REAL ACADEMIA GALEGA DE CIENCIAS

La Estadística Matemática en la era del “Big Data”.

Wenceslao González Manteiga

Discurso de Ingreso en la

Real Academia Galega de Ciencias

Santiago de Compostela, 19 de septiembre de 2018.

Índice

Prólogo y agradecimientos. 3

Los inicios de un estadístico matemático. 6

Los años previos a la tesis doctoral. 6

Los años posteriores a la tesis doctoral. 8

Un estadístico matemático en la década de los 80. 11

Un breve resumen de la evolución de la Estadística hasta la década de los 80. Algunas pinceladas históricas. 11

El inicio del maridaje entre la metodología y la computación. 13

Desde los años 90 hasta la actualidad. 19

La metodología de la Estadística Matemática en la era del "Big Data". 36

La Estadística Matemática gallega en la era del "Big Data". 44

Una breve conclusión. 48

Referencias bibliográficas. 49

Excmo. Sr.Presidente da Real Academia Galega de Ciencias, Reitor da Universidade de Santiago de Compostela, Reitor da Universidade de A Coruña, Comandante-Director de la Escuela Naval Militar de Marín, Concelleiro de Educación e Cidadanía do Concello de Santiago, Autoridades, Compañeiros e compañeiras, amigos e amigas, familiares:

Prólogo y agradecimientos.

Mi intervención se inicia con palabras de agradecimiento al presidente y a todos los miembros de la Real Academia Gallega de Ciencias (RAGC) por considerarme merecedor de formar parte como Académico en una institución tan prestigiosa de Galicia.

El nombramiento por parte de la Academia Gallega de Ciencias es para mí una gran satisfacción personal, ya que significa el reconocimiento a mi labor científica aquí, en mi tierra, tras otros éxitos recabados a nivel nacional e internacional.

Tengo que reconocer que el "azar", concepto fundamental en la evolución histórica y metodológica de una disciplina como la Estadística Matemática, juega también un papel central a la hora de motivar este discurso. El azar ha sido muy favorable en varias etapas de mi vida: utilizando terminología de la lógica matemática, hizo que una condición necesaria para el logro de esta distinción como académico, como es el trabajo duro y continuado durante años, también se transformase en suficiente, al acompañarme con fortuna en dos momentos cruciales de mi vida.

- En primer lugar, el de mi nacimiento en Sao Paulo, Brasil. Como hijo de emigrantes viví de primera mano una niñez marcada por tres viajes en barco surcando el Atlántico en compañía de mis padres. Ya en la adolescencia, en el tercero de estos viajes de medio mes de duración, tuve la grata sensación del retorno y de la acogida definitiva en mi querida Galicia.
- El segundo momento fue años más tarde, siendo estudiante de la licenciatura de Matemáticas, cuando conocí a Pili. Ella es hoy mi esposa y la madre de mis cuatro hijos: Helena, David, Javier y Lidia, así como abuela "jubilosa" de mis tres nietos: Julia, Leo y Alex.

Estos dos "azarosos" eventos tuvieron un papel fundamental en mi posterior ruta profesional. Sin los valores de la gente humilde que me transmitieron mis padres, y que siempre tomé como base esencial en las actuaciones de mi vida, y sin el apoyo inestimable de mi familia, no hubiesen sido posibles los logros académicos. En ese sentido debo reiterar el papel crucial y especial de mi esposa, que siempre estuvo a mi

lado, animándome en los momentos más complicados y ayudándome como soporte fundamental de la difícil logística que conlleva la educación de cuatro hijos. Especialmente durante los largos periodos de ausencia que la actividad docente e investigadora requiere.

Podríamos decir que el azar es también responsable de que, como matemático, presente hoy un discurso ligado a la Estadística y no a otras disciplinas de las Matemáticas; tal es el caso del "Análisis Matemático", que siempre me resultó muy atractivo y que para mí representa un soporte metodológico fundamental para el desarrollo de las Matemáticas y en especial de la propia Estadística Matemática.

Durante el primer ciclo de mis estudios de licenciatura en Matemáticas, mi opción preferente fueron las asignaturas de Cálculo Diferencial y Ecuaciones Diferenciales. El inicio de la impartición de la especialidad de Estadística en la licenciatura de Matemáticas en el curso 1976-1977, cuando yo todavía era estudiante, cambió definitivamente mi elección en el segundo ciclo de licenciatura. La Estadística Matemática se fundamenta en el Cálculo de Probabilidades y la Teoría de la Medida, requiriéndose por tanto un buen conocimiento del Análisis Matemático. Por otro lado, también es un puente natural hacia la modelización de fenómenos de la vida real. Esta opción de estudios resultó ser muy atractiva para una persona con mi perfil, que ya en sus inicios, antes de cursar la licenciatura, dudaba sobre la posibilidad de elegir los estudios de Física como ciencia motivadora de modelos matemáticos.

Tengo una inmensa satisfacción de ser el primer "Estadístico Matemático" de Galicia que se incorpora a este selecto grupo de cinco académicos numerarios dentro del campo de Matemáticas, Física y Física del Cosmos de la RAGC, que conforman los matemáticos Luis A. Cordero Rego (1979), Gerardo Rodríguez López (1979), Juan José Nieto Roig (2016) y los físicos Félix Vidal Costa (1992), Vicente Pérez Villar (1998) y José Rivas Rey (2017).

Precisamente cursando la asignatura "Análisis Funcional", que era obligatoria de quinto curso de la especialidad de Estadística e Investigación Operativa, conocí a uno de los mencionados Académicos, el Profesor Juan José Nieto Roig, compañero actual en el departamento de Estadística, Análisis Matemático y Optimización, y con quien además compartí toda una vida académica en la Facultad de Matemáticas. Hoy le agradezco profundamente su favorable disposición a contestar este discurso de ingreso en la RAGC.

Como ya se deja entrever en el título de esta conferencia, mi objetivo no es otro que transmitir la idea de lo que puede representar para un estadístico matemático la era actual del "Big Data". Para ello, tomando como base la experiencia científica acumulada en las últimas décadas, analizaré lo que estamos viviendo actualmente, con cierta proyección de futuro. Se trata por tanto de una apreciación personal, la de

alguien que llegó a esta disciplina a finales de los 70. Describiré mi propio viaje científico de las últimas décadas: cómo era la Estadística Matemática cuando yo la conocí, los protagonistas que fueron surgiendo en mi camino, los avances más notables de los que fui testigo y, cómo no, la experiencia maravillosa de haber dirigido 32 tesis doctorales, contribuyendo así en el surgimiento de una potente escuela de estadísticos en Galicia. Finalmente, concluiré con un análisis de la situación actual de la Estadística Matemática tras la irrupción del gran boom asociado al “Big Data”.

Los inicios de un estadístico matemático.

Los años previos a la tesis doctoral.

A mediados de septiembre del año 1979, tras terminar los estudios de la licenciatura de Matemáticas, con el correspondiente examen de grado, me disponía a aceptar una plaza de profesor contratado para impartir disciplinas del área de Análisis Matemático en el Campus de Lugo de la Universidad de Santiago de Compostela. En ese momento el azar reaparece en mi camino con el traslado a la Universidad de Granada del profesor Ramiro Melendreras, catedrático de Estadística Matemática y Cálculo de Probabilidades en la Universidad de Santiago de Compostela. Como consecuencia de este hecho, y respondiendo a una oferta de incorporación inmediata al departamento de Estadística Matemática, en un viernes de finales de septiembre de 1979, me veo obligado a tomar una decisión rápida que marcaría el destino académico siguiente de toda mi vida. Así, en ese mismo fin de semana, cambio de planes y cambio de destino; el desplazamiento original Ferrol-Lugo pasa a ser un Ferrol-Santiago de Compostela.

De los primeros años en el departamento, previos a la defensa de mi tesis doctoral, tengo un grato recuerdo por la ilusión científica con la que los abordé y por lo que representaron para mi posterior desarrollo profesional y humano.

El departamento de Estadística Matemática había permanecido integrado en el departamento de Algebra y Fundamentos, desde la fugaz estancia como Catedrático de la disciplina, del profesor Francisco Azorín. El profesor Azorín ejerció desde septiembre de 1961 hasta diciembre de 1962, cuando obtuvo una excedencia especial de 10 años por su tarea como asesor en cuestiones de muestreo estadístico al servicio de las Naciones Unidas. La Estadística queda entonces relativamente oculta en la Universidad de Santiago de Compostela hasta el año 1976. En ese año se incorpora como catedrático el profesor Ramiro Melendreras y crea la Especialidad de Estadística e Investigación Operativa. Como estudiante, paso entonces a formar parte de la segunda generación de los que cursan esa especialidad de la titulación de Matemáticas.

Cuando me integro en el departamento, a principios de los 80, éste estaba entonces formado por un conjunto de personas jóvenes y muy voluntariosas, que en su mayoría no habían tenido la oportunidad de cursar una especialidad de Estadística en sus respectivas licenciaturas como matemáticos. Casi todos se habían incorporado como profesores contratados por la Universidad para para poner en marcha el funcionamiento de la especialidad. Ello suponía impartir diversas disciplinas muy especializadas en condiciones realmente complejas y adversas.

