Informe final publicable de proyecto
Estadística para datos complejos

Código de proyecto ANII: FCE_3_2022_1_172289

Fecha de cierre de proyecto: 01/10/2025

MORENO ROMERO, Leonardo Fabian (Responsable Técnico - Científico)
BERRENDERO, José Ramón (Investigador)
CHOLAQUIDIS NOBLÍA, Alejandro (Investigador)
CUEVAS, Antonio (Investigador)
ESTRAMIL, Agustin (Investigador)
FRAIMAN, Ricardo (Investigador)
GAMBOA, Fabrice (Investigador)
HERNÁNDEZ BANADIK, Manuel (Investigador)
JOLY, Emilien (Investigador)
PATEIRO LOPEZ, Beatriz (Investigador)

UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN (Institución Proponente) \\ UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN \\ UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE CIENCIAS

Resumen del proyecto

En los últimos años ha aumentado mucho la cantidad de datos complejos y muy diversos, y eso exige nuevas herramientas para poder analizarlos de forma confiable. Ese fue uno de los principales desafíos de este proyecto. En ese marco, se lograron avances importantes en varias líneas de trabajo. Se desarrolló un nuevo método robusto de agregación, pensado para combinar información de manera estable incluso cuando los datos presentan valores atípicos o comportamientos poco usuales. También se amplió la inferencia conforme a problemas de regresión en espacios geométricos más generales, se propusieron nuevas formas de medir qué tan lejos está un conjunto de ser convexo, se construyó un estimador corregido de una propiedad geométrica llamada standardness y se introdujo el concepto de alcance polinomial junto con una forma de estimarlo a partir de datos.

Estos desarrollos se estudiaron tanto desde el punto de vista teórico como mediante simulaciones y aplicaciones con datos reales, mostrando su utilidad en distintos contextos. Además, el proyecto impulsó nuevas metodologías de clasificación no supervisada, fortaleciendo una línea de investigación que ya venía desarrollándose y abriendo nuevas posibilidades de trabajo futuro.

Por otra parte, los resultados no quedaron restringidos al ámbito de la investigación. El proyecto promovió activamente su difusión en congresos y jornadas científicas, y también incluyó actividades de divulgación y vinculación con otros sectores, como la industria y estudiantes de educación media.

En conjunto, el proyecto contribuyó a fortalecer las capacidades nacionales en metodología estadística y en su aplicación a problemas actuales cada vez más complejos.

Ciencias Naturales y Exactas / Matemáticas / Estadística y Probabilidad / Inferencia en espacios abstractos
Palabras clave: Estimación de conjuntos / Inferencia conforme / Estadística en variedades y con datos funcionales. /

Antecedentes, problema de investigación, objetivos y justificación.

En las últimas décadas, la investigación estadística ha debido ampliar de manera sustancial su campo de acción para responder a nuevas formas de producción y disponibilidad de datos. La masificación de los registros digitales, el aumento de la dimensión de los problemas y la aparición de observaciones con estructura geométrica compleja han puesto en evidencia las limitaciones de los enfoques clásicos, concebidos principalmente para datos multivariados en espacios euclidianos de dimensión finita. En muchos problemas actuales, los datos ya no pueden ser tratados adecuadamente como vectores reales ordinarios, sino que toman valores en espacios métricos generales, variedades riemannianas, espacios funcionales o conjuntos con propiedades geométricas específicas. Esta transformación del escenario metodológico ha impulsado el desarrollo de nuevas herramientas estadísticas capaces de incorporar la estructura intrínseca de los datos y de producir inferencias válidas bajo supuestos más realistas.

En ese contexto se inscribió el proyecto Estadística para datos complejos, cuyo propósito fue abordar problemas actuales y relevantes de la estadística matemática y aplicada cuando los datos, por su propia naturaleza o por transformaciones necesarias para su análisis, habitan espacios complejos. Desde su formulación, la propuesta se ubicó en un punto de convergencia entre la estadística en espacios abstractos, la inferencia no paramétrica, la geometría de conjuntos y la estadística sobre variedades. Esta articulación respondió a una necesidad concreta: avanzar hacia métodos que, sin abandonar el rigor teórico, fueran capaces de dialogar con aplicaciones reales en ámbitos donde la geometría y la complejidad estructural de los datos no pueden ser ignoradas. El proyecto se planteó explícitamente como objetivo base el desarrollo de resultados teóricos publicables en revistas de alto impacto, junto con su validación a través de simulaciones y aplicaciones concretas.

El punto de partida conceptual del proyecto fue reconocer que la complejidad de los datos no es un fenómeno marginal, sino una característica central de buena parte de los problemas científicos recientes. En medicina, biología, ingeniería, meteorología, bioinformática o reconocimiento de patrones, es frecuente encontrar datos funcionales, direcciones, subespacios, matrices de covarianza, curvas, trayectorias o imágenes cuyo análisis exige herramientas específicas. La literatura internacional ha mostrado un crecimiento sostenido en estas áreas, pero también la persistencia de vacíos relevantes, sobre todo en lo relativo a la construcción de procedimientos generales con respaldo teórico sólido, capaces de incorporar restricciones geométricas y de operar más allá de los marcos euclidianos tradicionales. En este sentido, la propuesta se apoyó en antecedentes bien establecidos, pero buscó a la vez avanzar sobre problemas concretos, de este marco más general, que permanecían abiertos o insuficientemente desarrollados.

Una primera línea central del proyecto fue la estimación de conjuntos y de funcionales geométricos asociados. Esta área cuenta con antecedentes importantes en la literatura de estimación de soporte y de conjuntos, donde trabajos pioneros mostraron cómo reconstruir o aproximar un conjunto a partir de una muestra de puntos vinculada con él. Sin embargo, el interés actual ya no se limita a estimar el soporte de una distribución, sino que se extiende a funcionales geométricos más delicados, como el borde, la superficie, el perímetro, el volumen o medidas que describen la forma del conjunto. El proyecto se planteó aportar en esta dirección mediante el estudio de nuevos estimadores y nuevas propiedades geométricas, con especial énfasis en problemas como la convexidad, las medidas de no convexidad y los conjuntos de volumen polinomial. La importancia de esta agenda radica en que permite vincular geometría, probabilidad y estadística de una manera fértil, y abre la puerta a aplicaciones en imágenes médicas, ecología espacial, trayectorias de individuos y otras áreas donde los objetos de interés tienen naturaleza geométrica.

