.67%	47.24%	-0.57%	0.01%	0.00%	6.46	23.71%	46.92%	46.67%	50.26%	43.58%
13-May	13	46.33%	47.28%	-0.94%	0.01%	0.01%	6.96	22.85%	46.98%	46.33%	50.32%	43.65%
14-May	14	46.96%	47.31%	-0.35%	0.00%	0.00%	7.46	23.65%	47.05%	46.96%	50.38%	43.71%
15-May	15	45.49%	47.35%	-1.85%	0.01%	0.03%	7.96	21.86%	47.11%	45.49%	50.45%	43.77%
16-May	16	47.19%	47.38%	-0.19%	0.00%	0.00%	8.45	24.30%	47.17%	47.19%	50.51%	43.84%
17-May	17	49.18%	47.41%	1.77%	0.00%	0.03%	8.95	25.44%	47.24%	49.18%	50.57%	43.90%
18-May	18	47.59%	47.45%	0.15%	0.00%	0.00%	9.45	22.85%	47.30%	47.59%	50.64%	43.97%
19-May	19	46.31%	47.48%	-1.18%	0.00%	0.01%	9.94	22.36%	47.37%	46.31%	50.70%	44.03%
20-May	20	48.34%	47.52%	0.82%	-0.01%	0.01%	10.44	25.04%	47.43%	48.34%	50.77%	44.09%
21-May	21	52.01%	47.55%	4.46%	0.04%	0.20%	10.94	27.69%	47.49%	52.01%	50.83%	44.16%
22-May	22	51.78%	47.58%	4.20%	0.19%	0.18%	11.43	25.61%	47.56%	51.78%	50.89%	44.22%
23-May	23	49.30%	47.62%	1.68%	0.07%	0.03%	11.93	23.24%	47.62%	49.30%	50.96%	44.29%
24-May	24	46.72%	47.65%	-0.93%	-0.02%	0.01%	12.43	21.91%	47.69%	46.72%	51.02%	44.35%
25-May	25	46.89%	47.69%	-0.80%	0.01%	0.01%	12.92	23.38%	47.75%	46.89%	51.08%	44.41%
26-May	26	47.63%	47.72%	-0.09%	0.00%	0.00%	13.42	24.04%	47.81%	47.63%	51.15%	44.48%
27-May	27	48.27%	47.75%	0.52%	0.00%	0.00%	13.92	24.31%	47.88%	48.27%	51.21%	44.54%
28-May	28	48.83%	47.79%	1.04%	0.01%	0.01%	14.41	24.54%	47.94%	48.83%	51.28%	44.60%
29-May	29	48.44%	47.82%	0.62%	0.01%	0.00%	14.91	23.87%	48.00%	48.44%	51.34%	44.67%
30-May	30	47.03%	47.86%	-0.82%	-0.01%	0.01%	15.41	22.66%	48.07%	47.03%	51.40%	44.73%
31-May	31	47.73%	47.89%	-0.16%	0.00%	0.00%	15.90	24.07%	48.13%	47.73%	51.47%	44.80%
1-Jun	32	48.92%	47.92%	1.00%	0.00%	0.01%	16.40	24.91%	48.20%	48.92%	51.53%	44.86%
2-Jun	33	45.89%	47.96%	-2.07%	-0.02%	0.04%	16.90	21.27%	48.26%	45.89%	51.60%	44.92%
3-Jun	34	43.92%	47.99%	-4.07%	0.08%	0.17%	17.39	20.83%	48.32%	43.92%	51.66%	44.99%
4-Jun	35	47.91%	48.02%	-0.12%	0.00%	0.00%	17.89	25.81%	48.39%	47.91%	51.72%	45.05%
5-Jun	36	49.63%	48.06%	1.58%	0.00%	0.02%	18.39	25.53%	48.45%	49.63%	51.79%	45.11%

Tabla 2

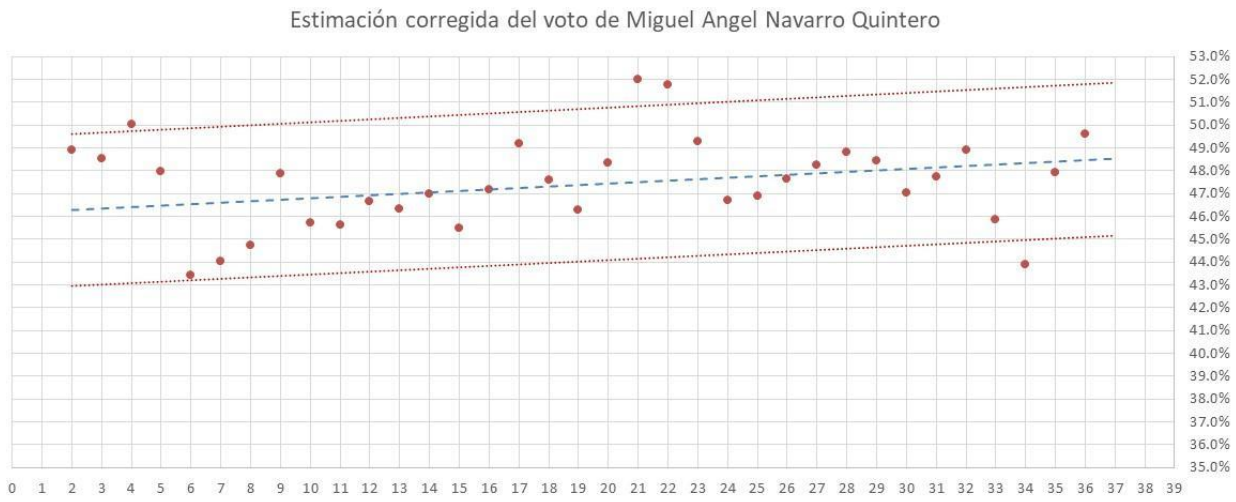


Gráfico 10

Modelo polinómico de orden 4: A efectos de interpretar la dinámica electoral se planteó un modelo polinómico de orden 4 aplicando las mismas correcciones enunciadas debido a los efectos de autocorrelación. Los resultados de observan en el gráfico 11 que ratifica la tendencia al valor 49%, sino que refleja el comportamiento ondulante del electorado y que refleja la opinión dubitativa de indecisos, una parte del electorado y sobre todo los indecisos fueron tomando diferentes posturas durante los últimos 37 días de campaña, lo cual refleja el impacto de la estrategia y de las modificaciones realizadas al mensaje de acuerdo con el seguimiento del comportamiento de los atributos motivadores del voto.

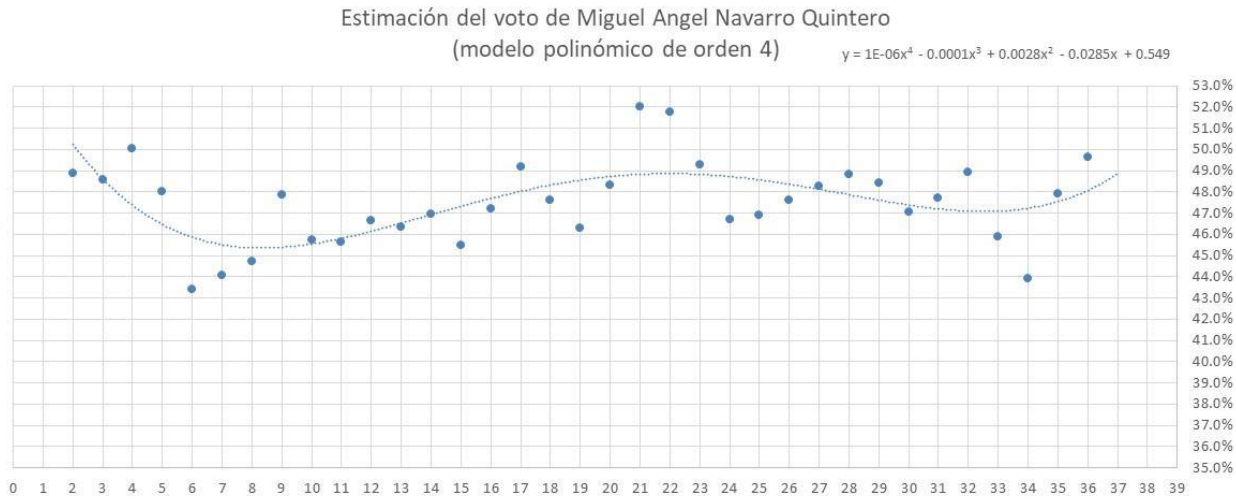


Gráfico 11

De igual manera se presentan los resultados del Tracking corregido de los dos candidatos que le siguen en orden de intención de voto, gráficos 12 y 13 obtenidos de las tablas 3 y 4 respectivamente.