En el curso 1979-1980 leo mi Tesina de Licenciatura en el campo del “Análisis Discriminante de datos discretos”, bajo la dirección del entonces recientemente incorporado como Profesor Agregado de Universidad, José Antonio Cristóbal. Una vez más el azar juega a mi favor y con este trabajo llego a la temática de “Estimación de Curvas”, que representaría posteriormente el núcleo de mi tesis doctoral. De hecho, todo es fruto de un hallazgo casual en la biblioteca de los departamentos de aquel momento, Teoría de Funciones y Ecuaciones Funcionales. Tengo la suerte de encontrar una revisión muy precisa sobre estimación de curvas de densidad. Este trabajo, que me resultó enormemente motivador, había sido publicado en el primer volumen de la revista Journal Nonlinear Analysis, Theory, Methods and Applications (1). Posteriormente, también sería de gran ayuda en el estudio de esta temática, la recepción de varios sobres cargados de trabajos publicados en aquellos años por los investigadores de referencia internacional, Luc Devroye y Gerard Collomb (este último, fallecido muy pocos años después), pertenecientes respectivamente a las Universidades McGill en Montreal (Canadá) y Paul Sabatier en Toulouse (Francia). Especialmente son de gran valor los ejemplares de sus tesis doctorales que guardo como un tesoro científico. Años más tarde, el profesor Luc Devroye y yo nos intercambiamos varias visitas científicas (Figura 1), participando en seminarios de investigación y en tribunales de tesis doctorales.

Figura 1: Con Luc Devroye en Montreal, Canadá, en el año 1998.

A principios de los 80 en el departamento no había tradición científica de contactos internacionales, los miembros del mismo investigaban de forma muy autónoma, en temas muy dispersos, y lo compatibilizaban con una gran carga de trabajo docente. El

profesor Cristóbal, como único profesor numerario en aquel momento, desarrollaba una labor francamente ardua: dirigía a la vez a media docena de investigadores en sus tesis doctorales que cubrían varios aspectos de Estadística Matemática y de la Investigación Operativa, como la "Teoría de Juegos", los datos categóricos, los modelos de fiabilidad, los modelos de decisión, los modelos de elección de cartera y las técnicas de clasificación. Eran años aquellos en los que el esfuerzo en la formación, que tenía que ser totalmente autónoma, dependía en gran medida del ánimo individual de cada persona y de la energía necesaria para hacer largas jornadas de dedicación a la investigación; jornadas por otro lado que incluían, más allá del horario de trabajo habitual, gran parte de las noches y de los fines de semana.

En noviembre de 1982, y tras seguir religiosamente estas estrictas pautas de trabajo, defiendo mi tesis doctoral dedicada a formalizar una clase general de estimadores no paramétricos de la función de densidad y de la función de regresión (2).

Llegado a este punto, me gustaría agradecer sinceramente al profesor Cristóbal, por haber revisado mi material de trabajo, a pesar de su escaso tiempo, y por haberme permitido defenderlo como tesis doctoral. Además, también debo reconocer de nuevo la gran suerte de haber podido acceder a los múltiples y muy motivadores artículos de investigación de los mencionados profesores, y científicos "gran class", Luc Devroye y Gerard Collomb. Sus envíos, por correo ordinario, de sobres llenos de material muy importante para mis avances, resultaron ser determinantes para mi finalización de trabajo de tesis doctoral.

Los años posteriores a la tesis doctoral.

Una vez finalizada mi etapa de doctorado, durante los cuatro años posteriores se suceden acontecimientos que me resultan, una vez más, muy favorables.

- **Una promoción académica muy rápida**. En ese corto período de tiempo promociono secuencialmente de Profesor Ayudante a Profesor Colaborador, luego a profesor Adjunto Contratado y finalmente a Titular de Universidad en julio de 1986.

- **Una integración transversal con otros equipos de investigación en la Universidad de Santiago de Compostela.** Como consecuencia de la gran demanda de apoyo estadístico, desarrollo en esos años una intensa actividad de consultoría a diversos investigadores, en diferentes campos de la Biología, la Química, la Física o la Medicina, entre otros. Empiezo a ser consciente de la potente herramienta que representan los modelos

de regresión y el análisis multivariante en la Estadística Matemática, no sólo por su gran componente metodológica de las Matemáticas, sino también como soporte científico al análisis de datos y a las correspondientes interpretaciones en los estudios experimentales.

- **La consecución del primer premio de investigación Ramiro Melendreras.** Hecho que tiene lugar en el año 1984 en el congreso nacional de Estadística e Investigación Operativa (SEIO) celebrado en la Universidad de Granada, en donde se presentan los resultados más importantes de mi tesis doctoral. Este premio, cuya existencia se inicia en ese mismo congreso, un año después del fallecimiento del profesor Melendreras, se otorga en cada edición de los congresos nacionales de la SEIO, desde entonces y hasta la actualidad.

- **El inicio de la internacionalización** Una publicación derivada de los resultados de mi tesis doctoral en la prestigiosa revista "Annals of Statistics" (4), en colaboración con mi director de tesis José Antonio Cristóbal y con mi compañero de departamento, el profesor Pedro Faraldo Roca, despierta un gran interés en el foro internacional y me sirve para iniciar colaboraciones con diversos investigadores de referencia. Esto se traduce en las primeras visitas a nuestro departamento, que por entonces ya se llamaba de Estadística e Investigación Operativa, de grandes personalidades en la materia.

  El primer contacto destacado en este período de tiempo es el del profesor Winfried Stute, al que considero realmente mi gran maestro en el terreno académico, por todo lo que he aprendido de él en herramienta probabilística a lo largo de los años. En particular con él me formé notablemente en el estudio de los procesos empíricos y sus aplicaciones en la obtención de resultados asintóticos de la Estadística. El profesor Stute empieza a visitarnos desde el año 1986, y sigue haciéndolo todos los años siguientes hasta la actualidad. La universidad de Santiago de Compostela le otorga el grado de "Doctor Honoris Causa" en el año 2008, (Figura 2). A lo largo de este extenso período de tiempo el profesor Stute imparte numerosos cursos, desde los primeros años en procesos empíricos hasta más recientemente en el cálculo estocástico en Ingeniería Financiera.

Figura 2: Foto de la izquierda, el profesor Stute en Alemania en mi primera visita a su Universidad en el año 1987. Foto de la derecha, con motivo de su "Honoris Causa" por la Universidad de Santiago en 2008.

Un estadístico matemático en la década de los 80.

En lo que sigue, describiré cuáles eran las coordenadas científicas de la situación de la Estadística Matemática a principio de los 80, en el período que abarca los años anteriores y posteriores a mi tesis doctoral. Para ello voy a remontarme previamente algunos siglos atrás con el fin de hacer una breve revisión de la evolución histórica de la Estadística hasta ese momento.

Un breve resumen de la evolución de la Estadística hasta la década de los 80. Algunas pinceladas históricas.

Los fundamentos de la historia de la Estadística Matemática se cimientan en la Teoría de la Probabilidad. Podemos considerar como creadores de los primeros desarrollos a Cardano (1501-1576), Galileo (1564-1642), las cartas entre Fermat (1601-1665) y Pascal (1623-1662), Bernoulli (1654-1705), Moivre (1667-1754), Laplace (1749-1827), Gauss (1777-1855) y Poisson (1781-1840). Todos ellos personificados a día de hoy en las funciones de distribución notables, tanto discretas como continuas, que llevan su nombre y que constituyen la base de muchos de los modelos probabilísticos: la campana de Gauss, la distribución de Bernoulli, la distribución de Poisson,…etc. Estos modelos de distribución sirven en la actualidad como soporte muy útil para el estudio de diversos eventos aleatorios, como por ejemplo el número de llamadas a una central telefónica, los eventos de carácter binario o la medición de aspectos morfológicos en la anatomía humana.

Desde la segunda mitad del siglo XIX y, sobre todo en la primera parte del siglo XX, tiene lugar la creación de diferentes escuelas y tendencias dedicadas al estudio de la Matemáticas en el campo de la Teoría de la Probabilidad. En este período son muy importantes los llamados “creadores de la escuela rusa” con Chebyshev (1821-1894), Markov (1856-1922), Lyapunov (1857-1918) y especialmente Kolmogorov (1903-1987), que desarrollan y perfeccionan las llamadas leyes débiles, leyes fuertes y los teoremas centrales del límite, herramientas cruciales para la explicación de cómo las frecuencias empíricas relativas convergen estocásticamente hacia la probabilidad de los eventos de interés.

La Estadística Matemática, como tal, se apoya en dichos fundamentos y su desarrollo es más reciente. Aunque Graunt (1620-1674), Thomas Bayes (1701-1761), Laplace (1749-1827) y Quetelet (1796-1874) son considerados como los precursores de la misma, sus estudios se limitaban a tablas de mortalidad, a recopilación de datos relativos de la Mecánica Celeste y, en general, al análisis descriptivo de datos de carácter observacional.

Los verdaderos artífices del punto de partida de la Estadística Matemática aparecen en el campo de la aplicación a la Biología y a las Ciencias Sociales. A destacar, Galton (1822-1911), Karl Pearson (1857-1936) y Fisher (1890-1962), de la llamada "Escuela Inglesa"; y Jeffreys (1891-1989), De Finetti (1906-1985) y L. Savage (1917-1971) , de la escuela de la Estadística Bayesiana. Con ellos surgen las distribuciones asociadas a las transformaciones de datos gaussianos, los primeros conceptos desarrollados formalmente de estimación puntual, la estimación por intervalos, los contrastes de hipótesis, los contrastes de bondad de ajuste, los modelos de regresión y del análisis de la varianza, que se explican actualmente en los cursos de Inferencia Estadística en las diversas titulaciones de grado. En este período se crea el "International Statistical Institute" (ISI) (1885), la revista Biometrika (1901) fundada por Karl Pearson y el "Journal of the Royal Statistical Society" (1887). El primero es un ejemplo de un organismo científico muy importante, gestor de grandes congresos mundiales cada dos años. Las otras dos representan revistas de máximo prestigio desde entonces y hasta nuestros días.