Dentro de esta misma línea, la propuesta otorgó particular relevancia al estudio de conjuntos cuyo volumen dilatado presenta comportamiento polinomial. Este problema no es meramente técnico: se vincula con resultados clásicos de Federer y con la posibilidad de interpretar los coeficientes del polinomio en términos de funcionales geométricos del conjunto, incluyendo invariantes topológicos. Desde una perspectiva estadística, esto genera una agenda de trabajo especialmente interesante, porque permite pensar en métodos de estimación, comparación y eventualmente contraste para propiedades globales de objetos geométricos a partir de observaciones muestrales. Asimismo, el proyecto contempló el análisis de esquemas de muestreo alternativos que pudieran mejorar el desempeño de los estimadores y reducir la dependencia respecto de la dimensión ambiente, un punto crucial cuando se trabaja con objetos inmersos en espacios de alta dimensión. Cabe destactar la fortaleza de los investigadores nacionales e internacionales del proyecto en las áreas de investigación mencionadas.

La segunda gran línea del proyecto fue la inferencia conforme en modelos de regresión con respuesta en variedades riemannianas. La estadística sobre variedades constituye hoy un área de fuerte desarrollo internacional, impulsada por aplicaciones donde las observaciones viven naturalmente en espacios no lineales. Este es el caso, por ejemplo, de datos direccionales, datos sobre cilindros o toros, matrices definidas positivas, subespacios asociados a análisis de componentes principales o representaciones geométricas en visión por computador. En todos estos casos, la ausencia de estructura vectorial lineal obliga a reformular conceptos y procedimientos estadísticos básicos. El proyecto tomó este desafío y se propuso desarrollar técnicas de inferencia conforme para contextos de regresión en los que la variable respuesta se encuentra sobre una variedad riemanniana, en particular sobre variedades como la de Stiefel o la Grassmanniana.

La elección de la inferencia conforme como eje metodológico respondió a su gran potencial para construir regiones o bandas de predicción con garantías de cobertura, bajo supuestos distribucionales mínimos. En un escenario donde una de las preocupaciones principales de los métodos predictivos es cuantificar la fiabilidad de sus salidas, la inferencia conforme ofrece una respuesta particularmente atractiva: permite producir conjuntos de predicción distribution-free o, al menos, con requerimientos muy débiles sobre la distribución conjunta de los datos. Aunque esta metodología ha tenido un desarrollo vertiginoso en los últimos años, su adaptación a contextos geométricos complejos y, en particular, a la regresión sobre variedades, seguía siendo un terreno escasamente explorado. El proyecto se propuso avanzar tanto en la formulación de modelos adecuados como en la construcción de bandas de confianza y en el estudio de propiedades asintóticas y tasas de convergencia.

La tercera línea de trabajo se concentró en el problema del clustering en espacios de alta dimensión y en datos funcionales. Tanto la estadística funcional como la estadística de alta dimensión han crecido de manera sostenida en lo que va del siglo, debido a la proliferación de datos registrados de forma continua y a la aparición de problemas donde el número de variables es grande y las asociaciones entre componentes son intensas. En esos contextos, los métodos de clasificación no supervisada ocupan un lugar central, pero presentan dificultades

conceptuales y prácticas relevantes. El proyecto se propuso justamente comenzar a estudiar nuevos métodos de agrupamiento que incorporaran restricciones geométricas sobre los grupos, por ejemplo en datos composicionales, buscando mayor interpretabilidad y mejores fundamentos metodológicos.

Desde el punto de vista de sus antecedentes científicos, el proyecto se sustentó en una base sólida construida por el propio equipo de trabajo y por sus colaboradores nacionales e internacionales. La propuesta destacó que el grupo cuenta con experiencia previa en estadística en espacios abstractos y, en particular, en estimación de conjuntos, estadística sobre variedades y datos funcionales. Entre los antecedentes reseñados se incluyeron aportes en profundidad estadística sobre variedades, detección de outliers, clasificación supervisada, estimación de densidades y conjuntos de nivel en variedades, análisis de sensibilidad en espacios métricos generales y diversos problemas de estimación de conjuntos, incluyendo estimación de borde, restricciones geométricas y aplicaciones al home-range de especies. También se señaló la relevancia del grupo en estadística funcional, con contribuciones reconocidas internacionalmente. Todo ello daba al proyecto un punto de partida especialmente favorable, tanto por la acumulación de conocimiento como por la articulación entre perfiles matemáticos y estadísticos.

En el plano nacional, la propuesta subrayó además el carácter singular del grupo, al señalar que se trata del único equipo en el país trabajando de manera sistemática en estadística en espacios abstractos y en los problemas específicos considerados por el proyecto. Esa singularidad constituye, al mismo tiempo, una fortaleza y una responsabilidad. Por un lado, posiciona al proyecto como una iniciativa estratégica para el desarrollo de capacidades científicas avanzadas en Uruguay; por otro, refuerza la importancia de consolidar masa crítica y formar nuevos investigadores en estas áreas. En esta dirección, la propuesta contempló explícitamente la participación de estudiantes de maestría y la articulación con instituciones nacionales relevantes.

A nivel internacional, el proyecto se insertó en una red de colaboración activa con investigadores de Francia, España y México, varios de los cuales integraron formalmente el equipo. Esta dimensión internacional fue relevante no solo por el respaldo académico que aporta, sino porque favorece la circulación de ideas, la construcción de agenda común y la proyección de los resultados hacia foros especializados de alto nivel. En un campo como el de la estadística para datos complejos, donde los desarrollos más significativos suelen surgir en la intersección entre teoría profunda y aplicaciones novedosas, la articulación con grupos consolidados del exterior constituye un componente importante para la calidad científica y visibilidad del Proyecto.