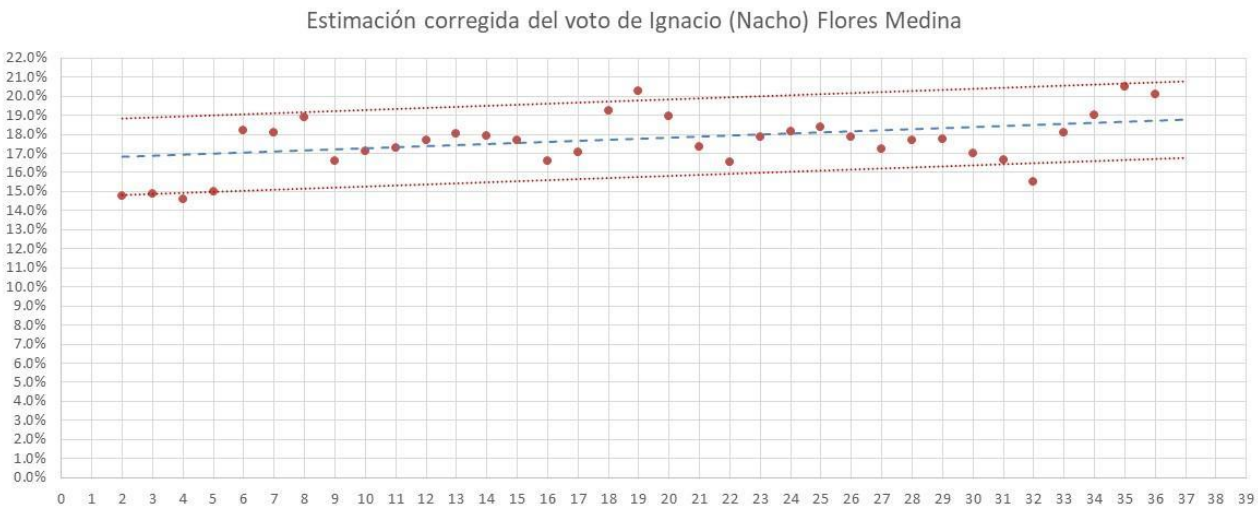


Gráfico 12

Ignacio (Nacho) Flores Medina												
Fecha	Xt	Yt	Estimación Lineal del voto	ut	ut . ut-1	ut2	X*	Yt*	Estimación corregida del voto	Voto observado	LS	LI
2-May	2	14.77%	16.43%	-1.66%					16.82%	14.77%	18.84%	14.80%
3-May	3	14.90%	16.49%	-1.59%	0.03%	0.03%	1.79	5.94%	16.88%	14.90%	18.90%	14.85%
4-May	4	14.59%	16.56%	-1.97%	0.03%	0.04%	2.18	5.55%	16.93%	14.59%	18.96%	14.91%
5-May	5	15.00%	16.62%	-1.63%	0.03%	0.03%	2.57	6.14%	16.99%	15.00%	19.01%	14.96%
6-May	6	18.18%	16.69%	1.49%	-0.02%	0.02%	2.97	9.08%	17.04%	18.18%	19.07%	15.02%
7-May	7	18.07%	16.76%	1.32%	0.02%	0.02%	3.36	7.04%	17.10%	18.07%	19.12%	15.08%
8-May	8	18.88%	16.82%	2.06%	0.03%	0.04%	3.75	7.91%	17.15%	18.88%	19.18%	15.13%
9-May	9	16.60%	16.89%	-0.29%	-0.01%	0.00%	4.14	5.14%	17.21%	16.60%	19.23%	15.19%
10-May	10	17.14%	16.96%	0.18%	0.00%	0.00%	4.54	7.06%	17.27%	17.14%	19.29%	15.24%
11-May	11	17.28%	17.02%	0.26%	0.00%	0.00%	4.93	6.88%	17.32%	17.28%	19.35%	15.30%
12-May	12	17.67%	17.09%	0.58%	0.00%	0.00%	5.32	7.18%	17.38%	17.67%	19.40%	15.35%
13-May	13	18.02%	17.16%	0.86%	0.00%	0.01%	5.72	7.29%	17.43%	18.02%	19.46%	15.41%
14-May	14	17.94%	17.22%	0.72%	0.01%	0.01%	6.11	7.01%	17.49%	17.94%	19.51%	15.47%
15-May	15	17.70%	17.29%	0.41%	0.00%	0.00%	6.50	6.81%	17.54%	17.70%	19.57%	15.52%
16-May	16	16.61%	17.36%	-0.75%	0.00%	0.01%	6.90	5.86%	17.60%	16.61%	19.62%	15.58%
17-May	17	17.04%	17.42%	-0.38%	0.00%	0.00%	7.29	6.96%	17.66%	17.04%	19.68%	15.63%
18-May	18	19.26%	17.49%	1.76%	-0.01%	0.03%	7.68	8.91%	17.71%	19.26%	19.73%	15.69%
19-May	19	20.29%	17.56%	2.73%	0.05%	0.07%	8.07	8.60%	17.77%	20.29%	19.79%	15.74%
20-May	20	18.93%	17.62%	1.31%	0.04%	0.02%	8.47	6.62%	17.82%	18.93%	19.85%	15.80%
21-May	21	17.32%	17.69%	-0.37%	0.00%	0.00%	8.86	5.83%	17.88%	17.32%	19.90%	15.86%
22-May	22	16.56%	17.76%	-1.19%	0.00%	0.01%	9.25	6.05%	17.93%	16.56%	19.96%	15.91%
23-May	23	17.88%	17.82%	0.06%	0.00%	0.00%	9.65	7.83%	17.99%	17.88%	20.01%	15.97%
24-May	24	18.14%	17.89%	0.25%	0.00%	0.00%	10.04	7.28%	18.05%	18.14%	20.07%	16.02%
25-May	25	18.38%	17.96%	0.42%	0.00%	0.00%	10.43	7.37%	18.10%	18.38%	20.12%	16.08%
26-May	26	17.86%	18.02%	-0.17%	0.00%	0.00%	10.83	6.70%	18.16%	17.86%	20.18%	16.13%
27-May	27	17.22%	18.09%	-0.87%	0.00%	0.01%	11.22	6.38%	18.21%	17.22%	20.24%	16.19%
28-May	28	17.70%	18.16%	-0.46%	0.00%	0.00%	11.61	7.25%	18.27%	17.70%	20.29%	16.24%
29-May	29	17.74%	18.22%	-0.48%	0.00%	0.00%	12.01	7.00%	18.32%	17.74%	20.35%	16.30%
30-May	30	17.01%	18.29%	-1.28%	0.01%	0.02%	12.40	6.24%	18.38%	17.01%	20.40%	16.36%
31-May	31	16.65%	18.36%	-1.70%	0.02%	0.03%	12.79	6.33%	18.43%	16.65%	20.46%	16.41%
1-Jun	32	15.52%	18.42%	-2.90%	0.05%	0.08%	13.18	5.41%	18.49%	15.52%	20.51%	16.47%
2-Jun	33	18.07%	18.49%	-0.42%	0.01%	0.00%	13.58	8.65%	18.55%	18.07%	20.57%	16.52%
3-Jun	34	18.99%	18.56%	0.44%	0.00%	0.00%	13.97	8.02%	18.60%	18.99%	20.63%	16.58%
4-Jun	35	20.47%	18.62%	1.85%	0.01%	0.03%	14.36	8.94%	18.66%	20.47%	20.68%	16.63%
5-Jun	36	20.09%	18.69%	1.40%	0.03%	0.02%	14.76	7.67%	18.71%	20.09%	20.74%	16.69%

Tabla 3

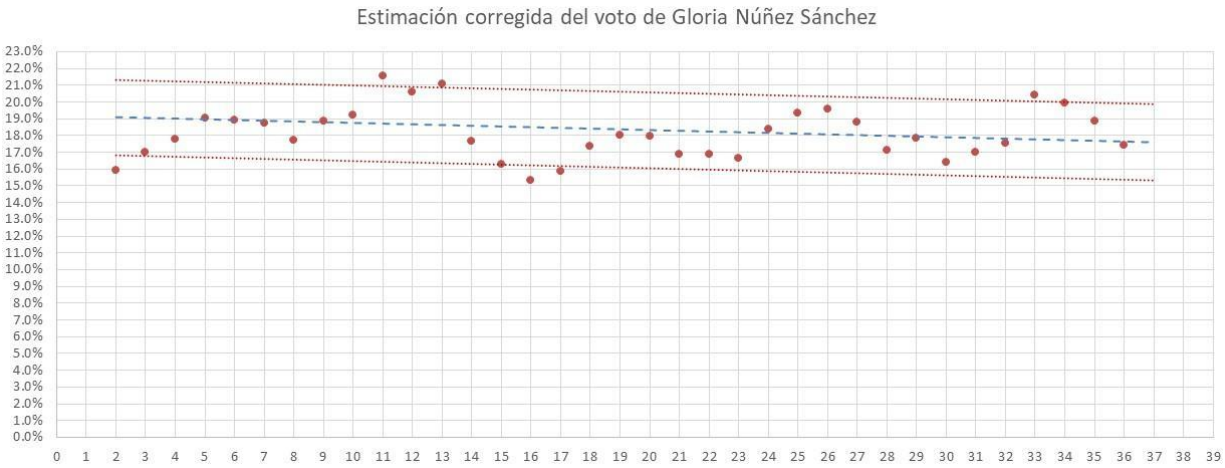


Gráfico 13

Gloria Núñez Sánchez												
Fecha	Xt	Yt	Estimación Lineal del voto	ut	ut . ut-1	ut2	X*	Yt*	Estimación corregida del voto	Voto observado	LS	LI
2-May	2	15.94%	18.13%	-2.18%					19.09%	15.94%	21.34%	16.84%
3-May	3	17.02%	18.12%	-1.10%	0.02%	0.01%	1.67	6.43%	19.05%	17.02%	21.30%	16.80%
4-May	4	17.78%	18.12%	-0.34%	0.00%	0.00%	2.01	6.47%	19.01%	17.78%	21.25%	16.76%
5-May	5	19.05%	18.12%	0.92%	0.00%	0.01%	2.34	7.24%	18.96%	19.05%	21.21%	16.72%
6-May	6	18.93%	18.12%	0.80%	0.01%	0.01%	2.68	6.27%	18.92%	18.93%	21.17%	16.67%
7-May	7	18.72%	18.12%	0.60%	0.00%	0.00%	3.01	6.14%	18.88%	18.72%	21.13%	16.63%
8-May	8	17.71%	18.12%	-0.42%	0.00%	0.00%	3.35	5.27%	18.84%	17.71%	21.08%	16.59%
9-May	9	18.83%	18.12%	0.71%	0.00%	0.01%	3.69	7.07%	18.79%	18.83%	21.04%	16.55%
10-May	10	19.19%	18.12%	1.08%	0.01%	0.01%	4.02	6.69%	18.75%	19.19%	21.00%	16.50%
11-May	11	21.55%	18.12%	3.43%	0.04%	0.12%	4.36	8.80%	18.71%	21.55%	20.96%	16.46%
12-May	12	20.60%	18.12%	2.48%	0.09%	0.06%	4.69	6.28%	18.67%	20.60%	20.91%	16.42%
13-May	13	21.09%	18.12%	2.98%	0.07%	0.09%	5.03	7.41%	18.62%	21.09%	20.87%	16.37%
14-May	14	17.64%	18.12%	-0.47%	-0.01%	0.00%	5.36	3.63%	18.58%	17.64%	20.83%	16.33%
15-May	15	16.26%	18.11%	-1.85%	0.01%	0.03%	5.70	4.54%	18.54%	16.26%	20.79%	16.29%
16-May	16	15.31%	18.11%	-2.81%	0.05%	0.08%	6.04	4.50%	18.50%	15.31%	20.74%	16.25%
17-May	17	15.88%	18.11%	-2.23%	0.06%	0.05%	6.37	5.72%	18.45%	15.88%	20.70%	16.20%
18-May	18	17.38%	18.11%	-0.74%	0.02%	0.01%	6.71	6.82%	18.41%	17.38%	20.66%	16.16%
19-May	19	18.03%	18.11%	-0.08%	0.00%	0.00%	7.04	6.48%	18.37%	18.03%	20.62%	16.12%
20-May	20	17.95%	18.11%	-0.16%	0.00%	0.00%	7.38	5.98%	18.33%	17.95%	20.57%	16.08%
21-May	21	16.91%	18.11%	-1.20%	0.00%	0.01%	7.71	4.98%	18.28%	16.91%	20.53%	16.03%
22-May	22	16.90%	18.11%	-1.21%	0.01%	0.01%	8.05	5.66%	18.24%	16.90%	20.49%	15.99%
23-May	23	16.64%	18.11%	-1.47%	0.02%	0.02%	8.39	5.42%	18.20%	16.64%	20.45%	15.95%
24-May	24	18.37%	18.11%	0.26%	0.00%	0.00%	8.72	7.32%	18.15%	18.37%	20.40%	15.91%
25-May	25	19.35%	18.11%	1.25%	0.00%	0.02%	9.06	7.15%	18.11%	19.35%	20.36%	15.86%
26-May	26	19.56%	18.10%	1.46%	0.02%	0.02%	9.39	6.71%	18.07%	19.56%	20.32%	15.82%
27-May	27	18.77%	18.10%	0.67%	0.01%	0.00%	9.73	5.78%	18.03%	18.77%	20.28%	15.78%
28-May	28	17.11%	18.10%	-0.99%	-0.01%	0.01%	10.06	4.64%	17.98%	17.11%	20.23%	15.74%
29-May	29	17.85%	18.10%	-0.25%	0.00%	0.00%	10.40	6.48%	17.94%	17.85%	20.19%	15.69%
30-May	30	16.41%	18.10%	-1.69%	0.00%	0.03%	10.74	4.55%	17.90%	16.41%	20.15%	15.65%
31-May	31	16.98%	18.10%	-1.12%	0.02%	0.01%	11.07	6.08%	17.86%	16.98%	20.11%	15.61%
1-Jun	32	17.54%	18.10%	-0.56%	0.01%	0.00%	11.41	6.26%	17.81%	17.54%	20.06%	15.57%
2-Jun	33	20.39%	18.10%	2.29%	-0.01%	0.05%	11.74	8.74%	17.77%	20.39%	20.02%	15.52%
3-Jun	34	19.93%	18.10%	1.84%	0.04%	0.03%	12.08	6.39%	17.73%	19.93%	19.98%	15.48%
4-Jun	35	18.85%	18.10%	0.76%	0.01%	0.01%	12.41	5.61%	17.69%	18.85%	19.93%	15.44%
5-Jun	36	17.44%	18.10%	-0.65%	0.00%	0.00%	12.75	4.92%	17.64%	17.44%	19.89%	15.40%

Tabla 4

APLICACIÓN PARA EL CONTROL DE CAMPAÑA

Se muestra a continuación un resumen de cómo se aplicó el tracking para el control de atributos formadores del voto para ajustes del mensaje y estrategia.