Los investigadores herederos de esta generación fueron en cierta medida los creadores de lo que hoy es el Análisis Multivariante, base fundamental para el posterior desarrollo de la metodología estadística en "Big Data" -la llamada Estadística de los datos multidimensionales-. A destacar en este ámbito, Harold Hotelling (1895-1973), gran estudioso de la técnica de las componentes principales y del análisis de correlación canónica y además uno de los artífices principales de los llamados métodos de reducción de la dimensión; P.C. Mahalanobis (1893-1972), estadístico hindú, que, motivado por lo problemas económicos y culturales en la India, es el fundador del "Indian Statistical Institute" y diseñador de los planes de desarrollo quinquenales posteriores a la independencia del país; John Wishart (1898-1956), ayudante de Pearson y de Fisher, que generaliza los estudios del análisis de la varianza (ANOVA), con la llamada "distribución matricial de Wishart" para los modelos del análisis de la matriz de covarianzas (MANOVA); Samuel Wilks (1906-1964) que bajo la influencia de Pearson y Wishart, crea la distribución de Wilks y es uno de los fundadores del "Institute of Mathematical Statistics" y editor de la revista "Annals of Mathematical Statistics", hoy subdividida en dos: "Annals of Statistics" y "Annals of Applied Statistics", ambas entre las mejores revistas de referencia de la Estadística Matemática y de la Estadística Aplicada actual respectivamente. Finalmente, también añadimos a

este grupo a Charles Spearman (1863-1945), psicólogo inglés, creador del primer modelo de Análisis Factorial.

Con un análisis detallado de toda la metodología desarrollada por estas dos generaciones de estadísticos, se entiende el perfil de lo que representa la Estadística, como herramienta matemática encaminada a explicar lo que dicen los datos y la modelización de los mismos. Es decir, utilizando terminología moderna, la Estadística es “la Ciencia Matemática de los datos”.

De forma complementaria, para entender correctamente el estado de la Estadística Matemática de la década de los 80, debemos mencionar también la influencia de John Tukey (1915-2000), por ser el creador de la transformada de Fourier rápida -crucial para la agilización computacional de los algoritmos-, de los nuevos métodos para el diagnóstico de modelos y estimación robusta y sobre todo, por ser el inventor de los términos conceptuales ligados al análisis moderno exploratorio de datos: el “boxplot“ (conocido como “diagrama de cajas” o “diagrama de bigotes”) es un ejemplo notable.

En lugar destacado debemos citar también a otro de los gigantes, el científico Calyampudi R. Rao (1920), discípulo de Mahalanobis y generador de muchos conceptos de la Estadística Matemática, como la famosa cota de error de Cramer-Rao. Este profesor fue premiado en Estados Unidos con la medalla nacional de la Ciencia. La “American Statistical Association” lo describe como una leyenda viva de la Estadística y el “The Times of India” lo identifica como uno de los mejores 10 científicos en la India de todos los tiempos.

Los aspectos detallados de la historia de la Estadística y sus preliminares podrían cubrir la totalidad de este discurso y mucho más. No quiero extenderme en el tema y recomiendo, para más detalles, la lectura del reciente libro sobre aspectos históricos de la Estadística y de la Probabilidad (4) del profesor y colega, Miguel Ángel Gómez Villegas, de la Universidad Complutense de Madrid y también presidente de la Sociedad de Historia de la Estadística y la Probabilidad de España desde agosto de 2009.

El inicio del maridaje entre la metodología y la computación.

Rápidamente, en mi etapa formativa, pude comprobar que la Estadística en aquellos momentos ya vivía una transición, con un cambio de paradigma hacia los períodos más recientes ligados a la computación. La creciente capacidad computacional de los

centros de cálculo y la aparición de los primeros ordenadores de mesa, que luego evolucionarían hacia los versátiles y ligeros portátiles actuales, hicieron que nuevos campos metodológicos de la Estadística Matemática se inspirasen en la computación y empezasen un desarrollo vertiginoso en esa década, que continuaría, por supuesto, en los años posteriores.

A destacar:

A) **Las técnicas de remuestreo**. De entre todas ellas destaco el "Bootstrap" que se empieza a popularizar en el año 1979 a partir de un artículo invitado en la revista "Annals of Statistics" (5), publicado por el profesor Efron de la Universidad de Stanford. Esto da paso a la estrategia de creación en ordenador de muestras artificiales, que permiten generar una Inferencia Estadística computacional. A partir de estos datos simulados se obtienen estimadores del sesgo y de la varianza asociados a estimadores puntuales de todo tipo, así como intervalos de confianza, mucho más precisos que los obtenidos en las décadas precedentes bajo la suposición de secuencias gaussianas.
Cabe mencionar, como ejemplo de la popularidad que pronto alcanzan estos procedimientos, una primera publicación de carácter divulgativo (6), del año 1983 en la revista Investigación y Ciencia, en donde se ilustra la potencialidad de estas técnicas con datos relativos a las notas de acceso en la prueba de selectividad de los alumnos de Stanford. En particular se generan intervalos de confianza para el coeficiente de correlación lineal entre las notas obtenidas por los alumnos, resultando intervalos más cortos que los usados por la aproximación normal y con mayor acuracidad (Figura 3).

Investigación y Ciencia
Julio 1983 Nº 82
ESTADÍSTICA
"Métodos estadísticos intensivos por ordenador"

Diaconis, Persi and Efron, Bradley

A costa de grandes cálculos, se prescinde de hipótesis convencionales sobre los datos. El método "autodocimante" ha hecho reconsiderar muchas estimaciones sobre la fiabilidad de inferencias

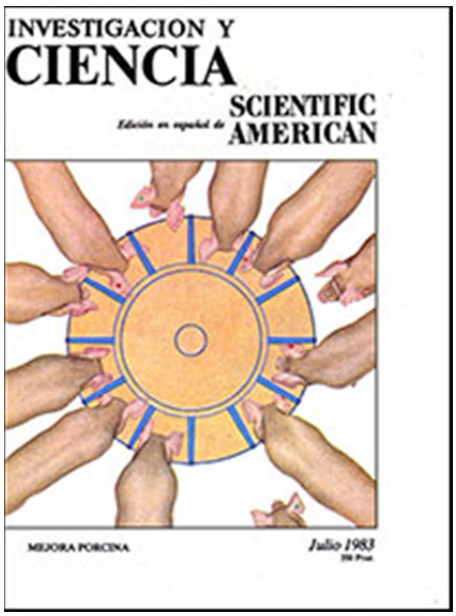


Figura 3: Portada de la publicación divulgativa del bootstrap en la revista Investigación y Ciencia.

científicas.

B) **El análisis exploratorio con técnicas de suavización.** Como consecuencia de la mayor versatilidad gráfica computacional, las estimaciones de las curvas notables usadas en el análisis exploratorio de los datos empiezan a ser representadas de forma relativamente rápida en pantallas de ordenador. Por citar algunos ejemplos, *la función de densidad* (para la exploración de la disposición de los datos de una distribución poblacional)*, la función de regresión* (para el análisis de la dependencia lineal o no lineal entre variables), *la función de razón de fallo* (para el seguimiento del riesgo de fallo en estudios de fiabilidad de sistemas), *la función de supervivencia* (como elemento gráfico para el análisis de duración de tiempos de vida en el ámbito biomédico o de modelos de duración en desempleo en el ámbito económico) , *la función de vida media residual* (de notable importancia en estudios actuariales),…etc. Comienza, por tanto, la explosión del uso gráfico de la Estadística con fines exploratorios, complementando así el carácter más cuantitativo y tradicional de la inferencia de las décadas anteriores del siglo XX.

En este período aparecen los primeros textos notables en esa dirección, como el libro de Wolfgang Hardle (1990) (7) o el de Bernard Silverman (1986) (8), ambos de referencia, aun hoy, en cursos de grado o máster. Curiosamente, el profesor Hardle, visita por primera vez la Universidad de Santiago de Compostela en el año 1988, pasando a ser, junto al profesor Winfried Stute antes mencionado, otra de las figuras científicas más influyentes en mis primeros desarrollos de relevancia en la inferencia no paramétrica en ese momento. Como anécdota, debo mencionar que fue con motivo de su visita cuando comencé a utilizar el correo electrónico con el envío de un e-mail a su dirección. Posteriormente, algunos de mis primeros estudiantes de doctorado visitarían las universidades de Bonn y de Giessen en Alemania, donde trabajaban respectivamente los mencionados profesores (Figura 4).

Figura 4: Con el profesor Hardle en uno de los "Oberwolfach Seminars" en Alemania en 1989.

C) **La generación de curvas ligadas a modelos de supervivencia**. La versatilidad gráfica también es muy determinante para el incremento en el uso de las curvas de supervivencia. Este ámbito es de gran impacto en las aplicaciones a la Biomedicina y a la Bioestadística en general. El estimador introducido por Kaplan y Meier en 1958 (9) y la extensión del mismo, que permite el uso de variables explicativas, e introducida posteriormente en un artículo no publicado, por el profesor Beran de la Universidad de Berkley en 1981, empiezan a ser de uso habitual en todos los estudios de seguimientos en los tiempos de vida, usando las diversas bases de datos del momento. De forma alternativa, y también con gran éxito, destacan las curvas de supervivencia que se derivan de los conocidos como "modelos de regresión de Cox" introducidos por el autor en 1972 (10).

D) **La explosión de los modelos de regresión flexibles generalizados**. Este es el momento en el que se introducen nuevas clases de modelos de regresión que permiten mayor flexibilidad que los restrictivos modelos de regresión lineal múltiple. Un texto que recoge los principales avances de esos años es el libro de Hastie y Tibshirani, con una primera edición en 1990 (11). Estos autores y su libro son los precursores de los diversos modelos que aparecerían años más tarde ligados a las técnicas de aprendizaje estadístico y que se pueden ver en los textos recientes de referencia.