Asimismo, desde su concepción, el proyecto no se limitó a la generación de resultados teóricos, sino que incorporó como componente sustantivo su difusión y divulgación. En efecto, se previó la comunicación de los avances y resultados en seminarios nacionales y congresos internacionales, con el propósito de integrarlos a la discusión académica. Al mismo tiempo, se contempló la divulgación de las herramientas desarrolladas hacia sectores no estrictamente académicos, incluyendo potenciales usuarios de áreas tecnológicas y otros actores sociales para los cuales estas metodologías pudieran resultar de utilidad. De este modo, la propuesta asumió que la producción de conocimiento en estadística para datos complejos debía articularse no solo con la investigación de frontera, sino también con su circulación, apropiación y proyección hacia ámbitos más amplios de la sociedad.

Metodología/Diseño del estudio

La metodología del proyecto se organizó a partir de una lógica común a todas sus líneas de trabajo: formular con precisión matemática el problema estadístico de interés, identificar los supuestos geométricos o probabilísticos mínimos que permiten tratarlo con rigor, construir estimadores o procedimientos nuevos, estudiar sus propiedades teóricas y, finalmente, evaluar su comportamiento mediante simulaciones y ejemplos de aplicación. Esta estrategia ya estaba explícitamente prevista en la propuesta original, donde se establecía que, para cada problema, se analizarían consistencia, comportamiento asintótico, desigualdades de concentración, factibilidad computacional y utilidad en datos reales.

El proyecto en general adoptó un enfoque no paramétrico o semiparamétrico, buscando construir herramientas válidas bajo supuestos débiles, pero suficientemente estructurados como para permitir demostraciones rigurosas y algoritmos implementables. En todos los casos, la metodología combinó cuatro planos: formulación abstracta del problema, desarrollo teórico, diseño computacional e ilustración empírica.

En la línea de agregación robusta, la idea metodológica central consistió en transformar un estimador potencialmente frágil en otro resistente a contaminación y valores atípicos. Para ello se siguió un esquema por bloques: la muestra total se divide en K submuestras independientes, se calcula un estimador en cada una y luego esos estimadores se agregan mediante un criterio robusto definido en un espacio métrico general. El procedimiento GROS fue formulado como un problema de minimización sobre dicho espacio, basado en la cercanía a una mayoría de estimadores parciales. A partir de esa construcción, el trabajo metodológico se concentró en demostrar cotas de tipo sub-gaussiano, estudiar el breakdown point en el sentido de Donoho y, además, proponer una versión computacionalmente viable en la que la minimización se restringe al conjunto finito de estimadores obtenidos en las submuestras. La evaluación metodológica de GROS se completó con cinco estudios de simulación en problemas heterogéneos: clustering, bandits, regresión, estimación de conjuntos y topological data analysis.

En la línea de inferencia conforme sobre variedades, la metodología partió de una dificultad de base: cuando la respuesta vive sobre una variedad riemanniana, no pueden trasladarse automáticamente los argumentos usuales de regresión e inferencia desde el caso euclidiano. Por eso, el trabajo comenzó por fijar un marco de hipótesis geométricas y analíticas suficientemente generales: variedad compacta o tratable, regularidad C^2
, existencia de densidades conjuntas y condicionales, continuidad tipo Lipschitz en la covariable y condiciones de regularidad sobre los conjuntos de nivel de la densidad condicional. Sobre esa base se adaptó el esquema de Lei y Wasserman al caso manifold-valued, reemplazando los objetos euclidianos por sus análogos geométricos. Metodológicamente, esto exigió demostrar primero la convergencia uniforme del estimador kernel de densidad sobre la variedad, luego definir regiones de predicción localmente válidas a partir de particiones del espacio de covariables y, finalmente, estudiar la convergencia casi segura de esas regiones empíricas hacia sus contrapartes poblacionales. La metodología incluyó también un tratamiento diferenciado de puntos cercanos al borde de la variedad, así como una discusión computacional para contextos de mayor dimensión o complejidad geométrica. Las simulaciones sobre la esfera y la variedad de Stiefel, junto con ejemplos reales asociados a viento y salidas probabilísticas en simplex, cumplieron el papel metodológico de contrastar cobertura, tamaño de región y factibilidad numérica.

En la línea dedicada a standardness, la metodología fue de naturaleza más fina y conceptual. El primer paso consistió en clarificar la definición de la propiedad y su relación con otras restricciones geométricas usadas en estimación de conjuntos, como convexidad, reach positivo, r-convexidad, rolling condition y cone-convexidad. A partir de esa revisión, el problema estadístico se reformuló en términos de un parámetro: la constante de standardness. Sobre esa base, se analizó primero un estimador plug-in ya sugerido en la literatura, demostrando su consistencia casi segura bajo condiciones generales. A partir de evidencia numérica se detectó una subestimación sistemática en ciertos escenarios, por lo que se introdujo una corrección de sesgo que incorpora información sobre la frecuencia con que el mínimo local es casi alcanzado en la muestra. Así, el desarrollo combinó argumento asintótico, diagnóstico numérico y corrección metodológica. Finalmente, el problema fue llevado a un plano más fundamental: se probó que, aun cuando la constante puede estimarse, no es posible decidir a partir de una muestra finita si una medida satisface o no la propiedad de standardness.