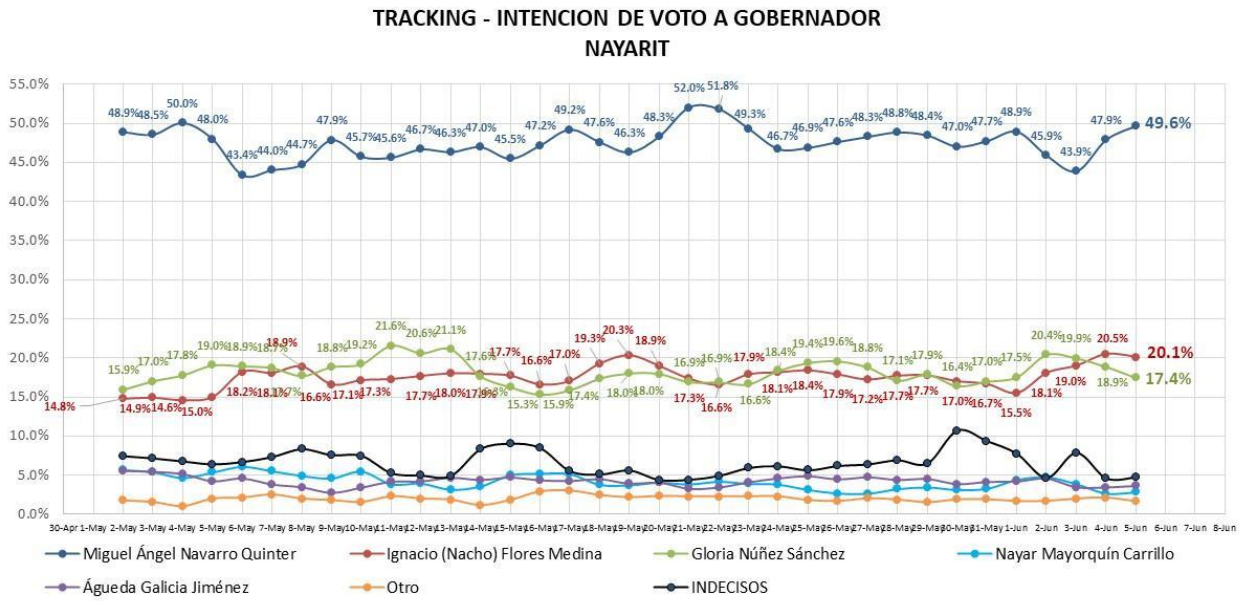


Gráfico 14

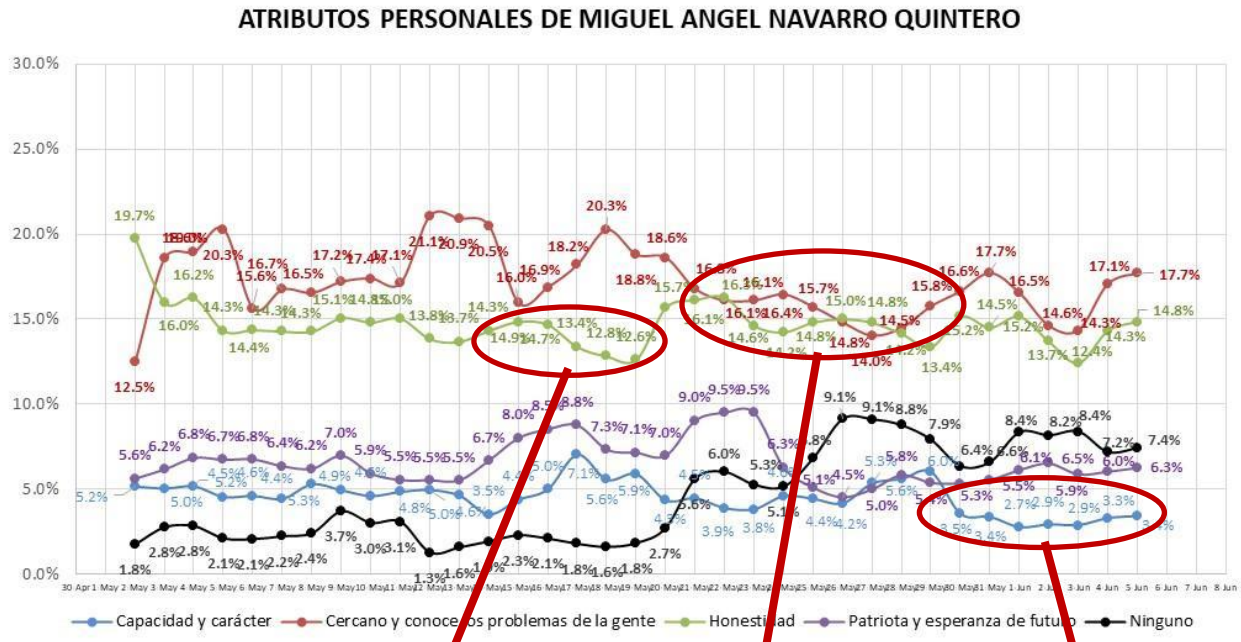


Gráfico 15

Mejorar Honestidad

Mejorar percepción de cercano y conoce los problemas de la gente

Evitar decrezca la percepción de capacidad y carácter

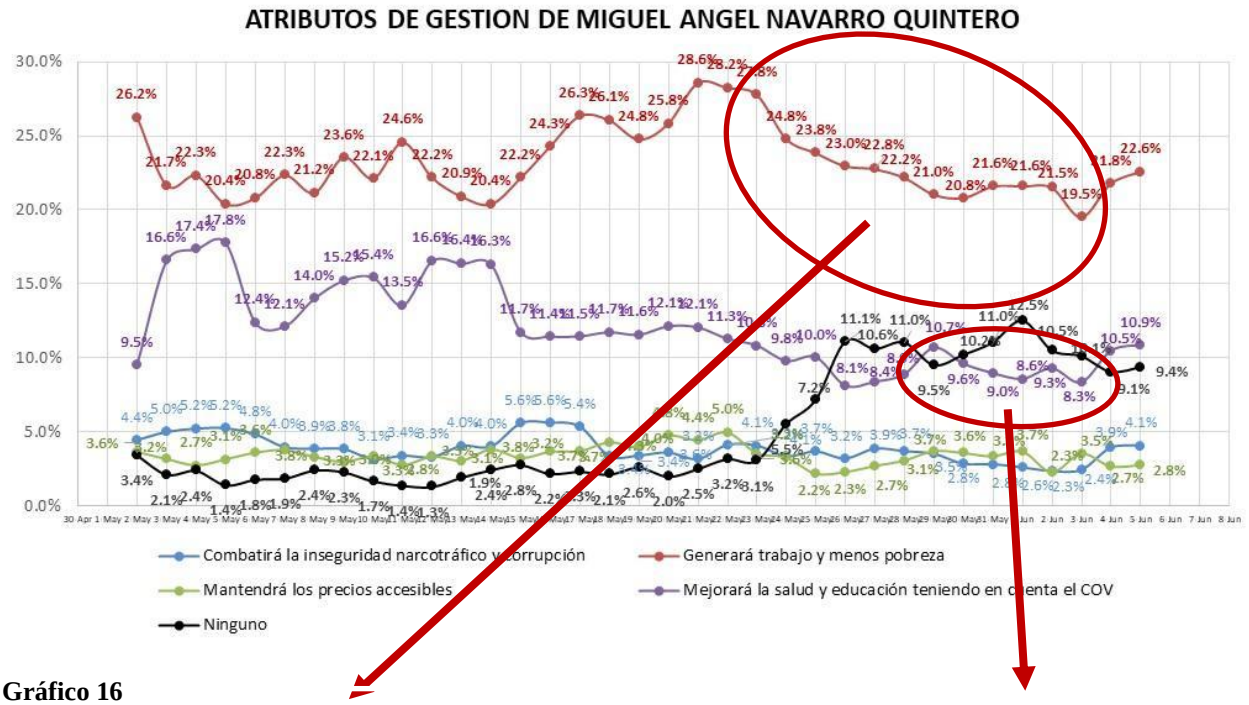


Gráfico 16

Recuperar la percepción de generación de trabajo y menos pobreza

Reforzar salud y educación en contexto de pandemia

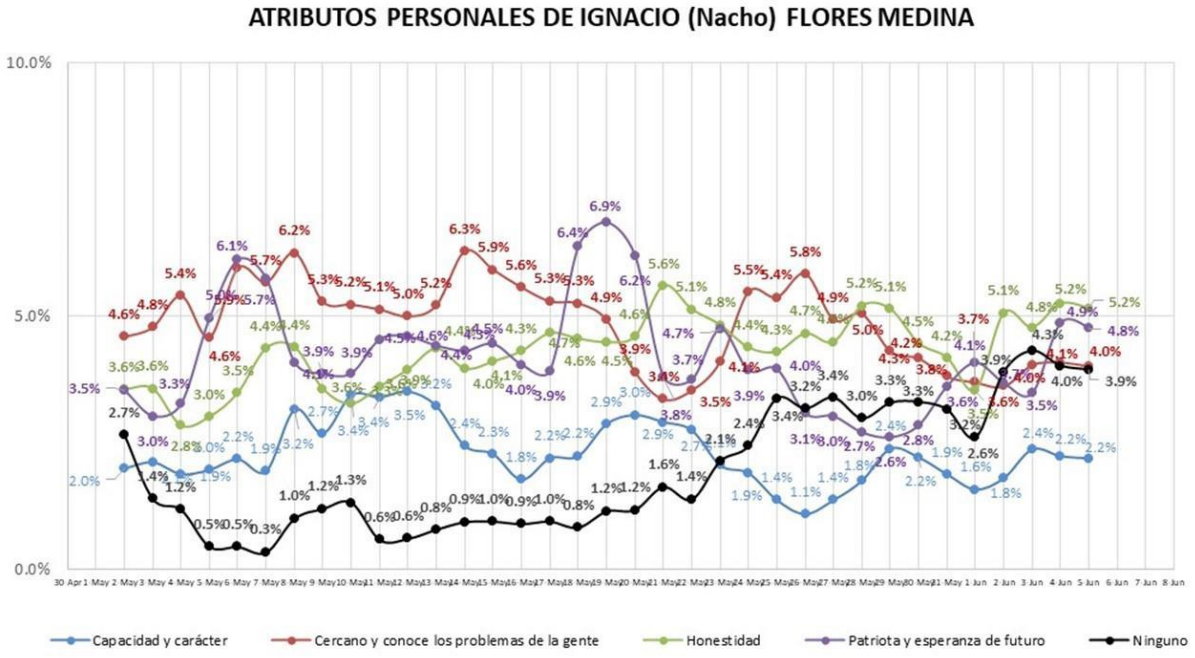


Gráfico 17

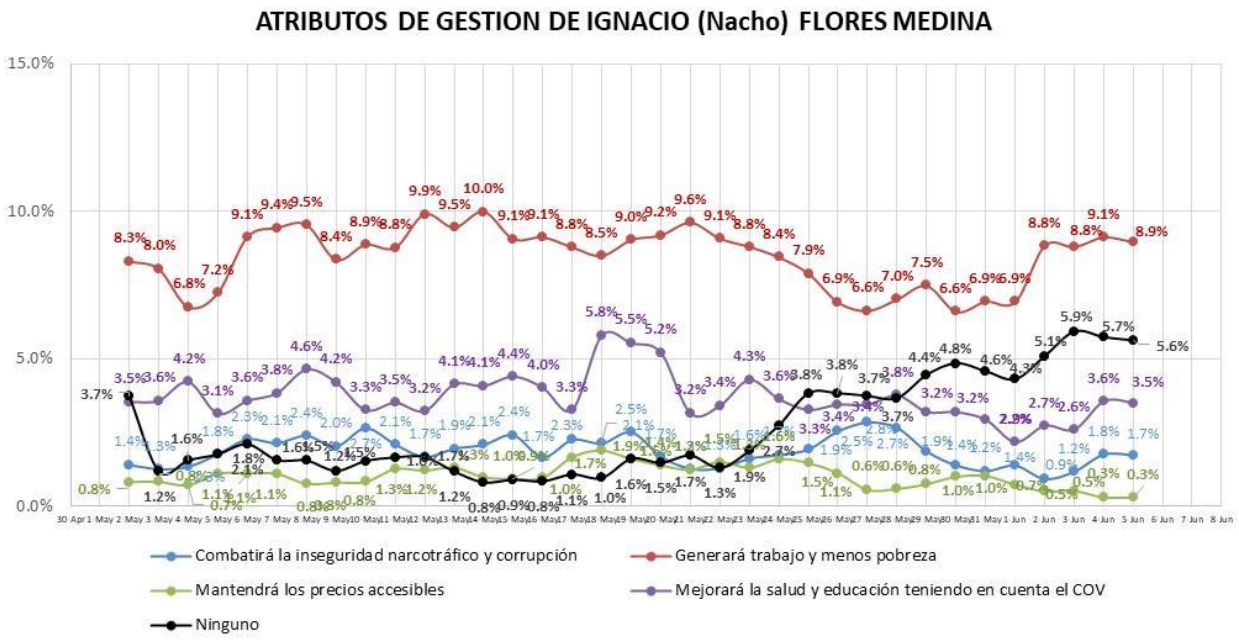

ATRIBUTOS DE GESTION DE IGNACIO (Nacho) FLORES MEDINA
15.0%
10.0%
5.0%
0.0%
Combatirá la inseguridad narcotráfico y corrupción
Generará trabajo y menos pobreza
Mantendrá los precios accesibles
Mejorará la salud y educación teniendo en cuenta el COV
Ninguno