La década de los 80 representó en gran medida el inicio de una era dorada de la Estadística. Buena prueba de ello está en el hecho de que los profesores David Cox y Bradley Efron, protagonistas especiales de aquella década, recibieran en el año 2017 el Premio “Fundación BBVA Fronteras del Conocimiento” en la categoría de Ciencias Básicas, por sus contribuciones en el campo de la Estadística y por convertirla en una herramienta "imprescindible" para el resto de las Ciencias, tal y como se recoge en el diario La Vanguardia, 24/01/2017, por “Revolucionar la Estadística”, (Figura 5).

El profesor David Cox, comentaba en una entrevista posterior a la entrega del premio: “Mi modelo estima la probabilidad de que ocurran desde terremotos a quiebras económicas”.

El profesor Bradley Efron, por su parte, decía que: “El “Bootstrap” no es la estrella de la Estadística, pero si el mejor actor secundario”.

Figura 5: Efron (derecha) y Cox (izquierda), premios “Fundación BBVA Fronteras del Conocimiento”, 2017.

Llegados a este punto y para resumir me gustaría destacar las tres circunstancias que se dieron y que confluyeron a distinto nivel:

a) A nivel internacional, el avance de los citados desarrollos en la metodología estadística.
b) A nivel local, la llegada de una generación de estudiantes brillantes, que acababan sus estudios de Matemáticas con la especialidad de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela, y a los que pude dirigir o codirigir, en años posteriores, sus tesis doctorales, y

c) A nivel personal, todas mis pequeñas aportaciones en el desarrollo científico de la Estadística Matemática, motivadas por tan buenas coincidencias y alimentadas con mucha energía e ilusión.

La combinación de estos tres factores propicia sin duda el nacimiento en nuestra comunidad de un gran colectivo de estadísticos que actualmente desarrollan su labor docente e investigadora en las tres universidades gallegas.

Es por eso que a partir de los años 90 se multiplican nuestras aportaciones con numerosos trabajos de tesis doctorales, cuyos avances más significativos detallo a continuación.

Desde los años 90 hasta la actualidad.

Los primeros avances en inferencia no paramétrica de curvas, para muestra con datos independientes, se completan en el trabajo correspondiente a la dirección de mi primera tesis doctoral, realizada por *Juan Vilar Fernández* (1987) (12), en el contexto de datos dependientes o de series de tiempo. A partir de entonces se empieza a interaccionar el contenido de nuestras ideas iniciales con los grandes campos de potencial crecimiento que la Estadística mostraba en aquellos años. Así, en conexión con los ámbitos ya mencionados anteriormente destacamos:

A) **En el Análisis de Supervivencia**.

La estimación de la curva de Kaplan Meier condicionada a la información suministrada por variables explicativas, es la base de la tesis doctoral de *Carmen Cadarso Suárez* (1990) (13), con grandes aplicaciones biomédicas.

La extensión de dicha estimación a modelos con datos que pueden presentar de forma simultánea el fenómeno de la censura y el truncamiento, se lleva a cabo en las tesis doctorales de *Carmen Iglesias Pérez* (2001) (14) y *César Sánchez Sellero* (2001) (15). Dichas situaciones de datos incompletos son habituales en los seguimientos de tiempos de vida asociados a observaciones biomédicas. Por ejemplo, la modelización con datos truncados tuvo especial relevancia por los seguimientos de los enfermos de SIDA en los años 80. Muchos de estos enfermos fallecieron antes del inicio de estudio del seguimiento de los tiempos de vida, no conociéndose la información de los mismos y representando el primer gran volumen histórico de lo que llamaríamos datos truncados.

En este mismo contexto del análisis de supervivencia destaca el estudio de algunos modelos notables muy relacionados con los modelos de Cox, como son los de censura proporcional estudiados en la tesis de *Jacobo de Uña Álvarez* (1998) (16).

B) **En Técnicas de Remuestreo**.

Fuimos pioneros en el uso de los procedimientos “Bootstrap” para el ámbito de la estimación de curvas, como se pone de manifiesto en la tesis doctoral realizada por *Ricardo Cao Abad* (1990) (17) (Figura 6) una parte de la cual es publicada en la revista “Annals of Statistics”.

Las extensiones metodológicas posteriores, en esta dirección, se presentan en las tesis doctorales de *Pilar García Soidán* (1994) (18), enfocada a la estimación suavizada de la función de distribución y de *María Celia Rodríguez Campos* (1994) (19) para curvas derivadas de la función de densidad y de regresión y con presencia de datos categóricos.

Figura 6: Viaje a Suiza en 1990 con mi compañero José Manuel Prada Sánchez y un jovencísimo doctorando Ricardo Cao Abad para la presentación de resultados de su tesis en un congreso internacional en homenaje a C.R. Rao.

C) **En Modelos de Regresión Flexibles Generalizados**.

Creamos diversos procedimientos que permitieron, por primera vez, la adaptación de la Estadística Robusta al campo de la estimación de curvas de regresión. La robustez en Estadística tuvo un desarrollo extenso en los años 60 y 70, con el objeto de diseñar procedimientos no sensibles a la presencia de datos atípicos, conocidos como "outliers". Este estudio estuvo limitado hasta entonces a la estimación de parámetros reales o vectoriales y no de curvas. Estos avances corresponden al desarrollo de las tesis doctorales de *Manuel Antonio Presedo Quindimil* (1991) (20) y *Carlos Pío del Oro Sáez* (1996) (21).

D) **En el Análisis Exploratorio con Técnicas de Suavización**.

Extendimos su uso a la Estadística Espacial, con una primera tesis defendida en el año (2003) (22) por *Rubén Fernández Casal.*

En ese momento comienza una serie de congresos de relevancia internacional, conocidos como METMA (Modelización espacio-temporal y aplicación al

medioambiente). Tuvimos la ocasión de acoger el quinto de esos congresos en la Universidad de Santiago en el año 2010. Previamente, ya habíamos organizado el "ISI Conference in Environmental Statistics and Health", congreso mundial satélite del ISI.

En todos estos eventos inicié y consolidé nuevos contactos científicos de máxima relevancia. Destaco, entre ellos, al profesor Noel Cressie, una de las mayores autoridades en el campo de la Estadística Espacial, (Figura 7).

Figura 7: Con Noel Cressie en 2001, con motivo de su primera visita a la Universidad de Santiago de Compostela.

La década de los 90 constituyó un apasionante punto de referencia científico de mi vida académica. Fueron varios los eventos azarosos que favorecieron el esplendor de este momento. A destacar la creación en el año 1989 de las Universidades de A Coruña y Vigo. Este hecho ayudó a que mis primeros estudiantes doctores se posicionasen en los departamentos recién creados de Estadística e Investigación Operativa en la Universidad de Vigo y de Matemáticas en la Universidad de A Coruña. También permitió iniciar y consolidar la existencia de una gran masa crítica de investigadores en Inferencia no Paramétrica en las tres Universidades Gallegas.

Trabajar en la Universidad de Santiago de Compostela en compañía de todos esos brillantes jóvenes investigadores significó para mí un privilegio en lo que a formación se refiere. Como colofón, en colaboración con varias de esas personas, creamos nuestro primer grupo de investigación, dando pasos en importantes avances metodológicos.

Muchos de esos avances se reflejaron en la organización de lo que denominamos “Seminarios de Inferencia no Paramétrica”, de reducido tamaño pero con participación de investigadores relevantes de talla internacional. Se organizaron formalmente dos seminarios en la Universidad de Santiago y uno en la Universidad de A Coruña en el año 1993. Informalmente fueron muchos más, con reuniones mensuales a las que asistíamos los investigadores de las tres Universidades gallegas.

Más tarde en los años 2000 (Santiago de Compostela), 2005 (Coruña) y 2008 (Vigo) esos seminarios ya serían de carácter internacional y de referencia mundial, con la participación de los mejores investigadores del momento en Inferencia no Paramétrica. Sería muy difícil hacer una selección de los investigadores de referencia participantes a lo largo de los diversos congresos organizados. No obstante, y claramente condicionado por mis preferencias personales, me gustaría mencionar a los profesores Steve Marron de la Universidad de Chapel Hill en USA y al profesor Peter Hall de la Universidad de Camberra en Australia. El primero, actualmente muy activo, es un investigador de referencia mundial de la Estadística Matemática orientada a los objetos -de gran importancia en la metodología del “Big Data”-. El segundo, desgraciadamente fallecido en el año 2016, dos años después de haber sido nombrado doctor honoris causa por la Universidad de Cantabria, (Figura 8), visitó las universidades gallegas varias veces y fue uno de los mejores estadísticos de los últimos tiempos con una gran productividad científica. De él aprendí mucho, tanto a través de la lectura de sus trabajos como de la interactuación oral. Obtuvo muchos premios y distinciones, cuya descripción omitiré por razones obvias de brevedad.

Figura 8: Nombramiento del profesor Peter Hall, posicionado en el centro, por la Universidad de Cantabria en 2014.

Paralelamente en aquellos años también se desarrolló con gran intensidad una línea de investigación en Teoría de Juegos, en nuestro departamento en la Universidad de Santiago, iniciada en los años 80 por el profesor José Manuel Prada Sánchez y fortalecida de forma definitiva por su discípulo, Ignacio García Jurado, que actualmente ejerce en la Universidad de A Coruña. Miembros de este grupo de investigación y del nuestro de Inferencia no Paramétrica se unieron en los años 90 para crear la Sociedad Gallega para la Promoción de la Estadística y la Investigación de Operaciones (SGAPEIO). Se trata de una sociedad científica de carácter autonómico con vocación divulgativa y que organiza cada dos años eventos de carácter científico relacionados con nuestras disciplinas, (Figura 9).