En la línea de medidas de no convexidad, la metodología combinó geometría, inferencia y computación. Primero se seleccionaron y formalizaron distintas nociones de alejamiento respecto de la convexidad, apoyadas en caracterizaciones complementarias del concepto: distancia al casco convexo, diferencia en medida, probabilidad de visibilidad y un nuevo índice basado en la inclusión del simplex generado por
$d+1$ puntos aleatorios dentro del conjunto. Luego se establecieron propiedades básicas de esas medidas: invariancia por traslación y escala, rango entre 0 y 1, y capacidad de detectar convexidad. En una segunda etapa se construyeron estimadores a partir de muestras aleatorias, en algunos casos mediante enfoques plug-in basados en estimadores del soporte, y en otros por aproximaciones Monte Carlo. La metodología incluyó además el estudio de consistencia y distribuciones límite, de modo de dotar a estos índices de una base inferencial y no meramente descriptiva. Finalmente, se desarrollaron procedimientos concretos de implementación: uso del casco convexo, r-convex hull, diagramas de Voronoi, búsqueda rápida de vecinos y algoritmos para contar puntos dentro de triángulos o símplices. El ejemplo con imágenes dermatoscópicas permitió mostrar cómo estos índices pueden operar como descriptores geométricos en problemas aplicados.

En la línea de alcance polinomial, la metodología estuvo orientada a convertir una propiedad geométrica profunda en un procedimiento estadístico operativo. El punto de partida fue la función de volumen dilatado $V(r)$, cuya forma polinómica en un intervalo inicial permite recuperar coeficientes con interpretación geométrica, en particular volumen y medida de borde. Sobre esa idea se definió formalmente el alcance polinomial como el mayor radio para el cual esa propiedad polinómica se mantiene. La estrategia metodológica consistió en estimar primero la función de

volumen a partir de una muestra interna del conjunto, sin requerir observaciones externas ni reconstrucción previa del soporte, y luego aproximarla por el polinomio de mejor ajuste en norma L^2. A partir de allí se desarrollaron dos vías para tratar el radio desconocido: un estimador consistente del alcance y, de manera más conservadora, un procedimiento de infraestimación, preferible en la práctica para evitar ajustar un polinomio fuera de su zona de validez. Una vez estimado ese intervalo útil, los coeficientes del polinomio se calcularon por mínimos cuadrados sobre una grilla. Las simulaciones en distintos conjuntos sintéticos se usaron para comparar precisión en las estimaciones.

En la línea de clasificación no supervisada, ver adjunto, se desarrolló además una metodología orientada a construir particiones interpretables en datos composicionales, es decir, observaciones restringidas al simplex y definidas en términos relativos. El enfoque partió de reconocer que, en este tipo de datos, las distancias euclidianas usuales no respetan adecuadamente su geometría, en particular cuando existen ceros o masa cerca de la frontera del simplex. Por ello, se trabajó con una adaptación del método CUBT (Clustering Using Unsupervised Binary Trees), que combina la lógica de partición recursiva de árboles binarios con reglas explícitas de separación, fácilmente traducibles a la escala original de las composiciones. La metodología consideró tres representaciones: una basada en log-ratios, otra basada en la distancia de Hellinger y una propuesta híbrida que interpola entre ambas según la cercanía a la frontera, de modo de preservar la interpretabilidad relativa en el interior del simplex y mantener estabilidad en presencia de ceros. El procedimiento incluyó etapas de crecimiento del árbol, poda y fusión de hojas, seguidas de una retrotraducción de las reglas al espacio original, lo que permitió obtener grupos no solo homogéneos, sino también descritos por condiciones transparentes sobre las componentes. La estrategia fue evaluada mediante simulaciones con mezclas Dirichlet y escenarios con ceros estructurales, y se ilustró con una aplicación a datos de turismo receptivo en Uruguay, donde permitió identificar perfiles de gasto con interpretación sustantiva clara.

Resultados, análisis y discusión

El proyecto alcanzó resultados relevantes y consistentes en todas las líneas de trabajo previstas, combinando desarrollos teóricos, validaciones por simulación, propuestas metodológicas originales y potenciales aplicaciones en la estadística. En conjunto, los avances obtenidos muestran una contribución significativa en áreas de creciente importancia, particularmente en contextos donde los datos presentan estructuras complejas, geometrías no euclidianas, contaminación por valores atípicos o supuestos clásicos difíciles de sostener en la práctica. Una característica transversal del proyecto fue la búsqueda de métodos robustos, flexibles y matemáticamente bien fundamentados, capaces de responder a desafíos reales en análisis de datos de alta complejidad.

Uno de los aportes más destacados se produjo en la línea de agregación robusta, donde se introdujo el método GROS. Esta propuesta se apoya en una idea general y poderosa: dividir la muestra en varios subconjuntos, calcular estimadores locales sobre cada uno de ellos y luego combinarlos mediante un procedimiento robusto, análogo en espíritu al esquema median-of-means. La ventaja principal de este enfoque es que permite transformar estimadores potencialmente sensibles a contaminación en procedimientos mucho más estables frente a la presencia de valores extremos. Desde el punto de vista teórico, se probó que el estimador agregado resultante posee comportamiento sub-gaussiano y un breakdown point elevado, lo que garantiza una resistencia sustancial a outliers. En términos prácticos, esto significa que el método conserva precisión incluso cuando una parte no despreciable de los datos se encuentra fuertemente contaminada.

Los experimentos realizados confirmaron estas propiedades teóricas. En simulaciones con distintos niveles de contaminación, GROS mantuvo errores de estimación reducidos, y el deterioro de su desempeño fue considerablemente menor que el observado en procedimientos clásicos. En escenarios de regresión con valores extremos, por ejemplo, el método superó a alternativas estándar y mostró una notable estabilidad al aumentar la proporción de observaciones aberrantes. Al mismo tiempo, cuando los datos no presentan outliers, GROS coincide esencialmente con los métodos usuales, por lo que no sacrifica eficiencia en situaciones favorables. Esta doble virtud (robustez cuando la contaminación existe y competitividad cuando no existe) constituye uno de los puntos fuertes del desarrollo. A ello se suma su generalidad, ya que puede construirse sobre cualquier estimador base definido en espacios métricos, incluyendo no solo contextos euclidianos, sino también problemas en espacios más abstractos de proximidad o similitud. El análisis de consistencia mostró además que, a medida que aumenta el tamaño muestral, el estimador converge al valor verdadero con alta probabilidad. Desde el punto de vista computacional, su estructura por bloques facilita la paralelización y lo vuelve especialmente atractivo para aplicaciones en ciencia de datos, finanzas u otros ámbitos donde se trabaja con grandes volúmenes de información. En el trabajo realizado se presentaron diversos ejemplos de aplicación que ponen de manifiesto el gran interés y la amplia proyección del método generado.