Gráfico 18

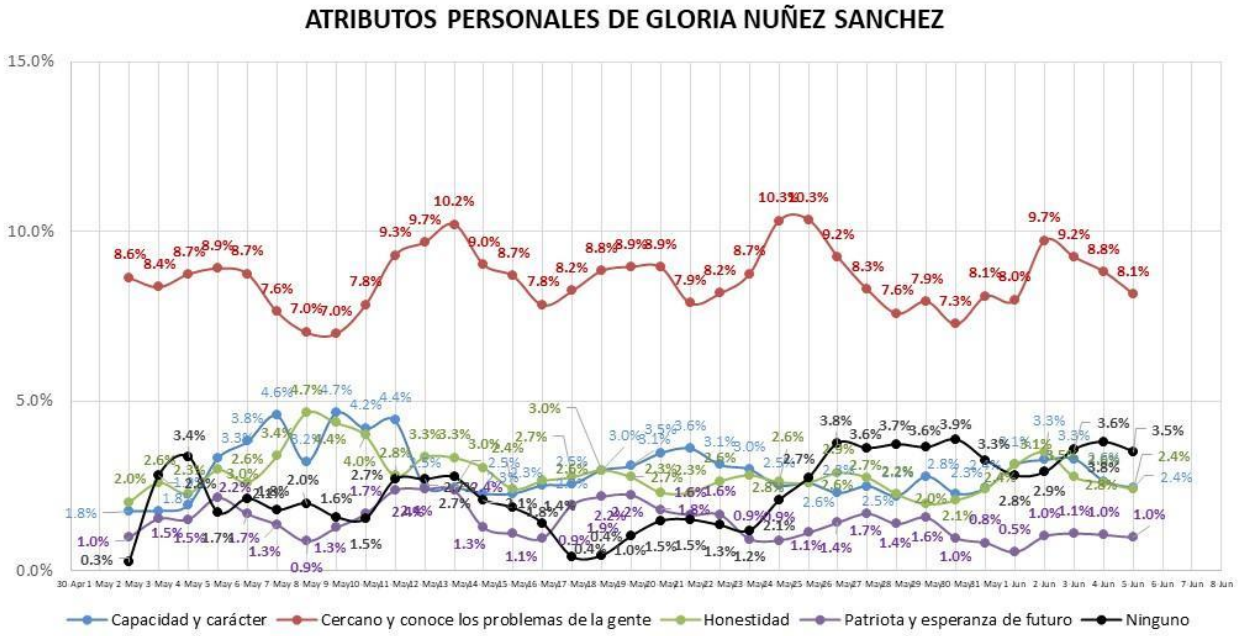


Gráfico 19

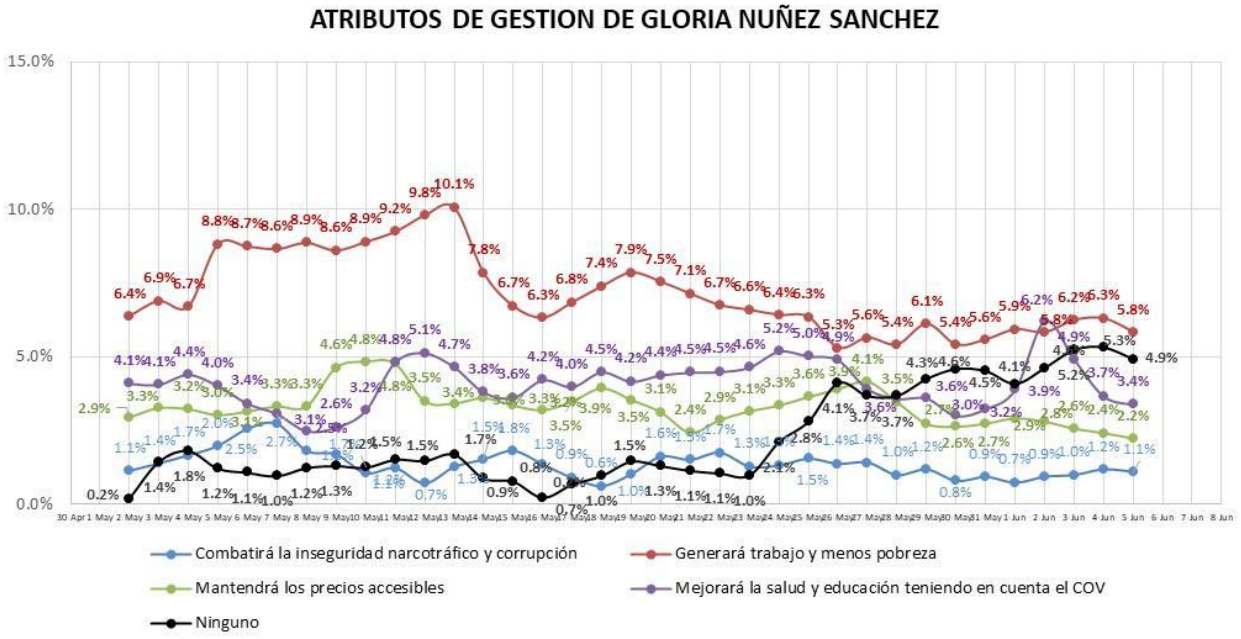


Gráfico 20

TRACKING POR ZONAS

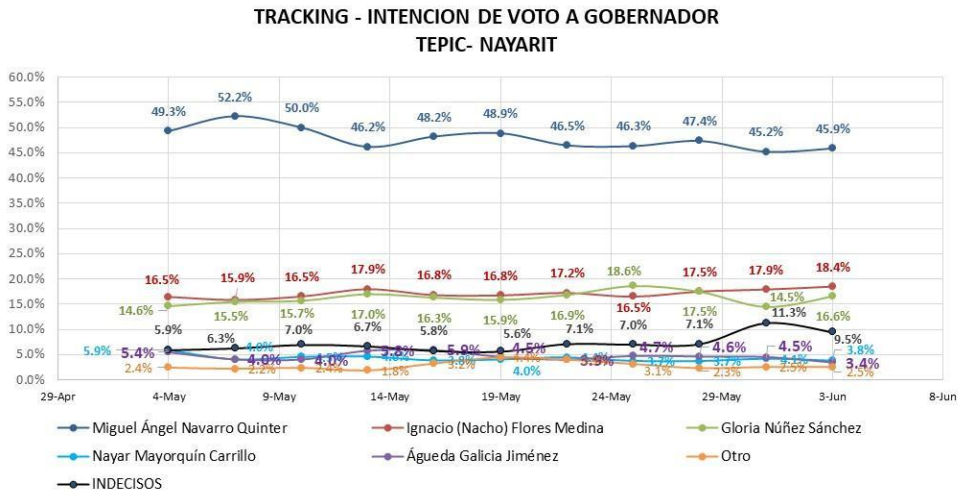


Gráfico 21

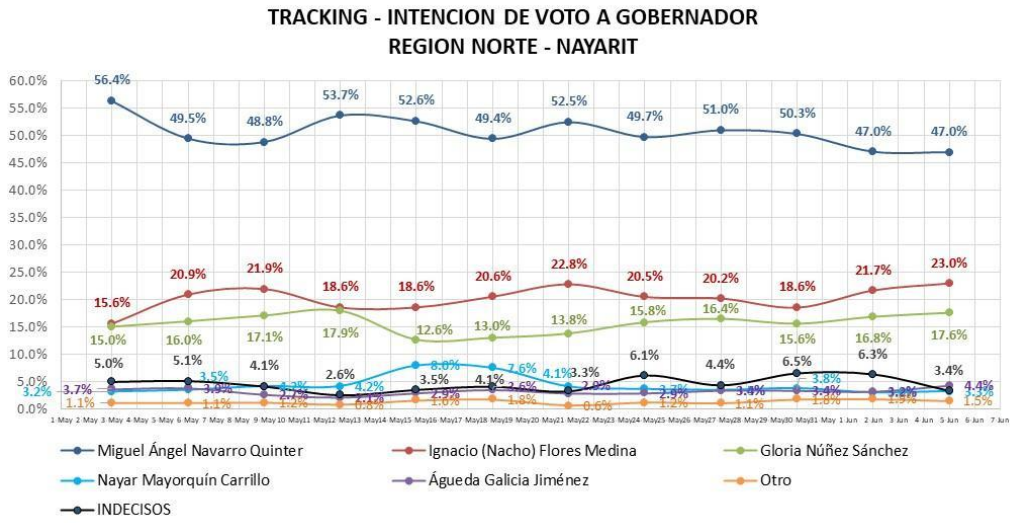


Gráfico 22

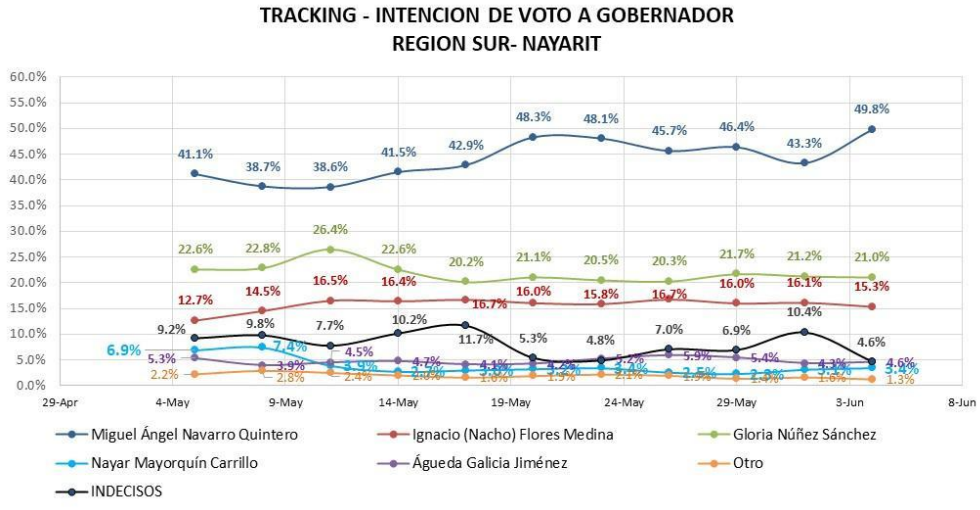


Gráfico 23

RESULTADOS EL 6 DE JUNIO

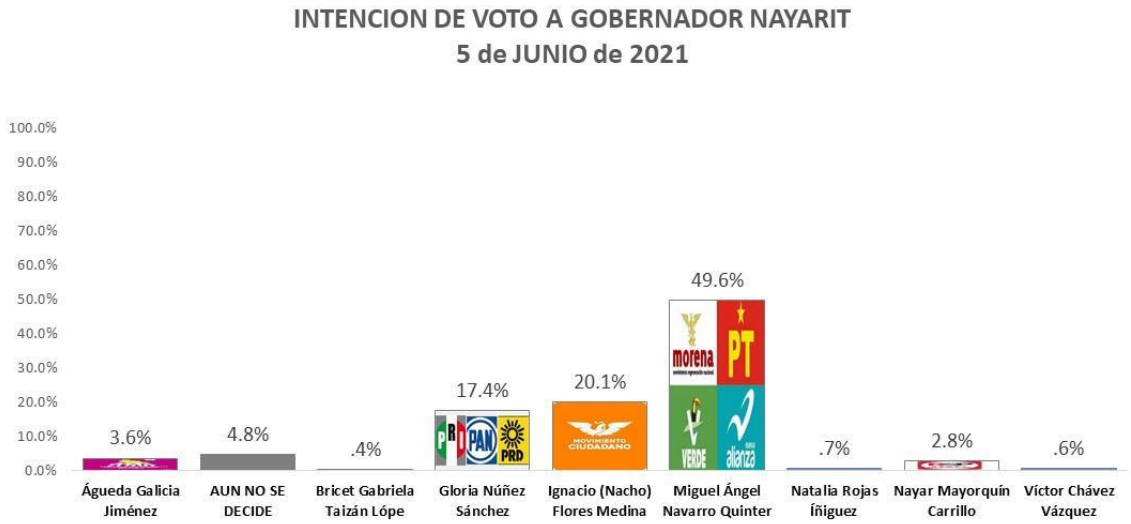


Gráfico 24

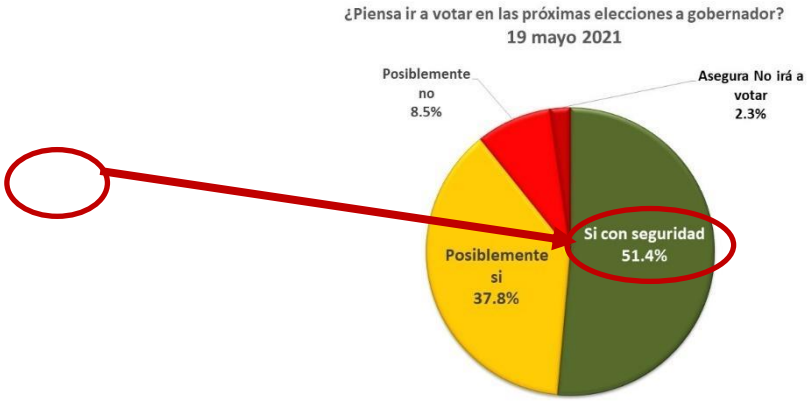


Gráfico 25

	Int Voto	Resultado	Error
Miguel Ángel Navarro Quintero	49.6%	49.3%	0.3%
Ignacio (Nacho) Flores Medina	20.1%	20.5%	0.4%
Gloria Núñez Sánchez	17.4%	17.7%	0.3%

CONCLUSIONES

Respecto a la última medición el error promedio observado es de 0.33 puntos, reflejando alta precisión del método usado. Las claves fueron:

- ✓ Filtrar abstencionistas
- ✓ Aplicar factores de ponderación basado en el padrón electoral
- ✓ Aplicar muestreo de control territorial para ajuste del muestreo robótico.
- ✓ División del estado en 3 macrozonas para control local.
- ✓ Aumento de la precisión por eliminación del efecto de autocorrelación.
- ✓ Aceptar y conocer que la opinión pública es ondulante y afecta la intención de voto, medida a través de un modelo polinómico de regresión.
- ✓ Analizar el comportamiento de indecisos mediante un análisis factorial en la etapa inicial de la campaña.
- ✓ Conocer el Humor social para establecer el estilo de campaña no agresivo y basado en conceptos de esperanza.
- ✓ Enfocar la alianza en el partido Morena e integrar a AMLO.
- ✓ Determinar los atributos formadores del voto clasificados en personales y de gestión.
- ✓ Controlar en el tracking los siguientes atributos:
    - Personales
        - Mejorar la percepción de cercano y conocer los problemas de la gente.
        - Honestidad.
        - Evitar que disminuya la percepción de capacidad y carácter del candidato.
    - Gestión
        - Recuperar la percepción de la capacidad para generar de trabajo y menos pobreza.
        - Reforzar la salud y educación en contexto de pandemia.

BIBLIOGRAFÍA

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). Estadística para administración y Economía. *10a.* Col. Cruz Manca, Santa Fe, D.F., México: Cencage Learning.

Hines, W. W., & Montgomery, D. C. (1996). Probabilidad y Estadística para Ingeniería y Administración. *3era.* Colonia San Juan Tlihuaca, D.F., México: Compañía Editorial Continental, S.A. de C.V.

Mendenhall, W. B. (2010). *Introducción a la probabilidad y estadística.* Santa Fe, D.F., México: Cencage Learning.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2009). Estadística matemática con aplicaciones. *7a.* Col. Cruz Manca, Santa Fe, México, D.F.: Cengage Learning Editores S.A. de C.V.

Walpole, R. E. (2012). *Probabilidad y estadística para ingeniería y ciencias.* México: Pearson Education.

IV Jornadas Internacionales de Estadística Aplicada

Estudio de las condiciones socioambientales y su vinculación con el delito en Jujuy.

Mariela Rodríguez, Nazarena Laureano, Gerardo Vargas, Norma Castro, Karen Navarro, Micaela Soria, Fabian López y Jesús Monne Escalante

Institución: Facultad de Ingeniería, Facultad de Humanidades y Ciencias Sociales, Universidad Nacional de Jujuy, Dirección General de Observatorio de Seguridad. San Salvador de Jujuy.

Contacto:
mariela.rodriguez@fi.unju.edu.ar

RESUMEN

El análisis espacial permite presentar las particularidades de los datos de un territorio, así como la manera de tenerlas en cuenta en el momento del análisis, estimula un correcto razonamiento espacial y un entendimiento adecuado tanto de las limitaciones como de la potencialidad de los datos espaciales como fuente de análisis geográfico. Con este estudio se pretende describir la relación entre las condiciones socioambientales con los hechos delictivos a partir de su ubicación en el espacio.

El trabajo de campo se realizó en barrios del departamento Dr. Manuel Belgrano de la Provincia de Jujuy, se aplicaron las técnicas de observación directa, no participativa y entrevistas semiestructuradas a referentes barriales, permitiendo conocer las siguientes variables: datos generales del barrio, infraestructura barrial, percepción sobre el escenario barrial, percepción de sobre la seguridad y participación ciudadana. Estas variables se interpolan con los hechos delictivos contra la propiedad registrados en el año 2021 brindados por el CIAC (Centro de Información y Análisis Criminal) del Ministerio de Seguridad de la Provincia de Jujuy.