Figura 9: 25 Aniversario de la Sociedad Gallega para la Promoción de la Estadística y la Investigación de Operaciones (SGAPEIO) (Fotografía tomada el 11/02/2017).

Otro hecho muy relevante, que tuvo lugar a la vez que los eventos mencionados, fue la transferencia a la industria. Desde finales de la década de los 80 y sobre todo en los primeros años de la década de los 90 se trabajó muy intensamente en el diseño de modelos de predicción semiparamétrica, con componentes de tipo paramétrico, propios de la famosa metodología Box-Jenkins, en combinación con estimadores no paramétricos de la función de regresión. Tal es el caso de los indicadores de polución del entorno de la central térmica de As Pontes de García Rodríguez. Esta colaboración, que mantenemos todavía en la actualidad, permitió transferir a la Central, entonces propiedad de ENDESA, una herramienta de seguimiento y predicción de eventos relacionados con episodios de polución de niveles críticos de SO_2 o NO_X. A este sistema lo denominamos "Sistema de Predicción Estadística de Inmisión" (SIPEI en las diversas versiones del software), (Figura 10).

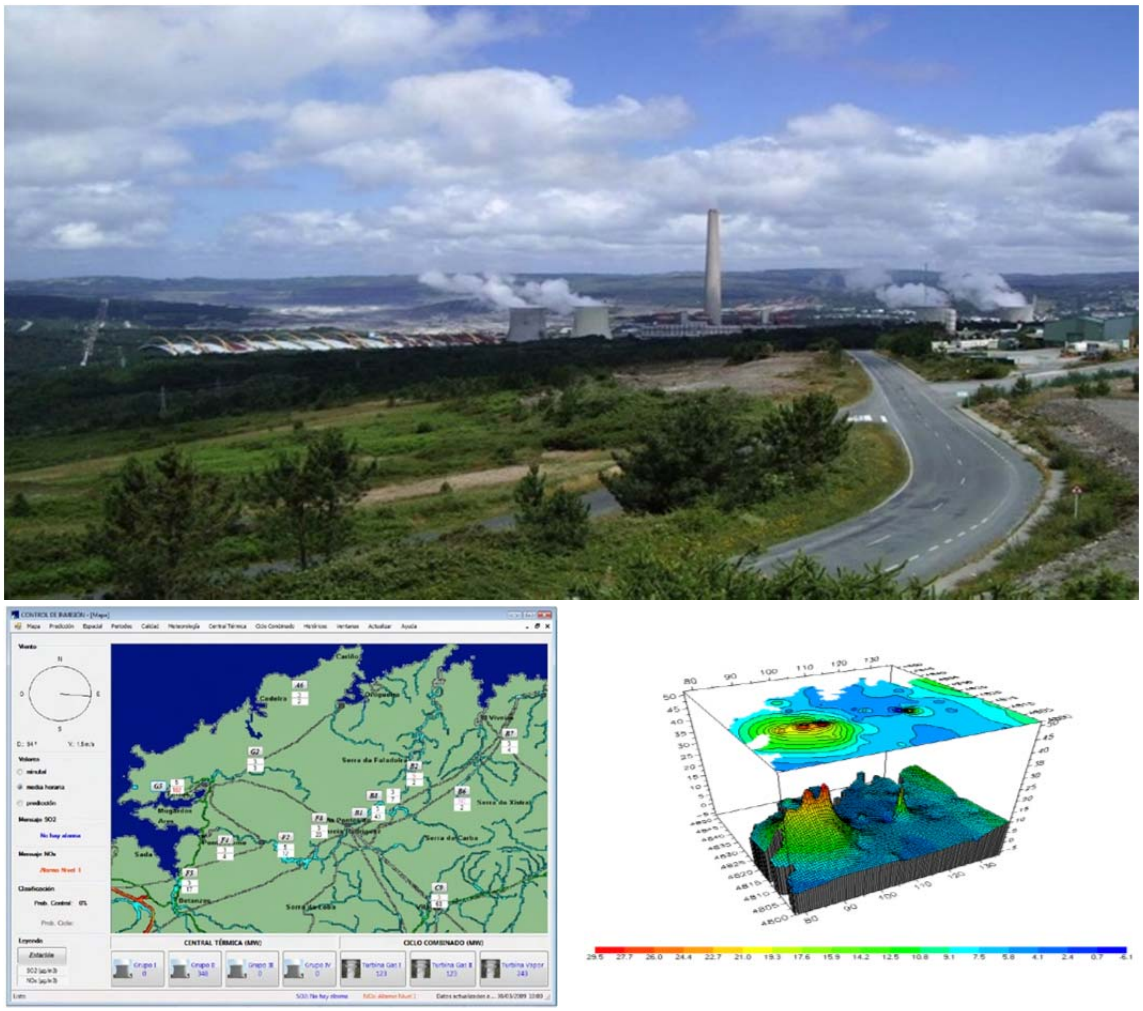


Figura 10: Parte superior: Imagen de la Central térmica de As Pontes de García Rodríguez. Parte inferior: (izquierda) Una de las pantallas ilustrativas del SIPEI. (Derecha) una superficie de predicción espacial de niveles de polución en torno a la central.

Sin duda este paso representó un avance extraordinario en herramientas estadísticas de gran utilidad y así queda reflejado en la tesis doctoral de *Manuel Febrero Bande* (1995) (23), uno de los grandes desarrolladores de este proyecto, y también en otras tesis doctorales posteriores relacionadas, como las de *Javier Roca Pardiñas* (2003) (24), *Tomás Cotos Yáñez* (2004) (25) y *Belén Fernández de Castro* (2004) (26).

Cabe destacar que en la metodología innovadora desarrollada, en relación a esta línea temática, y con antelación al boom del "Big Data" de la década siguiente, ya utilizábamos procedimientos muy próximos a los de uso en la era actual, como el concepto de la matriz histórica de información de los indicadores de polución. Esta nueva filosofía de la metodología estadística, consistente en usar de forma dinámica y cambiante la información disponible, tuvo un impacto muy importante en los estudios de la Estadística Aplicada al Medioambiente y a problemas relacionados.

Nuestra publicación en la revista “Technometrics” (1995) (27) y la organización del evento satélite del ISI, relativo a la aplicación al medioambiente y salud en el año 2003 (Figura 11), evidencian la relevancia de nuestros estudios.

Figura 11: Acto inaugural del “ISI Conference in Environmental Statistics and Health” en 2003.

De vuelta al plano metodológico, derivado de las ideas de nuestro primer trabajo de investigación, previamente mencionado, publicado en “Annals of Statistics” (1987) (3), se generó un línea de actividad científica muy productiva dedicada a los contrastes de bondad de ajuste para clases genéricas de modelos de regresión. La lista de tesis doctorales dirigidas o codirigidas en este contexto es amplia: *Luis Ramil Novo* (1998) (28), *Ana Pérez González* (2003) (29), *Juan Carlos Pardo Fernández* (2005) (30) y *Mercedes Conde Amboage* (2017) (31). En estos trabajos de investigación se cubrieron diversos aspectos de relevancia, como el uso de suavizadores tipo “spline”, el estudio de los contrastes de bondad de ajuste en datos incompletos y la modelización adaptada a la Estadística robusta

Ya a finales de la década de los 90 y primeros años del siglo XXI, la investigación que se desarrollaba en nuestro grupo tenía el gen previo de la investigación metodológica de los años 80 pero con una dosis importante de proyección a las aplicaciones y a la transferencia al sector productivo y económico.

Por ejemplo, en contextos de importancia para la Estadística Oficial, destacamos la dirección de tesis motivadas por las colaboraciones con el Instituto Gallego de Estadística (IGE). En especial los estudios derivados en "las técnicas de muestreo" y las diversas aplicaciones a la modelización de indicadores socioeconómicos, como por ejemplo el mapa de tasas de desempleo en la comunidad de Galicia. Así pues, fueron muy relevantes las aportaciones metodológicas derivadas de las tesis doctorales de *María José Lombardía Cortiña* (2002) (32) y *Carlos Iglesias Patiño* (2014) (33), este último, a su vez, también profesional del IGE.

Las tesis doctorales con corte de tipo aplicado tuvieron especial relevancia en otros ámbitos. Entre ellos, destacamos el estudio que se hizo con clasificadores innovadores, tipo SVM ("Support Vector Machine"), para datos genéticos de alta dimensión, desarrollados en la tesis de *Manuel García Magariños* (2010) (34) y en colaboración con el investigador de referencia en genética humana Antonio Salas.

Igualmente ponemos de relieve los avances en modelos predictivos en el entorno de la ingeniería y las finanzas con la tesis realizadas por *José María Matías Fernández* (2003) (35), en colaboración con investigadores en ingeniería de minas de la Universidad de Vigo y el trabajo de *Abelardo Monsalve Cobis* (2011) (36), de aplicación a la modelización de modelos de tipos de interés, en especial a datos del Euribor.

Y ya finalmente, no podemos dejar de mencionar los estudios realizados para la Hidrología, la Aerobiología y la Ingeniería Química, que se reflejaron en el trabajo de *María Castellano Méndez* (37). En esta última tesis, de marcado carácter trasversal, se usaron métodos predictivos diseñados con redes neuronales en un amplio marco de colaboración. A destacar la participación de miembros de la empresa Ferroatlántica, de la profesora María Jesús Aira Rodríguez, directora del grupo de Biodiversidad y Botánica Aplicada así como de sus diversos colegas de grupo, y finalmente del también académico de la RAGC, el profesor Juan Lema, director del grupo de Biotecnología Medioambiental de la Universidad de Santiago de Compostela.