Una segunda línea de resultados especialmente relevante fue la inferencia conforme en variedades. En este caso, el proyecto abordó el problema de construir regiones predictivas cuando la variable respuesta toma valores en una variedad riemanniana, es decir, en un espacio geométrico curvo donde las nociones usuales de promedio, distancia o dispersión no pueden tratarse del mismo modo que en el espacio euclidiano. Esta cuestión es de gran interés actual, dado que numerosos datos modernos (direcciones, orientaciones, formas, trayectorias, gestos o configuraciones geométricas) viven naturalmente en variedades. El desarrollo realizado permitió extender el paradigma de la inferencia conforme a este contexto, obteniendo regiones predictivas distribution-free, es decir, válidas sin necesidad de imponer supuestos paramétricos sobre la distribución de los datos.

El carácter distribution-free es particularmente valioso, porque otorga garantías de cobertura que no dependen de una correcta especificación del modelo probabilístico subyacente. En el proyecto se demostró que las regiones construidas presentan cobertura asintótica correcta, y esta propiedad fue respaldada por estudios de simulación y por ejemplos aplicados. Se obtuvieron regiones predictivas más ajustadas y menos conservadoras que aquellas derivadas de aproximaciones euclidianas ingenuas. Además, se verificó no solo la cobertura promedio global, sino también un comportamiento sólido frente a heterocedasticidad y distribuciones con colas pesadas, escenarios en los que los métodos paramétricos tradicionales suelen fallar. El enfoque desarrollado, por lo tanto, amplía de forma sustancial el alcance de la inferencia conforme y abre posibilidades concretas en aplicaciones geoespaciales, biomédicas y de análisis multivariado con estructura geométrica. También se desarrollaron algoritmos que funcionan de forma correcta cuando la dimensión del espacio es elevada.

En la temática de standardness, el proyecto produjo avances de naturaleza más teórica, pero con consecuencias metodológicas importantes para el Análisis Topológico de Datos. La condición de standardness aparece como una hipótesis geométrica relevante en varios procedimientos de reconstrucción y estimación de conjuntos, ya que controla, de algún modo, la regularidad local del soporte de la distribución. En este marco, se propuso un estimador plug-in consistente para el parámetro asociado a dicha condición, aprovechando información geométrica del conjunto observado. Adicionalmente, se diseñó una corrección de sesgo orientada a mejorar el desempeño en muestras pequeñas, con resultados numéricos alentadores en simulaciones.

Más allá del valor puntual del estimador, uno de los aportes más significativos de esta línea fue la identificación de un límite intrínseco del problema: se mostró que no es posible decidir de manera concluyente, a partir de datos finitos, si un conjunto satisface o no la condición de standardness. Este resultado negativo tiene gran importancia conceptual, porque delimita con claridad qué puede y qué no puede esperarse de cualquier procedimiento inferencial sobre esta propiedad. En otras palabras, no se trata solo de una dificultad técnica de un método particular, sino de una imposibilidad estadística más profunda. Sin embargo, el hecho de que la propiedad no sea testeable en sentido estricto no le quita interés al parámetro de standardness: su estimación aproximada sigue siendo informativa y útil, ya que permite describir cuán regular es la distribución local de masa en el soporte y anticipar cómo se comportan ciertas características topológicas al dilatar el conjunto. El estudio realizado también mostró que la precisión de la estimación mejora en regiones con mayor densidad de observaciones, lo que conecta la noción de standardness con una idea de uniformidad local de volumen. De este modo, el proyecto aportó tanto herramientas concretas de estimación como una mejor comprensión de los alcances y límites de la inferencia geométrica en este problema.

Otra línea donde se obtuvieron resultados sólidos fue la de medidas de no convexidad. Allí se introdujo un nuevo índice y se estudiaron, en paralelo, varios índices conocidos. El objetivo general fue cuantificar, de manera rigurosa y estadísticamente operativa, cuánto se aparta un conjunto de la convexidad. Este tipo de problema tiene interés no solo geométrico, sino también aplicado, ya que la convexidad o su ausencia pueden servir para caracterizar formas, detectar irregularidades y construir descriptores útiles en visión computacional, ingeniería o control de calidad.

Los resultados teóricos mostraron que todos los índices considerados toman valores en el intervalo entre 0 y 1, y que, además, detectan correctamente la convexidad en el sentido esperado: por ejemplo, uno de ellos se anula si y solo si el conjunto es convexo. A partir de esta base, se construyeron estimadores consistentes a partir de muestras aleatorias y se obtuvieron distribuciones límite, lo que proporciona un marco inferencial completo para trabajar con estas cantidades. En el plano computacional, los índices fueron implementados y probados sobre datos reales de nubes de puntos irregulares, mostrando buena factibilidad numérica. El nuevo índice propuesto evidenció una sensibilidad particular para detectar huecos y bordes internos, rasgo que lo distingue favorablemente respecto de algunas alternativas previas. Esta capacidad puede resultar especialmente valiosa cuando se busca caracterizar defectos geométricos, discontinuidades o anomalías estructurales. Asimismo, la posibilidad de utilizar estos índices como variables explicativas o rasgos geométricos dentro de métodos de aprendizaje automático

abre una vía muy interesante de articulación entre teoría geométrica e inteligencia artificial. En particular, su incorporación en tareas de clasificación podría mejorar la detección de patrones anómalos o la diferenciación entre clases con estructuras espaciales complejas.

Finalmente, en la línea vinculada al alcance polinomial de un conjunto, el proyecto desarrolló una propuesta original con gran potencial metodológico. Se definió el alcance polinomial como el mayor intervalo inicial en el cual el volumen dilatado del conjunto crece de manera polinómica. Esta noción puede entenderse como una generalización del reach de Federer en un contexto estadístico y se relaciona directamente con fórmulas geométricas clásicas, como el teorema de Steiner. A partir de esta definición, se propuso un procedimiento para estimar dicho alcance utilizando únicamente datos internos del conjunto, sin necesidad de introducir parámetros de suavizado.