En cuanto a los resultados del trabajo, se logró un análisis exploratorio de datos espaciales por barrio, obtenido a partir de la aplicación de diversas técnicas de visualización espacial tales como: Google Maps, cuya herramienta permitió ubicar puntos concretos sobre las variables estudiadas logrando las primeras aproximaciones al estudio y, QGIS desde el cual se generaron mapas digitales georreferenciados.

Esto posibilitó la construcción de diagnósticos preliminares sobre la situación de los diferentes barrios, aportando a simple vista, una densidad importante de datos que confluyen y refieren a un determinado sector de la realidad que permiten analizar la seguridad en cada uno de ellos.

Palabras Clave: estadística espacial, condiciones socio ambientales, delitos.

INTRODUCCIÓN

De acuerdo a diversas teorías criminológicas (Anitua, 2010) los delitos son multicausales, ya que involucran diferentes factores sociales, ambientales, económicos, culturales. Por lo cual, indagar sobre las condiciones socioambientales territoriales y su vinculación con el delito resulta relevante ya que permitirá, por un lado, describir y comprender esta correlación de manera científica y por el otro lado, fundamentar y proponer Políticas Públicas en materia de seguridad.

Entendemos las Políticas de Seguridad como un conjunto de intervenciones de carácter público que tienen como finalidad controlar los riesgos, conflictos violentos y delictivos que lesionan los derechos y las libertades de las personas, mediante la prevención, conjuración y/o represión de los mismos. Cabe mencionar, además, que desde el paradigma de la Seguridad Ciudadana se hace énfasis en la participación de diferentes actores sociales y estatales para la elaboración de diagnósticos participativos y en acciones para la prevención del delito y la violencia, que en ocasiones exceden la labor de las fuerzas de seguridad.

Siguiendo este enfoque diversos autores (Sozzo, 2008) distinguen tres tácticas alternativas de prevención del delito, tal como fueron desarrollándose en los distintos países. Vale referirse a la prevención *situacional-ambiental, la prevención social y la prevención comunitaria.* Una vez más, la distinción es analítica, pues estas estrategias de intervención se mezclan en el terreno y se combinan de diversas maneras. (Varela, 2010) Conforme a este enfoque, es que se considera que las condiciones ambientales, sociales y comunitarias favorecerían el acontecimiento de hechos delictivos en los distintos barrios de la ciudad de San Salvador de Jujuy.

Desde esta estrategia se pretende construir diagnósticos de los distintos escenarios barriales caracterizados por la vulnerabilidad y conflictividad social, que contribuyan al diseño e implementación de políticas públicas en materia de seguridad para el territorio provincial, obtener información confiable y compleja cuali-cuantitativa, respecto a la situación social y delictual (estadística oficial policial y cifras negras) de cada espacio poblacional seleccionado, generar mapas situacionales que permitan caracterizar cada escenario barrial y su vinculación con los niveles de conflictividad social y los índices de delitos.

El proyecto se ejecutó a partir del convenio de colaboración entre el Ministerio de Seguridad y el Instituto de Educación Superior N° 7 “Populorum Progressio In.Te.La.”, específicamente para la instancia de trabajo de campo, en el cual intervinieron estudiantes de la carrera Tecnicatura Superior en Trabajo Social.

Este trabajo investigativo, es innovador en el sentido que se logra como producto la construcción de diagnósticos preliminares sobre determinado territorio en materia de seguridad usando herramientas de análisis espacial que permiten obtener un panorama visual digitalizado sobre las condiciones socio-ambientales que caracterizan a cada escenario barrial estudiado del Departamento Dr. Manuel Belgrano en relación con el acontecimiento de hechos delictivos (denunciados y de acuerdo a la percepción de referentes locales).

A partir de la aplicación de las herramientas de análisis espacial, los resultados obtenidos contribuirán al diseño de Políticas Públicas en materia de seguridad pública (acciones públicas orientadas a producir niveles aceptables de convivencia) y seguridad ciudadana (seguridad orientada al ejercicio de derechos y obligaciones como ciudadanos), en tanto se apunta a que desde el Estado se generen acciones preventivas diferencias que contribuyan a disminuir/limitar la criminalidad.