Ya en el siglo XXI, encuadrados en un contexto en el que la información suministrada a nivel macro o micro es masiva, las líneas de investigación confluyeron en el desarrollo de procedimientos con gran complejidad:

- En aquellos problemas en los que la información de los datos se recibe cronológicamente en alta frecuencia, surgen los llamados datos funcionales. Aquí destaca el trabajo de tesis doctoral desarrollado por *Adela Martínez Calvo* (2013) (38) con la construcción y estimación de modelos de regresión de variables explicativas funcionales. En esta línea de investigación nuestro grupo adquiere mucha relevancia internacional. En especial, cabe destacar la creación de la librería en R: "fda.usc", por parte de Manuel Febrero Bande y en colaboración con su

estudiante de doctorado Manuel Oviedo de la Fuente. Esta librería tuvo más de 30000 descargas en los últimos cuatro años, (Figura 12).

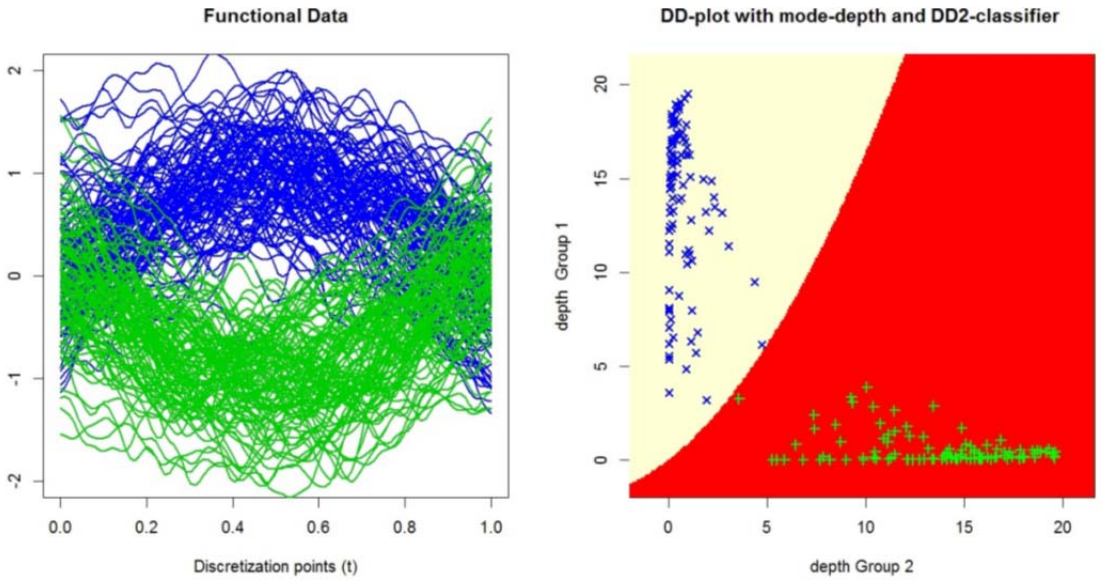


Figura 12: Clasificación de dos conjuntos de datos funcionales mediante el procedimiento "DD-plot" realizado con la librería "fda.usc".

- En los problemas en los que la información es recibida con coordenadas georreferenciadas, como por ejemplo los puntos de ignición de un incendio forestal, los puntos de localización de una especie en un "habitat", las coordenadas de localización de un delito en una ciudad (de importancia para la criminología) o las coordenadas de aparición de un nido de avispas velutinas, como ejemplo de gran impacto actual. Esta fuente de datos representa en los últimos años un gran impulso para la modelización de la Estadística Espacial o Espacio Temporal y de los llamados Procesos Puntuales. En las tesis doctorales de *Rosa María Crujeiras Casais* (2007) (39), *Isabel Fuentes Santos* (2016) (40) y *María Isabel Borrajo García* (2018) (41) se llevan a cabo diversos avances en esa línea temática.

  Con un enfoque dirigido a los datos de carácter espacial o espacio temporal en el dominio de las frecuencias, en la tesis de Rosa, se derivan diversas aplicaciones encaminadas al diseño de mapas de concentración de niveles de marcadores de polución. Este estudio fue realizado en colaboración con los compañeros del departamento de Edafología de la Universidad de Santiago de Compostela.

En un marco de colaboración muy reciente con el Instituto de Matemática Pura y Aplicada (IMPA) de Río de Janeiro, se aplica la modelización de los procesos puntuales al diseño de mapas de eventos criminales con datos suministrados por la red social “Fogo cruzado”.

Por otro lado, los trabajos de Isabel y María Isabel, se centran en la creación de mapas de riesgo de incendios forestales usando procesos puntuales y son realizados en colaboración con el profesor Manuel Marey del departamento de Ingeniería Agroforestal, (Figuras 13a, 13b y 13c).

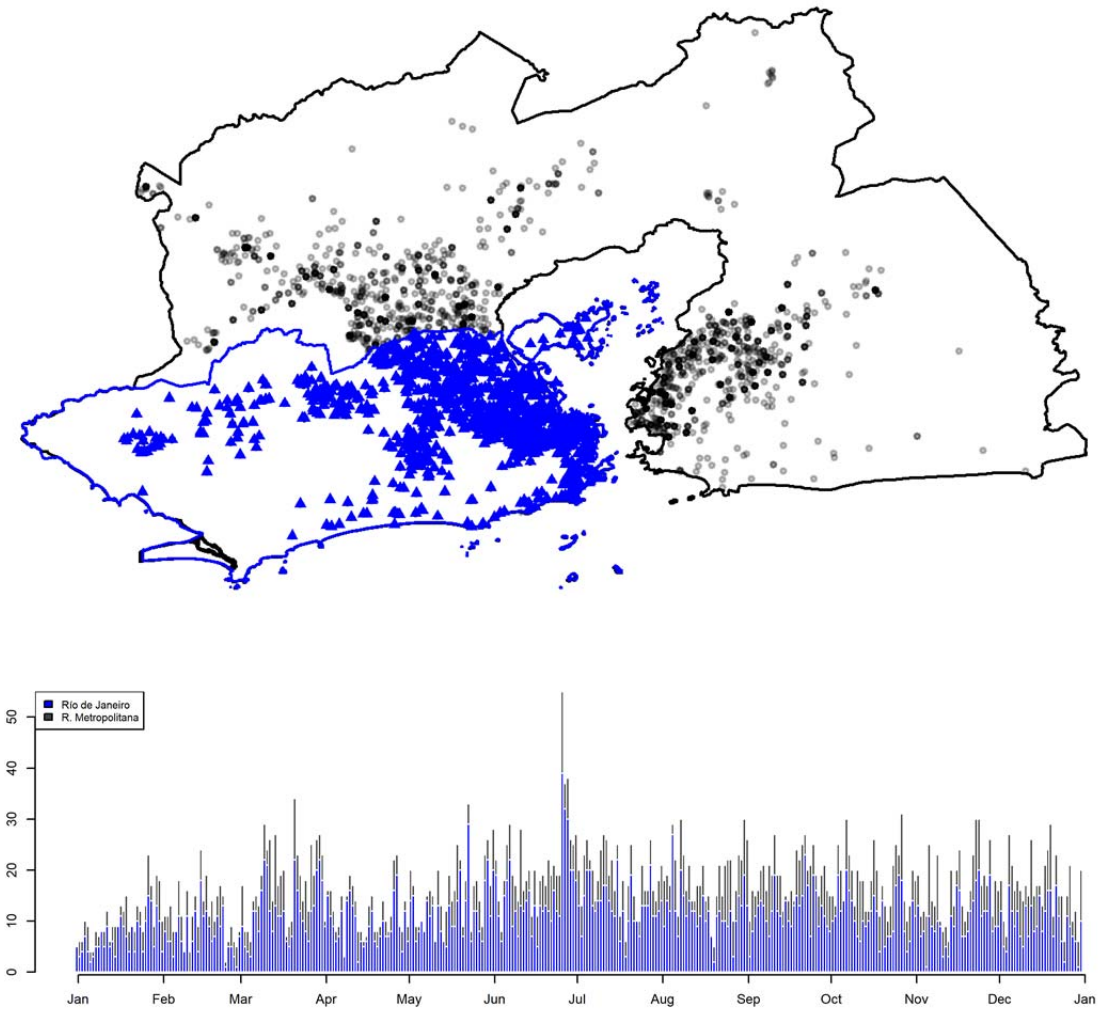


Figura 13a: Mapa superior con datos georreferenciados de crímenes en Río de Janeiro durante el 2017. Gráfico inferior con la evolución del número de crímenes a lo largo de ese año.