La ausencia de parámetros de suavizado representa una ventaja importante, porque evita un problema recurrente en estadística geométrica: la necesidad de calibrar cuidadosamente un parámetro que puede afectar de manera sensible la estimación final. En este caso, el método permite obtener estimaciones precisas de cantidades geométricas como volumen y perímetro de manera más directa y estable. Las aplicaciones sintéticas mostraron que el procedimiento logra recuperar con gran precisión valores teóricos en conjuntos tridimensionales simulados. Desde una perspectiva aplicada, esto sugiere utilidad en escenarios donde la precisión geométrica es central, como la microfotografía, la inspección de objetos tridimensionales, la visualización científica o ciertos problemas de visión computacional. Conceptualmente, el desarrollo también resulta valioso porque vincula una propiedad geométrica profunda con un procedimiento estadístico concreto y operativo. Así, el proyecto no solo amplió el repertorio de herramientas disponibles para la inferencia sobre conjuntos, sino que lo hizo mediante una metodología elegante, interpretable y de implementación relativamente simple.

Por otro lado, en relación con la línea de clasificación no supervisada, el proyecto también promovió desarrollos metodológicos en clustering, en particular orientados a incorporar interpretabilidad mediante restricciones geométricas sobre los grupos obtenidos. Si bien esta línea tuvo avances conceptuales y metodológicos durante el período de ejecución, no llegó a traducirse en artículos publicados dentro del plazo del proyecto. No obstante, se adjunta un trabajo aún no sometido a revisión pero desarrollado en el proyecto, en el que se explora este enfoque en el contexto de datos composicionales, es decir, datos cuyas observaciones se encuentran en el simplex unitario. En ese marco, el objetivo es construir herramientas de agrupamiento que, además de su buen desempeño, permitan una interpretación más clara de la estructura de los datos respetando su geometría natural.

Consideradas en conjunto, estas líneas de resultados permiten afirmar que el proyecto cumplió de manera integral los objetivos planteados y, en varios aspectos, los superó. En primer lugar, se obtuvieron aportes originales en problemas centrales de estimación de conjuntos, robustez e inferencia predictiva, todos ellos respaldados por análisis matemáticos rigurosos. En segundo lugar, se desarrollaron herramientas con vocación aplicada, capaces de trasladarse a contextos reales donde los supuestos tradicionales no son adecuados. En tercer lugar, el proyecto generó una base conceptual y metodológica que habilita nuevas líneas de investigación, algunas de las cuales ya comenzaron a delinearse, por ejemplo en clasificación no supervisada y clustering robusto inspirado en los principios de agregación desarrollados.

La articulación entre teoría y aplicación fue, en este sentido, una fortaleza distintiva. No se trató únicamente de producir resultados abstractos, sino también de evaluar su comportamiento mediante simulaciones, analizar su factibilidad computacional e identificar problemas concretos donde pueden resultar útiles. Este equilibrio es especialmente importante en proyectos de estadística avanzada, donde el impacto científico depende tanto de la originalidad conceptual como de la capacidad de convertir esa originalidad en herramientas utilizables.

Otro resultado importante del proyecto fue la consolidación de redes de colaboración académica, tanto en el plano nacional como internacional. El trabajo desarrollado favoreció el intercambio sostenido entre investigadores, fortaleció vínculos existentes y generó nuevas interacciones que permiten proyectar continuidad más allá del período financiado. Este aspecto es especialmente valioso para un sistema científico como el uruguayo, donde la construcción de masa crítica y redes estables de investigación resulta fundamental para sostener líneas de trabajo de frontera. En ese sentido, el proyecto no solo produjo resultados específicos, sino que también contribuyó a fortalecer capacidades instaladas.

En materia de difusión, los resultados del proyecto tuvieron una circulación amplia y diversificada. Por un lado, fueron presentados en congresos y encuentros académicos nacionales e internacionales, lo que permitió someterlos a discusión especializada y obtener retroalimentación de investigadores de otras instituciones. Desde el proyecto se realizo el Congreso denominado "New Trends in Mathematical Statistics II" que reunio investigadores nacionales e internacionales. Por otro lado, se impulsaron actividades dirigidas a públicos no estrictamente académicos. Entre ellas se destacan jornadas de estadística aplicada realizadas en el interior del país, espacios de intercambio entre academia y sector productivo, y actividades de divulgación orientadas a estudiantes de enseñanza media. En particular, la jornada "Un puente entre la academia y la empresa", organizada desde el Proyecto, constituyó una instancia relevante para mostrar la pertinencia de los desarrollos metodológicos en problemas concretos de empresas y organizaciones, y para visibilizar el valor estratégico de la estadística en la toma de decisiones. Del mismo modo, las charlas en liceos y otras acciones de acercamiento al público joven contribuyeron a fortalecer la cultura científica y a promover vocaciones tempranas en áreas cuantitativas.

Conclusiones y recomendaciones

El balance general del proyecto es claramente positivo. Los resultados obtenidos muestran que se cumplieron de manera satisfactoria los objetivos propuestos y, en varios aspectos, se lograron avances que incluso exceden las metas iniciales. A lo largo de su ejecución, el proyecto permitió desarrollar nuevas herramientas metodológicas, profundizar en problemas teóricos de relevancia actual y consolidar líneas de trabajo que combinan estadística robusta, geometría, inferencia y análisis de datos complejos. En ese sentido, una de sus principales fortalezas fue la capacidad de articular preguntas matemáticas profundas con problemas estadísticos contemporáneos, generando resultados originales, rigurosos y potencialmente transferibles a distintos contextos de aplicación.