METODOLOGÍA

Se trata de una investigación aplicada, descriptiva, sincrónica y empírica. Desde lo metodológico es cuanti-cualitativa, ya que se articularon datos estadísticos y cualitativos generados a partir de las siguientes técnicas: observación semiestructurada de los espacios poblacionales seleccionados, entrevistas (referentes comunitarios, miembros de organizaciones gubernamentales y de la sociedad civil, entre otros.), análisis de datos

estadísticos sobre los delitos más frecuentes y su relación con las condiciones socio-ambientales de los escenarios barriales (Vieytes, 2004).

Técnicas de investigación:

Los instrumentos seleccionados guardan correspondencia con los objetivos de la Investigación y permiten el análisis de distintas categorías que se consideran trascendentes al momento de realizar un estudio de esta magnitud a fin de describir las condiciones socioambientales concretas.

- Observación semi estructurada
- Entrevistas
- Registros cartográficos
- Técnicas de estadística espacial

Estadística Espacial

La estadística espacial es el conjunto de técnicas estadísticas que cuantifican aspectos relacionados con la estructura de las distribuciones espaciales. La característica distintiva del análisis estadístico de datos espaciales es que el patrón espacial de las localizaciones (objetos espaciales), la asociación espacial entre los valores observados en diferentes localizaciones (dependencia espacial) y la variación sistemática del fenómeno en las distintas localizaciones (heterogeneidad espacial) se convierte en el mayor foco de investigación. (Miranda Salas & Condal, 2003)

Esta rama de la ciencia permite realizar un análisis exploratorio de datos que es un instrumento indispensable al momento de realizar las primeras aproximaciones al estudio de la estructura de la información socioespacial en una determinada área de estudio (Buzai & Baxendale, 2012).

Para el desarrollo del trabajo se utilizará el software Qgis que es un sistema de información geográfica de código abierto. Admite diversos formatos de datos ráster y vectoriales, pudiendo añadir nuevos formatos usando la arquitectura de complementos. Proporciona una creciente gama de capacidades a través de sus funciones básicas y complementos. Puede visualizar, gestionar, editar y analizar datos, y diseñar mapas imprimibles.

Localización y selección de la muestra

El criterio para la selección de los espacios poblacionales responderá fundamentalmente a la tasa delictiva registrada en el CIAC (Centro de Información y Análisis Criminal) de la provincia de Jujuy. Se tomaron aquellos escenarios que presentan alto índice de delitos y de conflictividad social, ubicados sobre la capital de la provincia.

Como primer momento se iniciará el proceso de análisis sobre algunos barrios del Dpto. Dr. Manuel Belgrano: Mariano Moreno, Cerro Las Rosas, San Pedrito, Punta Diamante, El Chingo, Belgrano, Chijra y Alto Gorriti

Criterios Muestrales

Se trata de una muestra intencional, considerando que el proceso investigativo se desarrollará en escenarios sociales multi problemáticos conforme a los índices delictivos y al nivel de conflictividad social detectado en los mismos. Datos que surgen de entidades oficiales dependientes del Ministerio de Seguridad (CIAC).

La base muestral, como se advierte previamente, está constituida por los ciudadanos que están representados por referentes barriales con legitimidad social y por actores externos pero que forman parte activa de cada una de las comunidades, seleccionadas porque desarrollan

sus funciones en distintas organizaciones gubernamentales y no gubernamentales emplazadas en las jurisdicciones escogidas.

Variables y dimensiones del estudio

- **Datos generales del Barrio:** nombre del Barrio, Accesos, límites, características topográficas: naturales (zanjones, hondonadas, etc.) y artificiales (puentes, pasarelas, etc.)
- **Infraestructura barrial:** servicios públicos, características habitacionales, organizaciones públicas, organizaciones de la Sociedad Civil, actividades sociales, actividades comerciales, uso del espacio público y privado libre.
- **Percepción sobre el escenario barrial:** descripción del barrio, situaciones problemáticas que afectan al barrio, priorización, frecuencia, estado de higiene y cuidado del medio ambiente (percepción de la higiene barrial).
- **Percepción sobre la Seguridad:** Valoración sobre la seguridad en el barrio, señales de vandalismo, predictores de movimiento, rutas de escape, senderos de circulación, circuito de transporte, condiciones ambientales favorecedoras del delito, Iluminación, lugares propicios para esconderse, lugares considerados peligrosos, presencia Policial, medidas de Autoprotección Vecinal, propuestas para mejorar la seguridad.
- **Participación ciudadana:** disponibilidad de colaboración en las acciones del barrio, organizaciones preocupadas por lo que sucede en el barrio y actuación de los vecinos frente a situaciones de emergencia.

DESARROLLO

En este primer avance del proceso de análisis de datos, se describirán las características generales de los barrios estudiados, la infraestructura barrial en relación con la percepción de seguridad y su vinculación con el delito, específicamente en los barrios Mariano Moreno, Alto Gorriti y Chijra. Dicha selección se fundamenta en la densidad de los datos obtenidos a partir de la aplicación de los instrumentos de recolección de datos y de la técnica de análisis estadístico-espacial utilizado y la relación encontrada entre las variables en estudio.

Características generales

San Salvador de Jujuy constituye la región de mayor concentración de población (unos 300.000 habitantes). Hidrográficamente, está recorrida por dos ríos: el Grande y el Xibi Xibi o Chico. Ambos torrentes confluyen en la ciudad de San Salvador y continúan unidos, rumbo noreste hasta los ríos Lavayen – San Francisco, estos afluentes otorgan cierta particularidad de la edificación de los barrios que se ubican en la ciudad.

Por las particularidades de las zonas en las que se ubican cada uno de ellos, es interesante poder describirlos según sus características topográficas naturales y artificiales, que le otorgan características particulares.

Al noreste se encuentran los barrios Belgrano, El Chingo y Punta Diamante. Todos ellos se encuentran edificados en forma paralela al Río Grande y en un nivel de territorio bajo, respecto a la zona céntrica de la ciudad. Es por esta razón que se observa una diversidad de características topográficas que le otorgan una particularidad especial para comprender las distintas variables definidas para el estudio.

Entre las características topográficas artificiales, encontramos el Puente Gral. San Martín y el Puente Senador Pérez, a su vez, al margen de la Avda. Italia y Avda. Gdor. José María Fascio, 13 escaleras y 1 pasarela que permite el acceso peatonal, mientras que el acceso vehicular se realiza por la calle Lugones y Juanita Moro, para ingresar a los barrios Belgrano y El Chingo

y por calle Jade del Sur, para ingresar al barrio Punta Diamante. Entre las características topográficas naturales, los barrios de esta zona se distinguen por la proximidad al Río Grande, con la particularidad que en el barrio Punta Diamante, en el cual confluye con el Río Xibi Xibi.

De esta zona también se estudió el barrio Chijra, este barrio se caracteriza por estar edificado al margen de los cerros (lo cual implica la construcción de viviendas y calles en pendiente), limitar con el Río Grande y contar con la afluencia de un arroyo que atraviesa el barrio, en forma paralela a la calle Armanini.

Hacia el noroeste de la ciudad, se ubican los barrios Cerro Las Rosas y Mariano Moreno. Ambos tienen la particularidad de estar edificados en la zona alta de la ciudad. Entre las características topográficas artificiales, se observa la edificación de la Ruta Nacional N°9 que circula en forma paralela a ambos barrios, un puente y una pasarela que permiten la circulación de peatones entre ambas zonas y un canal fluvial que atraviesa el barrio Cerro Las Rosas. En cuanto a las características topográficas naturales, este último barrio, está edificado en proximidad al cerro y por una gran cantidad de terrenos baldíos (10 diez identificados en mapa). Los accesos al barrio se realizan en su mayoría, a través de calles que conectan con los barrios Cuyaya, Alto Gorriti, Bajo Gorriti, Luján y San Cayetano y la Ruta Nacional N°9 y por la Avda. Horacio Guzmán que permite la conexión con la zona céntrica de la ciudad.

Del centro de la ciudad, se estudió los barrios que se ubican próximos al centro de la ciudad: Alto Gorriti. En este caso, el barrio Alto Gorriti, tiene la particularidad de estar edificado en una zona elevada respecto al centro de la ciudad y se encuentra limitado por el barrio Mariano Moreno, compartiendo la Avda. Horacio Guzmán como uno de los principales accesos. En condición diferente se encuentra el barrio Luján, éste se presenta en una superficie plana y paralela respecto del centro de la ciudad y cuenta con diferentes avenidas que garantizan un acceso peatonal y vehicular al mismo: Avda. Gral. Savio, Avda. Pueyrredón, Avda. El Éxodo y Avda. Pte. Perón. Entre las características topográficas artificiales existentes en esta zona se encuentran: pasajes y escaleras, estas últimas presentes en el límite entre el barrio Alto Gorriti y Bajo Gorriti (más conocido como la zona Vieja Terminal de Ómnibus). En cuanto a las características topográficas naturales, la zona está caracterizada por zanjones y terrenos baldíos.

Análisis exploratorio de datos espaciales de las capas de Infraestructura barrial, percepción de seguridad y su vinculación con el delito

El análisis exploratorio de datos espaciales que se realiza a continuación cuenta con una descripción de las capas en estudio: Infraestructura barrial, percepción de seguridad y delitos contra la propiedad que permite descubrir los factores que son favorecedores del delito. Si bien el estudio se realizó en nueve barrios de la ciudad de San Salvador, para esta descripción se detallan tres de ellos como son San José de Chijra, Mariano Moreno y Alto Gorriti basado en el alto índice delictivo de delitos contra la propiedad que muestra la figura 1.

Figura 1: Mapa de calor de hechos delictuales de la Ciudad de San Salvador de Jujuy

Análisis descriptivo del Barrio San José de Chijra

Los principales accesos al barrio (figura 2), se realizan por el Pte. San Martín, Avda. Las Vicuñas, Avda. Mosconi y calle Armonía. Según datos de la observación, el acceso a los servicios públicos básicos se da sólo en algunos sectores. Actualmente se realizan obras de pavimentación en calle Las Corzuelas y existen zonas de asentamientos en sus márgenes.

La infraestructura social del barrio está constituida por la existencia de organizaciones públicas: Escuela, el Centro de Rehabilitación “Santa María”, Secretaría de Ciencia y Técnica de Jujuy, Centro de Desarrollo Infantil, Centro de Salud “La Viña”, Dirección de Educación Técnico Profesional, Dirección de Investigaciones, Centro de Participación Vecinal y Comisaria Seccional N° 3; y organizaciones de la sociedad civil: la Fundación “Conexos”, Templo de Testigos de Jehová, Iglesia San Bartolomé y el Centro Vecinal de “La Viña, ubicado en este barrio.

Se caracteriza por estar edificado al margen de las montañas (lo cual implica la construcción de viviendas y calles en pendiente), limita con el Río Grande y cuenta con la afluencia de un arroyo que atraviesa el barrio, en forma paralela a la calle Armanini. Se observa una zona comercial bien definida sobre las Avenidas Las Vicuñas y Las Corzuelas y de alta circulación por los ciudadanos, siendo esta condición favorecedora para la comisión de delitos contra la propiedad. Según datos del CIAC, estas arterias se configuran como lugares con alta densidad de delitos reportados.

En cuanto al consumo de estupefacientes se evidencia que se produce en zonas aledañas a las orillas del Río Grande y zonas que dan al margen de la montaña. Por otro lado, los ciudadanos marcan como lugares peligrosos y propicios para esconderse a aquellos que se encuentran cercanos al Río Grande y cercanos al Río Chijra. En forma paralela al Río Chijra se evidencia un asentamiento entre calle El Yapan y Los Quebrachales, lugar donde los referentes identifican el consumo de bebidas alcohólicas.