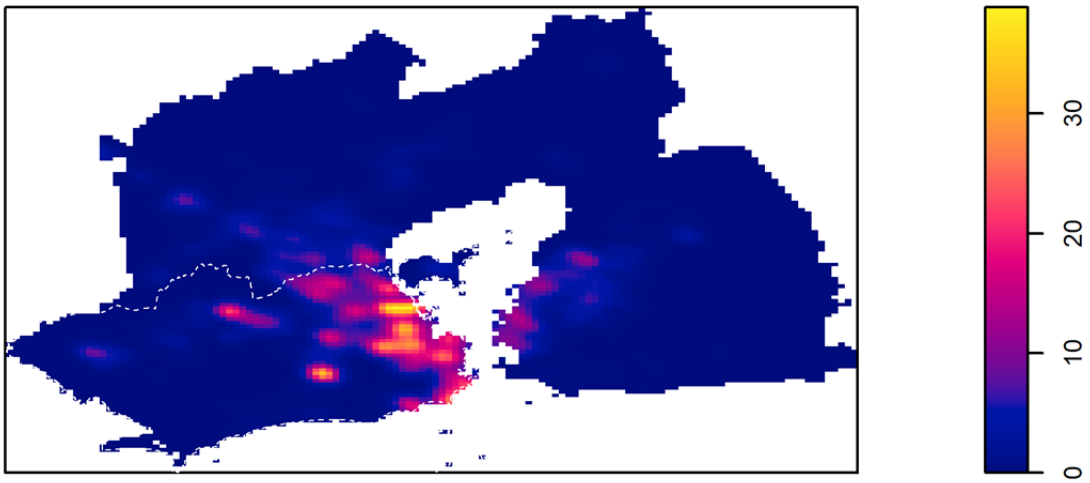
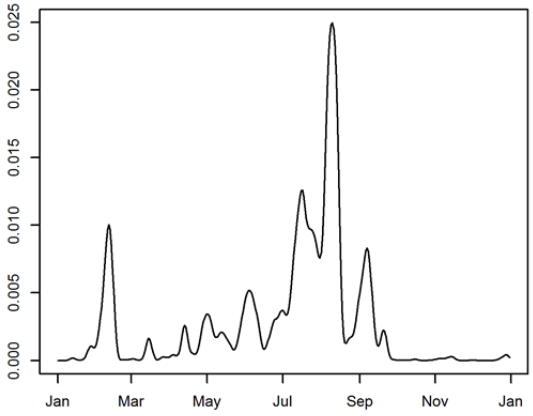
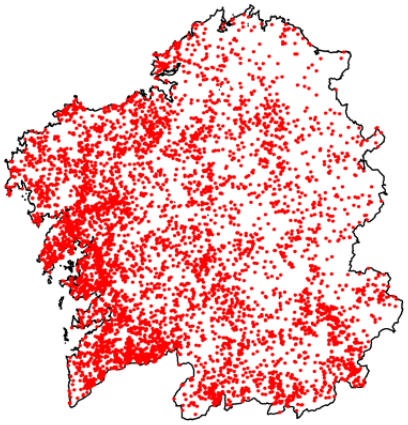


Figura 13b: Mapas de intensidad estimada del número de crímenes en Río de Janeiro en base a los datos de la figura 13a.





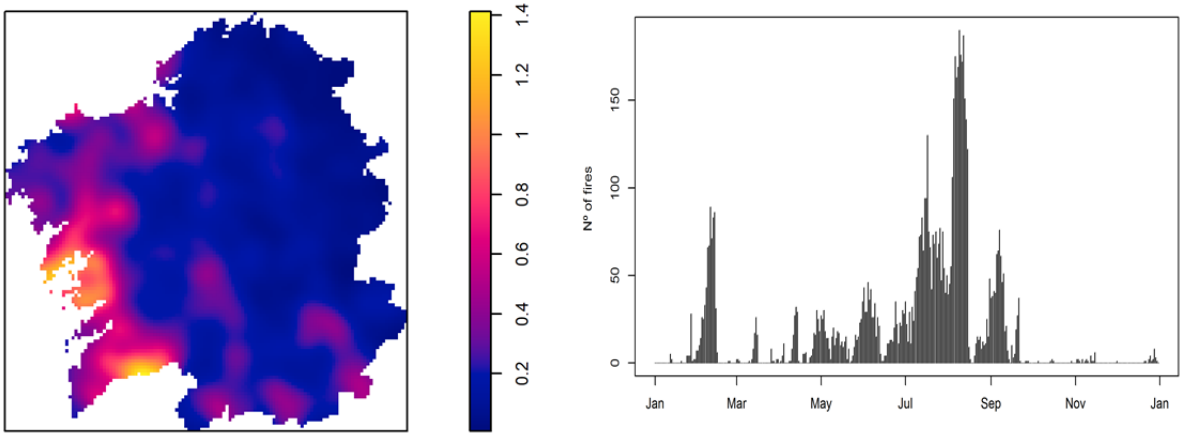


Figura 13c: Parte derecha, datos de incendios de Galicia, año 2006, vistos temporal (abajo) o espacialmente (arriba). Parte izquierda, estimación de los mapas de intensidad espacial (abajo) y temporal (arriba) correspondientes a dicho año.

El gran desarrollo de la información suministrada para campos de gran relevancia actual, como el "reconocimiento de imágenes", también representó un gran impulso y motivación para la línea de investigación de nuestro grupo en el contexto de estimación de conjuntos. Son dignos de mencionar los trabajos de las tesis doctorales realizadas por *Alberto Rodríguez Casal* (2003) (42) y *Paula Saavedra Nieves* (2015) (43). Este campo de investigación está siendo actualmente muy fructífero, tanto por las publicaciones de prestigio que se están consiguiendo en revistas de máximo nivel, como por el desarrollo de software y aplicaciones.

Es digno de mencionar la librería en R: "alphahull", creada por Beatriz Pateiro López, con el uso de desarrollos metodológicos de su tesis dirigida por Alberto y que también tuvo numerosas descargas, al igual que la "fda.usc", desde su creación (Figuras 14 y 15).

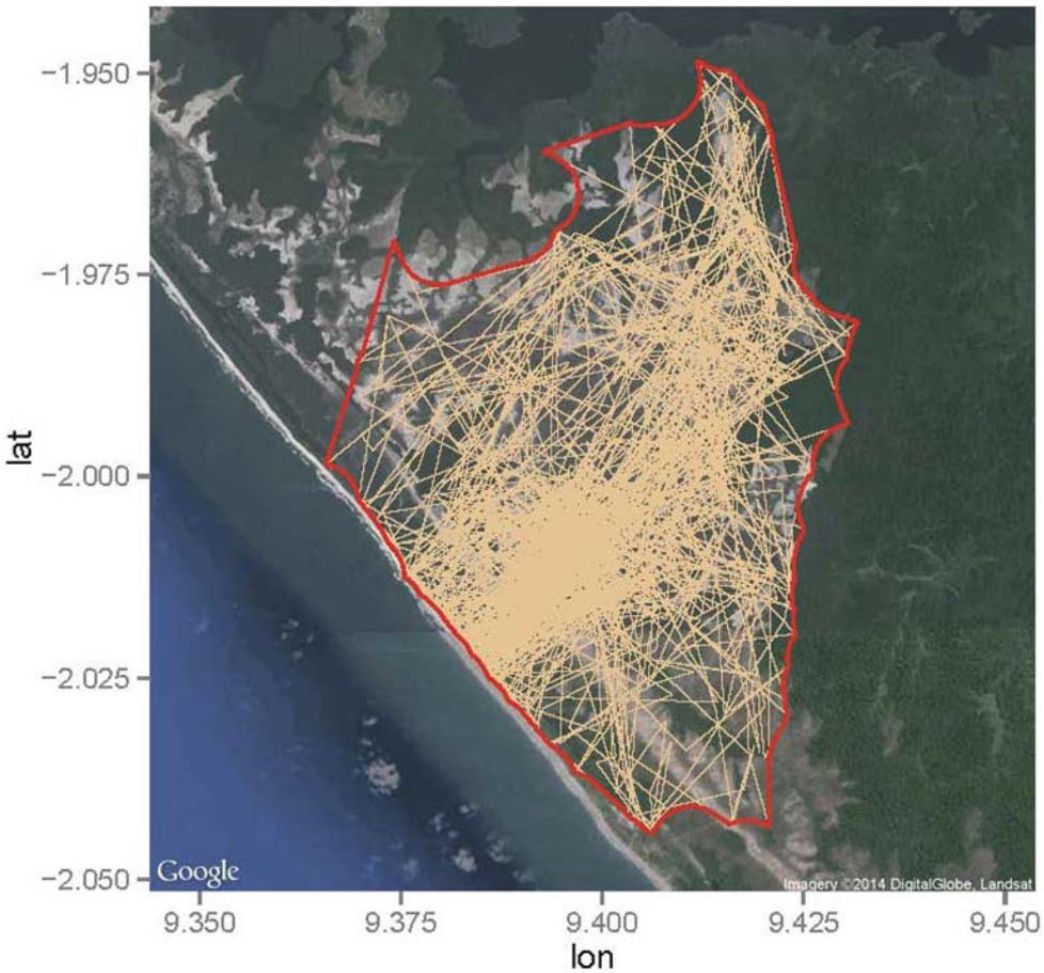


Figura 14: Estimación del "home range" del movimiento de un elefante con n=1633 posiciones. En rojo, la frontera del estimador "r-convex hull". Gráficas obtenidas mediante el paquete "alphahull".

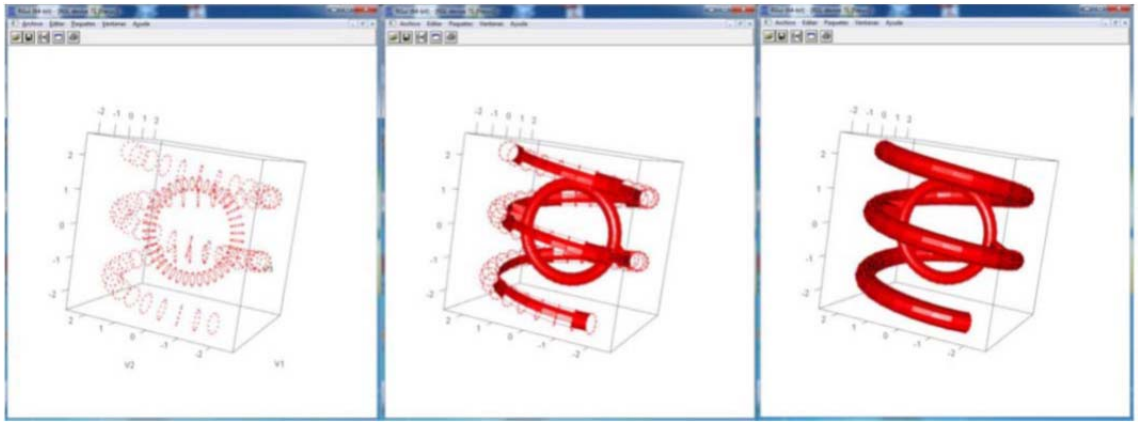


Figura 15: Estimación con el paquete "alphashape3d" de un conjunto en base a una nube de puntos en el espacio euclídeo tridimensional.