Uno de los aportes más relevantes del proyecto fue la introducción de nuevos métodos robustos para el análisis de datos en presencia de contaminación, valores atípicos o estructuras alejadas de los supuestos clásicos. En particular, el desarrollo de GROS constituyó un avance significativo dentro de la línea de agregación robusta. Este método permitió mostrar que es posible combinar estimadores construidos sobre submuestras y obtener, como resultado, un procedimiento estable, resistente a outliers y con garantías teóricas sólidas. El hecho de que GROS mantenga propiedades sub-gaussianas y un breakdown point elevado lo vuelve especialmente atractivo para problemas donde la contaminación de datos no es excepcional, sino parte habitual de la práctica. Más aún, el método presenta la virtud de ser muy general, en el sentido de que puede aplicarse a distintos estimadores base y en contextos métricos diversos. Esto amplía notablemente su alcance y sugiere que no se trata de una solución puntual para un problema específico, sino de una estrategia metodológica que el equipo de invetigación sigue explorando y desarrollando en otros paradigmas.

En paralelo, el proyecto logró extender marcos teóricos ya consolidados hacia escenarios más generales y exigentes. Un caso particularmente destacado fue la inferencia conforme en variedades. La posibilidad de construir regiones predictivas confiables cuando la variable respuesta pertenece a una variedad riemanniana constituye un aporte de gran importancia, tanto desde el punto de vista conceptual como aplicado. Hoy en día, muchos datos de interés científico no viven naturalmente en espacios euclidianos: esto ocurre, por ejemplo, con datos direccionales, formas, objetos geométricos, trayectorias o configuraciones que requieren una descripción intrínsecamente no lineal. En ese contexto, extender la inferencia conforme a espacios curvos representa un paso importante hacia una estadística capaz de responder a la complejidad real de los datos modernos.

Otro resultado importante fue el avance en torno a propiedades geométricas fundamentales de los conjuntos y su estimación a partir de datos. En particular, el estudio de la standardness permitió clarificar tanto las posibilidades como las limitaciones de la inferencia estadística en este ámbito. El proyecto no se limitó a proponer un estimador consistente del parámetro asociado a dicha propiedad, sino que además avanzó en la comprensión teórica de su naturaleza. La demostración de que la standardness no puede testearse con datos finitos constituye un resultado de especial valor conceptual. Este tipo de hallazgo tiene un papel muy importante en la investigación matemática y estadística, porque no solo aporta nuevas herramientas, sino que también delimita con precisión qué es posible y qué no es posible inferir a partir de la información observada.

En la misma dirección, los resultados obtenidos en torno a medidas de no convexidad constituyen otro ejemplo del aporte del proyecto a la estadística geométrica. La introducción de nuevos índices, junto con el estudio sistemático de otros ya existentes, permitió avanzar hacia una cuantificación más precisa de cuánto se aparta un conjunto de la convexidad. Esto no solo tiene interés teórico, sino también aplicaciones potenciales en áreas como visión computacional, control de calidad, análisis de formas y aprendizaje automático. El hecho de haber obtenido estimadores consistentes y distribuciones límite para estos índices muestra que el proyecto no se quedó en el plano descriptivo o exploratorio, sino que construyó una verdadera base

inferencial para su uso. Además, la posibilidad de interpretar estas cantidades como rasgos geométricos útiles para detectar irregularidades o anomalías sugiere líneas de continuidad muy prometedoras.

Asimismo, la noción de alcance polinomial desarrollada en el proyecto representa una contribución original con potencial metodológico considerable. La definición de este concepto y la propuesta de un método para estimarlo a partir de datos internos, sin necesidad de parámetros de suavizado, muestran nuevamente la combinación de profundidad teórica y operatividad práctica que caracteriza al proyecto. En estadística geométrica, la dependencia de parámetros de suavizado suele ser una dificultad importante, tanto desde el punto de vista metodológico como computacional. Haber logrado un procedimiento que evita esa dependencia constituye una ventaja clara. Además, la conexión con resultados clásicos de la geometría, como el teorema de Steiner, da cuenta de la solidez conceptual del enfoque. En suma, esta línea de trabajo amplía las herramientas disponibles para la estimación de propiedades geométricas de conjuntos y abre una vía interesante para futuras aplicaciones en contextos tridimensionales y de análisis de imágenes.

Un aspecto que merece destacarse especialmente es que cada uno de estos desarrollos fue acompañado por una validación rigurosa, tanto teórica como empírica. No se trató de propuestas meramente intuitivas o exploratorias, sino de métodos cuidadosamente fundamentados. Se probaron propiedades de consistencia, convergencia, cobertura y comportamiento asintótico, según correspondiera en cada caso. A ello se sumaron estudios de simulación que permitieron evaluar el desempeño efectivo de las metodologías bajo distintos escenarios, incluyendo situaciones con contaminación, heterogeneidad, estructuras geométricas complejas o distribuciones con colas pesadas.

Más allá de los resultados científicos específicos, el proyecto tuvo también un efecto importante en términos de fortalecimiento de capacidades de investigación. La ejecución del trabajo permitió consolidar un espacio de producción académica en torno a problemas de estadística moderna, con fuerte contenido matemático y alto potencial de desarrollo futuro. Se fortalecieron vínculos entre investigadores, se promovió el intercambio académico y se generaron condiciones para la continuidad de varias de las líneas abiertas. En ese sentido, el proyecto no solo produjo artículos, métodos y resultados teóricos, sino que también contribuyó a ampliar las capacidades analíticas instaladas a nivel nacional y a posicionar mejor estas temáticas dentro del sistema científico.

En relación con las perspectivas futuras, el proyecto deja planteadas múltiples direcciones de continuidad. Una de las más claras es la implementación sistemática de los métodos desarrollados en software estadístico, particularmente en entornos como R o Python. Este paso resulta natural y necesario para facilitar su uso por parte de otros investigadores y eventualmente por equipos técnicos de distintos sectores. La disponibilidad de implementaciones accesibles no solo amplía el impacto de los resultados, sino que también permite contrastarlos en nuevos problemas y enriquecerlos a través de la práctica. En algunos casos, la implementación también exigirá optimizaciones computacionales, especialmente si se busca aplicar estos métodos a muestras grandes, datos de alta dimensión o problemas de complejidad geométrica elevada. Por ello, una agenda futura razonable combina el trabajo teórico con desarrollos algorítmicos y computacionales.