Es destacable nombrar también los hechos que suceden entre las calles el Amor y Armanani donde se puede visualizar otro asentamiento próximo al margen de la montaña paralela a la calle Armanani. Esta zona presenta un alto índice delictivo y coincide con un lugar calificado como peligroso e identificado como sector donde se consumen bebidas alcohólicas.

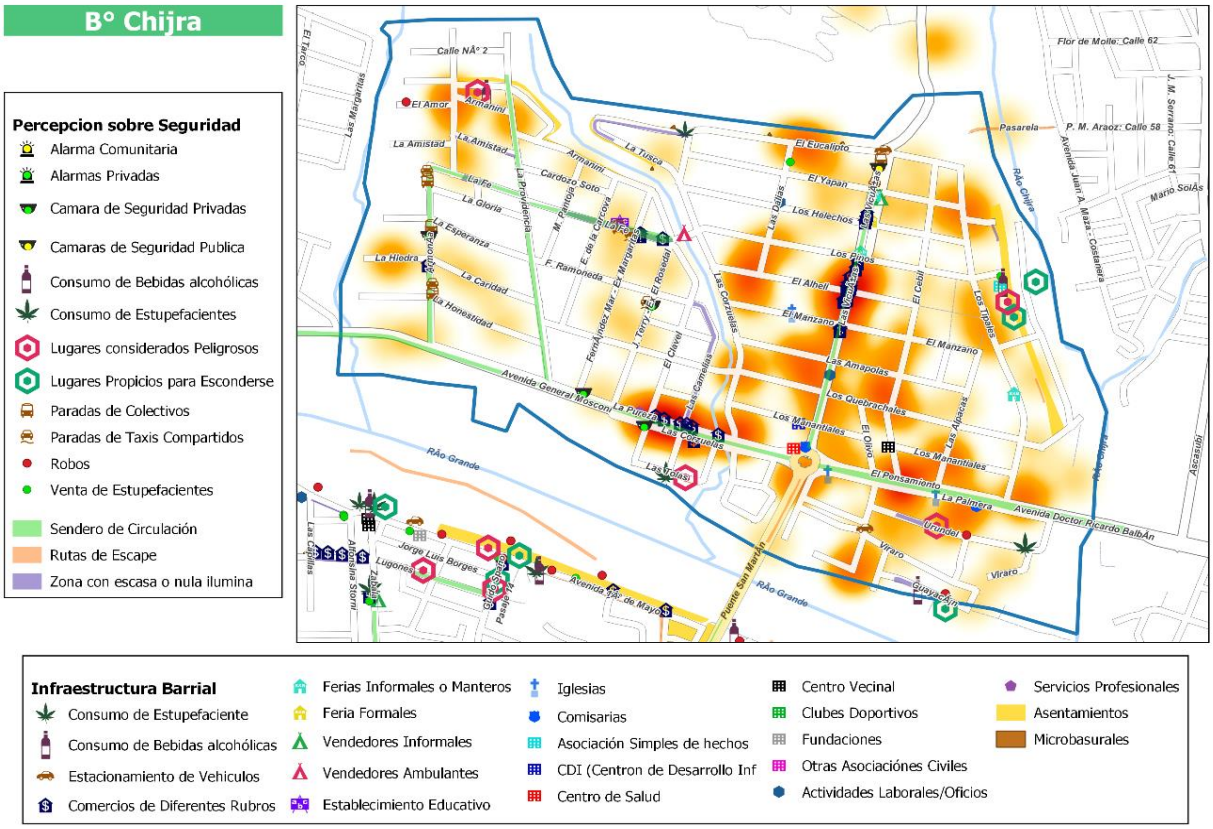


Figura 2: Mapa de estadístico espacial de Barrio de San José de Chijra

Análisis descriptivo del Barrio Alto Gorriti

En cuanto a su infraestructura barrial, podemos analizar según la figura 3, que el barrio se caracteriza por estar urbanizado y contar con los servicios básicos de luz, agua, gas natural, transporte público de pasajeros y red cloacal en toda su extensión. Las calles en su mayoría están pavimentadas y cuentan con iluminación, con lo cual se evidencia el abastecimiento a la población de ciertos productos considerados esenciales.

Respecto a la infraestructura social, se observa la existencia de organizaciones públicas como Establecimientos Educativos, CDI (Centros de Desarrollo Infantil), Comisarías y Centro de Salud, Además cuenta con una variedad de organizaciones de la sociedad civil como Iglesias, Asociaciones simples o de hecho (Comedores y merenderos) Club Deportivo, fundaciones y otras asociaciones civiles como Centro Vecinales, todas ellas distribuidas por distintas zonas del barrio.

Las actividades comerciales y laborales de diferentes rubros se encuentran concentradas principalmente en la calle Humahuaca entre Rinconada y Tumusla, y en la calle Cerro Aguilar entre Avda. El Éxodo y calle Humahuaca, si bien en el barrio se observa una importante actividad comercial en forma distribuida, en la intersección de las calles Humahuaca y Cerro Aguilar coincide con la densidad de datos por delitos contra la propiedad (zona caliente), según datos oficiales del CIAC.

Este barrio, resulta particular al momento de describir su conexión con el Barrio Mariano Moreno a través de la existencia de la feria formal e informal y la presencia de vendedores ambulantes ubicados en la Avda. El Éxodo y Horacio Guzmán.

Para analizar la percepción de seguridad en el barrio es importante tomar en cuenta la confluencia de diferentes variables, en el mapa se observa una importante concentración de actividades sociales, uso del espacio y predictores de movimiento que ponen de manifiesto la dinámica barrial y su vinculación con el delito. Así se puede observar que los vecinalistas circulan con mayor frecuencia por las calles: Caseros, Humahuaca y Avdas. Pueyrredón y El Éxodo, esto estaría coincidiendo con el uso del transporte público de pasajeros y taxis compartidos identificadas en la Avda. Pueyrredón y calle Caseros y con el acontecimiento de diversos delitos contra la propiedad (índice delictivo en color naranja y rojo).

La percepción de seguridad de los ciudadanos de este barrio, no estaría coincidiendo con las zonas con mayor registro de hechos delictivos las cuales están ubicadas en calle Cerro Aguilar y Humahuaca y con menor densidad, en calles Caseros y Avda. El Éxodo. Ya que las zonas percibidas por los referentes barriales como lugares peligrosos, se localiza en el pasaje ubicado entre calle Cerro Chañi y Avda. Pueyrredón, en la cual se identifica la confluencia de datos tales como: zona con escasa o nula iluminación, ruta de escape, consumo de alcohol y estupefacientes y lugar propicio para esconderse.

Alto Gorriti resulta particular al momento de analizar su conexión con el barrio Barrio Mariano Moreno a través de la existencia de la feria formal conocida como el “Big Mall” y la presencia de vendedores ambulantes ubicados en la zona comprendida entre la Avda. El Éxodo y Horacio Guzmán.

En la zona limítrofe con el barrio Mariano Moreno, más específicamente el Pasaje 44, los residentes de ambos barrios coinciden en que esta zona es peligrosa por la venta y consumo de estupefacientes, por constituirse en un lugar propicio para esconderse y por funcionar como zona de escape ante hechos delictivos.

Las zonas con escasa o nula iluminación se perciben en las calles Cerro Chañi, San Antonio y San Francisco, las cuales están más vinculadas a las zonas descritas anteriormente, de las cuales solo la Cerro Chañi y San Francisco son percibidas, por los referentes, como “inseguras”.

En cuanto a las medidas de seguridad pública, se observó la instalación de cámaras de seguridad públicas, éstas están ubicadas en distintos puntos del barrio, sin embargo, no se evidencian cámaras en las zonas con mayor índice delictivo (Humahuaca y Cerro Aguilar), excepto la cámara ubicada en el Pasaje 44.

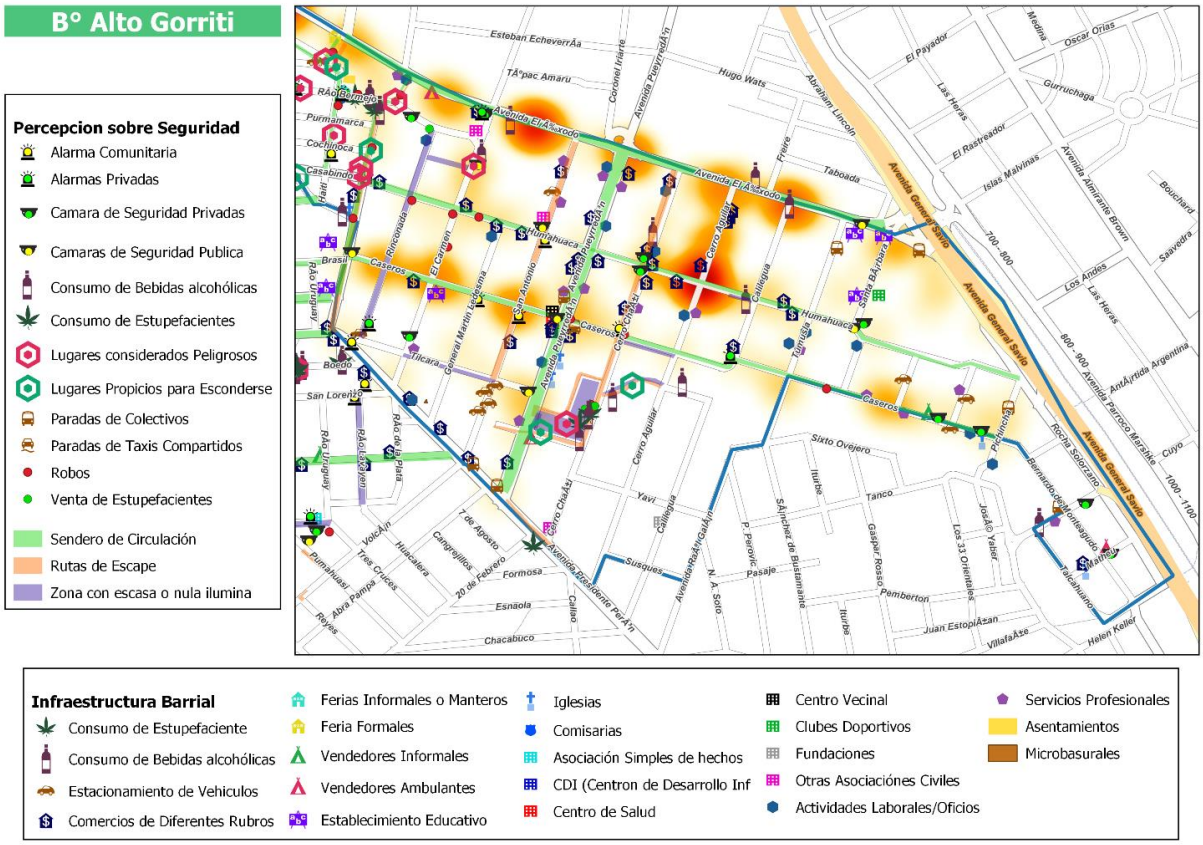


Figura 3: Mapa de estadístico espacial de Barrio de Alto Gorriti

Análisis descriptivo de Barrio Mariano Moreno

Los principales accesos al barrio (figura 4) se realizan por las calles Horacio Guzmán, Ejército del Norte, Falucho, República Dominicana y Colectora paralela a RN 9 a la altura de las calles Valdivia y Falucho, tiene la particularidad de contar con calles en pendiente, sobre todo a la altura de la Colectora paralela a RN 9. En general los servicios públicos están garantizados en todo el barrio, salvo algunos sectores en los cuales las calles no están pavimentadas. Cabe destacar que en calle San Lorenzo, actualmente se realizan obras de pavimentación.

El barrio se caracteriza por contar con un importante número de organizaciones públicas, sobre todo de establecimientos educativos de nivel inicial, primario y secundario tales como: Esc. Primaria N°171, Colegio Antonia María Gianelli, Esc. N° 100 Fco. de Argañaraz, Bachillerato Provincial N°21, Escuela Técnica N°2 “Prof. Jesús Raúl Salazar”, Jardín “Madre de Familia”, Centro de Formación Profesional N°1, Esc. Primaria Nocturna N° 397, Esc. Especial N°1 “Oscar Orias”, además cuenta con Seccional de Policía N°30 y el Centro de Salud.

Entre las organizaciones de la sociedad civil, el barrio cuenta con: la Iglesia “Sagrado Corazón de Jesús, Centro de día para Adultos Mayores, Centro Vecinal (sin edificio propio, pero con identificación por parte de referentes), Fundación “Gauchadas” y Fundación “Sonrisa”, Club Atlético Gral. Lavalle y Asociaciones civiles o de hecho como el Merendero “Arroz con Leche”, Radio Visión Jujuy y Canal 7 y, Radio Nacional.

Tiene una importante actividad comercial y laboral (comercios de diferentes rubros, oficios y servicios profesionales) distribuida por todo el barrio, pero principalmente concentrada en la Avda. Ejército del Norte y en la Avda. Horacio Guzmán. En ésta última se observa la presencia de comercios de diferentes rubros, de ferias formales como el “Big Mall” y de vendedores ambulantes, principalmente ubicados en el acceso al barrio que conecta la Vieja Terminal de Ómnibus y el barrio Alto Gorriti.

En este sentido es que los barrios Mariano Moreno y Alto Gorriti son considerados interesantes para ser abordados en el presente estudio, ya que es en este punto de encuentro se evidencia la zona de mayor densidad delictiva, específicamente en la zona comprendida entre calles Guayana, Cochinoca (Alto Gorriti), Panamá (Mariano Moreno) y Santa Catalina (Alto Gorriti).

Es importante destacar los distintos sectores que se presentan y cómo se relacionan con la infraestructura y percepción de los vecinos. En las calles Tobago y Santa Catalina se encuentra una plaza detrás de una zona comercial del barrio (descripta anteriormente). Este lugar es aprovechado para el consumo de estupefacientes, por tal razón es considerado como lugar peligroso y como ruta de escape, al encontrarse cerca de una gran zona comercial como es la zona de ex terminal y paradas de transporte público ubicados sobre Avda. El Éxodo. Asimismo, los datos oficiales registran un alto índice de delitos contra la propiedad, según mapa de calor (CIAC, 2021).

El Pasaje Boedo presenta menor circulación de peatones por sus alrededores, se caracteriza por el consumo de bebidas alcohólicas y de estupefacientes y es considerado por los referentes, como un lugar peligroso. Como medida de autoprotección vecinal, se observó la instalación de alarmas comunitarias localizadas en las calles Río Lavayen esq. San Lorenzo, Río Lavayen esq. Boedo, El Salvador esq. San Lorenzo y Horacio Guzmán esq. Boedo, todas ellas cercanas a esta zona.

En la zona limítrofe al barrio, sobre calle Iguazú, los referentes barriales la identifican como una zona en la cual se produce el consumo de bebidas alcohólicas y de estupefacientes, se añade a ello la escasa iluminación que favorece el acontecimiento de hechos delictivos y contravencionales.

Por último, es necesario destacar que en la plaza principal del B° Mariano Moreno se observa la circulación del transporte público de pasajeros (colectivos y taxis compartidos), sus calles aledañas se presentan con escasa iluminación (México y Colombia) y constituye un lugar para el consumo de drogas y bebidas alcohólicas (México, Yécora, Colombia y Perú). Esta situación estaría vinculada con la percepción que tienen los vecinos respecto al lugar, identificándose como zona peligrosa. Este dato cobra relevancia al momento de vincular esta situación con la presencia de distintas organizaciones ubicadas en la zona: la Esc. N° 100 Francisco de Argañaraz y el Bachillerato Prov. N° 21; la Iglesia Sgdo. Corazón de Jesús, la Seccional N° 30 y el Centro de Salud del barrio. En este sentido, este espacio recreativo y deportivo se encuentra atravesado por actividades ilícitas y contravencionales constantes que no se dispersan a pesar de la presencia de entidades públicas y de la sociedad civil, esto nos lleva a pensar en la necesidad de medidas urgentes que contribuyan a la prevención de estas situaciones.

Por último, es de destacar que las calles ubicadas alrededor de esta plaza central del barrio, tales como El Mollo, Ecuador y Yécora, se encuentran muy afectadas por delitos contra la propiedad, lo cual está próximo a situaciones que se manifiestan en calles Yecora y El Mollo, caracterizadas por el consumo de sustancias psicoactivas (droga y alcohol).

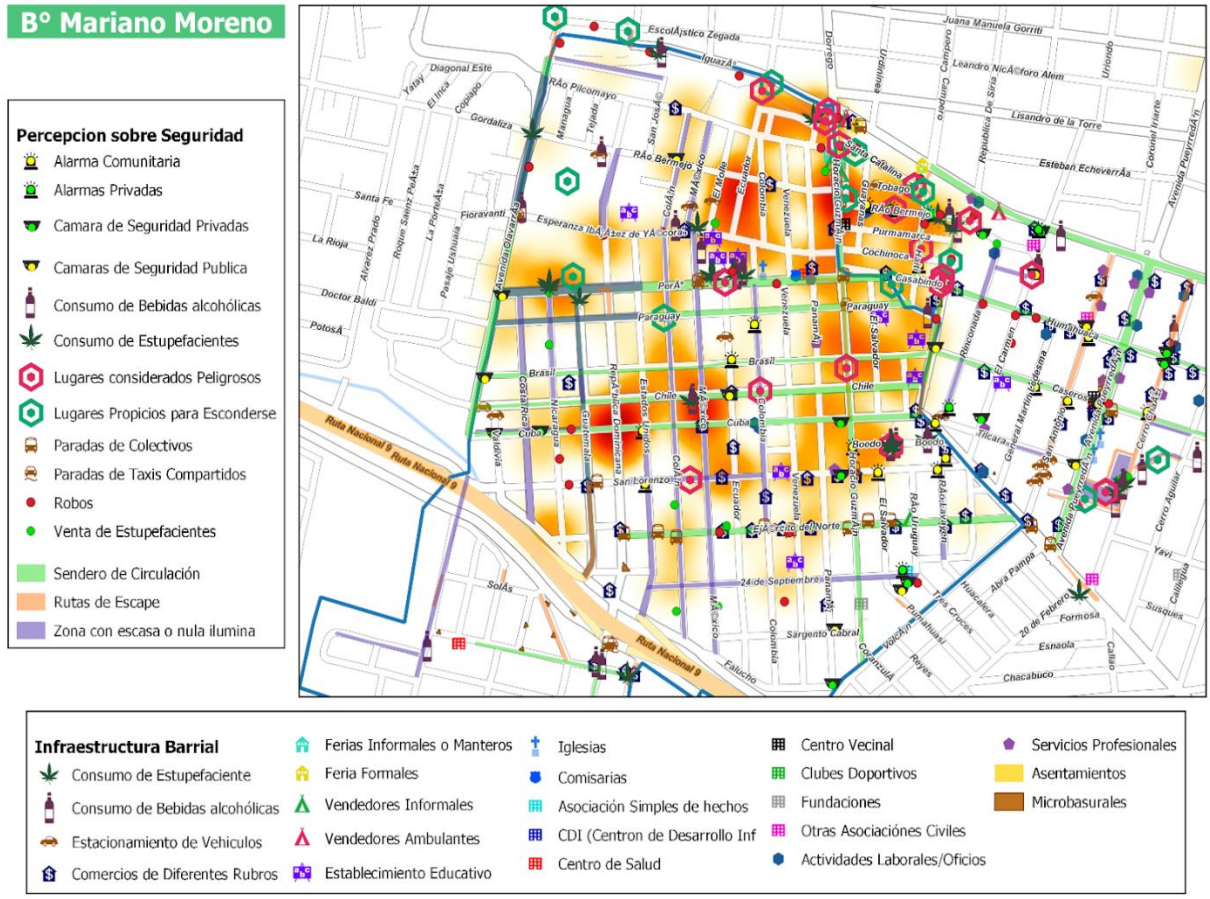


Figura 4: Mapa de estadístico espacial de Barrio de Mariano Moreno

CONCLUSIONES

El presente estudio se muestra como un primer avance sobre la vinculación entre las condiciones socioambientales y su vinculación con el delito. Permitió vincular las técnicas de observación en sitio, entrevistas a referentes vecinales en variables como son datos generales, infraestructura, percepción sobre el escenario barrial, percepción de sobre la seguridad y participación ciudadana con las fuentes oficiales de hechos delictuales.

Permitió arribar a algunas conclusiones preliminares que no solo contribuyen al estudio en sí de la inseguridad, sino a mejorar las técnicas de recolección y análisis de datos empleadas y al proceso que implica llegar a un conocimiento confiable.

Al ser datos recabados a partir de fuentes primarias, es necesario garantizar una mayor precisión de los datos obtenidos, esto requiere de un seguimiento y control permanente de los datos que garantice la fiabilidad de estos.

La técnica de estadística espacial es positiva en cuanto permite adquirir una mirada general de la situación, pero tiene como limitación que los datos identificados con precisión están estrechamente vinculados con el registro en el trabajo de campo, con lo cual la obtención de los datos debe ser lo más rigurosa posible, de tal manera que el producto obtenido en los mapas refleja de manera fiel el trabajo de campo.

En cuanto al objetivo del estudio, se logró construir diagnósticos preliminares sobre la situación de los diferentes barrios, si bien en esta presentación solo se describió algunos de los barrios seleccionados, se puede evidenciar que el estudio aportará a simple vista una densidad importante de datos que confluyen y refieren a un determinado sector de la realidad que permiten analizar la seguridad en cada uno de ellos.

Como se observa en los primeros análisis, cada barrio ofrece particularidades que lo hacen único, sin embargo, se pudo evidenciar cómo ciertas condiciones que hacen a la percepción de la seguridad por parte de los ciudadanos, se vinculan con el acontecimiento de hechos delictivos. Las actividades comerciales, la afluencia de personas en ciertos sectores y la venta y consumo de sustancias psicoactivas, son factores que confluyen al momento de construir una percepción en la comunidad.

Por otra parte, se considera que a mayor cantidad de datos que puedan aportar los ciudadanos, generará una mayor cantidad de información que permitirá lograr un conocimiento más certero.

Se considera que el resultado de esta investigación proporcionará información valiosa que contribuirá a los organismos gubernamentales a generar políticas públicas tendientes a la prevención, no solo de los hechos de inseguridad, sino también al abordaje de distintas situaciones problemáticas que afectan a las comunidades estudiadas en particular.

BIBLIOGRAFÍA

Anitua, G. (2010). Historia de los pensamientos criminológicos. Editores del Puerto.

Buzai, G., & Baxendale, C. (2012). Análisis socioespacial con sistemas de información geográfica. *Ordenamiento Territorial. Temáticas de base vectorial*, 315.

Miranda Salas, M., & Condal, A. (2003). Las técnicas de estadística espacial en la investigación salubrista. Caso síndrome de Down. *Bosque (Valdivia)*.

Sozzo, M. (2008). Gobierno local y prevención del delito en Argentina. *En Urvio Regista Latinoamericana de Seguridad Ciudadana*.

Varela, C. (2010). Los modelos preventivos. *Cuaderno de Seguridad. Comunidad y Seguridad. Ministerio de Justicia, Seguridad y Derechos Humanos*.

Vieytes, R. (2004). *Metodología de la Investigación en organizaciones, mercado y sociedad, epistemología y técnicas.* Buenos Aires: De las Ciencias.