También parece especialmente valioso continuar profundizando en el diálogo entre teoría geométrica e inferencia estadística. La experiencia acumulada en el proyecto muestra que muchas preguntas contemporáneas del análisis de datos requieren precisamente esta combinación: una comprensión fina de la estructura geométrica del objeto y, al mismo tiempo, herramientas estadísticas con garantías rigurosas. En ese sentido, el proyecto deja instalada una orientación conceptual muy potente, que puede nutrir trabajos futuros en áreas emergentes como análisis de datos en variedades, estadística sobre espacios métricos, topología aplicada o inferencia robusta en contextos no euclidianos.

En síntesis, puede afirmarse que el proyecto generó conocimiento sólido, original y bien documentado, con aportes concretos al estado del arte en varias áreas de la estadística contemporánea. Sus resultados avanzan en problemas de robustez, inferencia predictiva, geometría de conjuntos y cuantificación de propiedades estructurales, siempre con un equilibrio destacable entre profundidad teórica y pertinencia metodológica. La validación rigurosa de las propuestas, la coherencia de las extensiones respecto a la teoría clásica y la apertura de nuevas líneas de trabajo refuerzan la calidad global del proyecto.

Por todo lo anterior, las conclusiones del proyecto son francamente alentadoras. Se alcanzaron los objetivos propuestos, se produjeron desarrollos innovadores, se clarificaron límites fundamentales de la inferencia en ciertos problemas geométricos y se fortalecieron capacidades de investigación con proyección futura. A la vez, quedaron planteadas líneas de continuidad concretas, tanto en términos de implementación y optimización como de expansión hacia nuevos dominios de aplicación.

Productos derivados del proyecto

Tipo de producto	Título	Autores	Identificadores	URI en repositorio de Silo	Estado
Artículo científico	On the notion of polynomial reach: A statistical application	Alejandro Cholaquidis, Antonio Cuevas, Leonardo Moreno	10.1214/24-EJS2278	https://silo.uy/vufind/Record/COLIBRI_195b0ebd5895673f325ac4ec23f26164/Details	Finalizado
Artículo científico	Statistical analysis of measures of non-convexity	Alejandro Cholaquidis, Betriz Pateiro-López, Ricardo Fraiman, Leonardo Moreno	https://doi.org/10.1007/s11749-023-00889-4	https://silo.uy/vufind/Record/COLIBRI_0f94631f4802463d1668fa3c98d8c28b/Description	Finalizado
Artículo científico	On standardness: estimation of the standardness constant and decidability aspects	Alejandro Cholaquidis, Beatriz Pateiro-López, Leonardo Moreno	https://doi.org/10.1080/10485252.2025.2591341	https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/48543	Finalizado
Artículo científico	Conformal inference for regression on Riemannian manifolds	Alejandro Cholaquidis, Fabrice Gamboa, Leonardo Moreno	10.1214/25-EJS2478	https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/48533	Finalizado
Artículo científico	GROS: A General Robust Aggregation Strategy	Alejandro Cholaquidis, Emilien Joly, Leonardo Moreno	https://www3.stat.sinica.edu.tw/ss_newpaper/SS-2024-0414_na.pdf	https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/48533	Finalizado

Referencias bibliográficas

Cholaquidis, A., Joly, E. y Moreno, L. GROS: A General Robust Aggregation Strategy. arXiv preprint, 2024.
Cholaquidis, A., Gamboa, F. y Moreno, L. Conformal inference for regression on Riemannian manifolds. Electronic Journal of Statistics, 19 (2025), 6141–6166.
Cholaquidis, A., Moreno, L. y Pateiro-López, B. On standardness and the non-estimability of certain functionals of a set. arXiv preprint, 2023.
Cholaquidis, A., Fraiman, R., Moreno, L. y Pateiro-López, B. Statistical analysis of measures of non-convexity. arXiv preprint, 2022.
Cholaquidis, A., Cuevas, A. y Moreno, L. On the notion of polynomial reach: A statistical application. Electronic Journal of Statistics, 18 (2024), 3437–3460.
Devroye, L. y Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. SIAM Journal on Applied Mathematics, 38(3), 480–488.
Cuevas, A. y Rodríguez-Casal, A. (2004). On boundary estimation. Advances in Applied Probability, 36, 340–354.
Cuevas, A., Fraiman, R. y Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. The Annals of Statistics, 35(3), 1031–1051.
Fraiman, R., Gamboa, F. y Moreno, L. (2019). Connecting pairwise geodesic spheres by depth: DCOPS. Journal of Multivariate Analysis, 169, 81–94.
Fraiman, R., Gamboa, F. y Moreno, L. (2020). Sensitivity indices for output on a Riemannian manifold. International Journal for Uncertainty Quantification, 10(4).
Gamboa, F., Klein, T., Lagnoux, A. y Moreno, L. (2021). Sensitivity analysis in general metric spaces. Reliability Engineering & System Safety, 212, 107611.
Cholaquidis, A., Fraiman, R. y Moreno, L. (2022). Level set and density estimation on manifolds. Journal of Multivariate Analysis, 189, 104925.
Lei, J. y Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. Referencia metodológica central para la extensión conforme utilizada en variedades.
Vovk, V., Gammerman, A. y Shafer, G. (1998). Trabajo fundacional de inferencia conforme, citado como punto de partida de la línea desarrollada en variedades.
Federer, H. (1959). Trabajo clásico sobre reach positivo y fórmulas de tipo Steiner, base conceptual del alcance polinomial.
Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference. Referencia general importante para la componente funcional y de alta dimensión del proyecto.
Berrendero, J. R., Cuevas, A. y Torrecilla, J. L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. Journal of the American Statistical Association, 113(523), 1210–1218.
Berrendero, J. R., Bueno-Larraz, B. y Cuevas, A. (2020). On Mahalanobis Distance in Functional Settings. Journal of Machine Learning Research, 21(9), 1–33.