Finalmente, y dentro de la Estadística orientada a objetos, nos centramos en el desarrollo de procedimientos de inferencia para datos observados en variedades diferenciables en general, como pueden ser los datos en la esfera, el cilindro o el toro,

entre otros. En esa línea de investigación destaca la tesis doctoral realizada por *Eduardo García Portugués*, el cual desarrolló metodología inédita en el contexto de las técnicas de suavización para datos en variedades. En la parte práctica de su estudio se llevó a cabo una colaboración con la profesora Josefina Ling-Ling de nuestra Facultad de Matemáticas (2014) (44) con aplicación a datos de Astronomía (como ejemplo de datos esféricos) y con el profesor Pereira, de la Universidad de Lisboa, en el estudio de bases de datos de incendios en Portugal (como ejemplo ilustrativo de datos cilíndricos, con la consideración de las variables dirección e intensidad del viento en la propagación de un incendio), (Figuras 16 y 17).

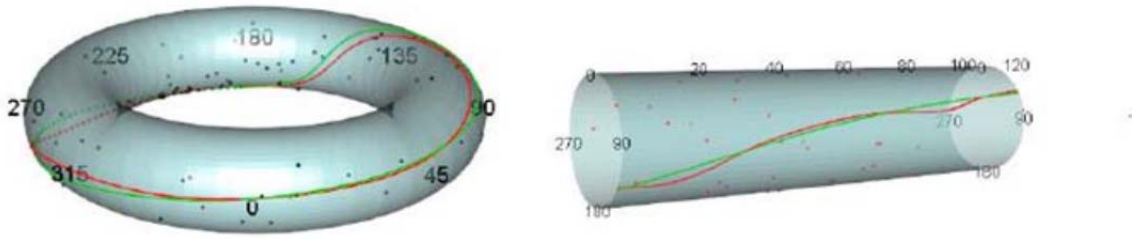


Figura 16: Ejemplos de estimaciones de funciones de regresión con respuesta circular y covariable circular (parte izquierda) y respuesta lineal con covariable circular (parte derecha).

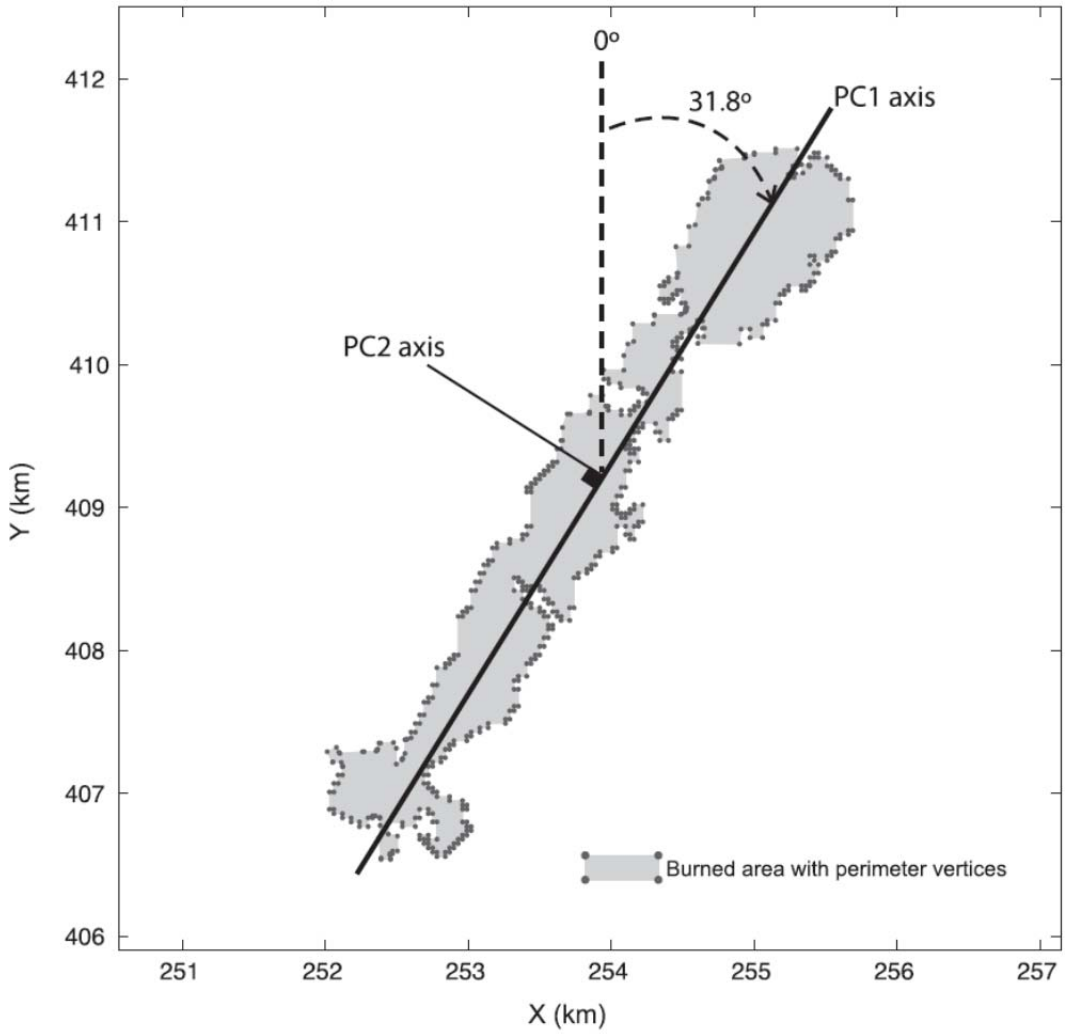


Figura 17: Ejemplo de dato direccional-lineal. La dirección dominante del viento como dato angular y la superficie quemada como dato de respuesta escalar.

Un aspecto dominante en la mayor parte de las tesis doctorales realizadas en los últimos años es la internacionalización de las mismas, con doctorados de carácter europeo o con mención internacional. Varias de estas tesis fueron codirigidas con profesores de relevancia internacional, como es el caso de los profesores Vieu y Ferraty de la Universidad de Toulouse (hijo y nieto científicos respectivamente del mencionado al principio de este discurso, profesor Collomb) o de la profesora Ingrid Van Keilegom (actualmente en la Universidad de Leuven, Bélgica), colaboradora científica con nuestro grupo de investigación y actualmente galardonada con una posición de “advanced grant” en Matemáticas, para desarrollos en procedimientos de inferencia no paramétrica, (Figura 18).

Figura 18: Tesis doctoral de Adela Martínez Calvo (posicionada en el centro). De izquierda a derecha, los profesores Frederic Ferraty, Ingrid Van Keilegom y Philippe Vieu junto a otros miembros de la comisión.

En este viaje maravilloso de treinta años, que acabo de describir, participando activamente en la dirección o codirección de las 32 tesis doctorales, hay muchos protagonistas. En especial, destaco a los estudiantes de doctorado como actores principales; le siguen los codirectores de las tesis y muchos otros investigadores además de los ya mencionados.

Mi mayor agradecimiento por lo que aprendí de todos ellos.

La metodología de la Estadística Matemática en la era del “Big Data”.

La aparición de los ordenadores supuso para la Estadística un cambio con implicaciones en el paradigma metodológico. Los avances científicos, derivados de las mencionadas generaciones de estadísticos de la primera mitad del siglo XX, estaban dirigidos a comprender la información de las bases de datos existentes. Estas eran de tamaño medio y, en muchos casos, logradas con gran esfuerzo.

Un ejemplo lo tenemos en los estudios en agricultura de Fisher, quien desarrolló metodología de diseño de experimentos y tuvo que esperar varios meses para el análisis estadístico de sus datos, mientras evolucionaba el terreno de las parcelas sometidas a cambios en sus condiciones.

Otro ejemplo más reciente lo constituye el estudio de las llamadas curvas de crecimiento en pediatría, analizadas por Gasser y otros (1984) (45), derivado de una base de datos longitudinales iniciada en los años 50 con la colaboración de la clínica pediátrica de la Universidad de Zurich. En este análisis se efectúa el seguimiento individualizado de más de 15 años en grupos de personas, desde su nacimiento hasta la adolescencia, para posteriormente inferir sobre la velocidad y la aceleración del crecimiento de las mismas en función del sexo.

La Estadística desarrollada en la primera parte del siglo XX estaba encaminada al diseño de algoritmos (como por ejemplo, el ajuste de mínimos cuadrados o el cálculo de los valores de un estadístico de bondad de ajuste) en donde el esfuerzo computacional era considerable para una capacidad de cálculo muy limitada. Se estimaban parámetros reales, y no las curvas, por la escasa capacidad gráfica. Se analizaba cuál podía ser el comportamiento de las distribuciones de los estadísticos, bajo ciertas hipótesis, por la imposibilidad obvia de tener tamaños de muestra grandes y reflejarlos en un ordenador.

En los años 80, la introducción versátil de los ordenadores generó nuevos procedimientos estadísticos basados en la simulación de datos artificiales, como el mencionado “Bootstrap”.

Pero es realmente en las décadas siguientes, con la generalización del uso de Internet y de las nuevas tecnologías de la información, con el desarrollo del cálculo distribuido y en paralelo, y con la disminución de los costes computacionales para el almacenamiento y procesado de las grandes bases de datos, cuando podemos establecer el comienzo de la era del “Big Data”.

Una pregunta obvia es a qué nos estamos refiriendo cuando hablamos del término lingüístico “Big Data”. En general, diremos que un conjunto de datos es del orden “Big Data” cuando es muy difícil tratarlos con metodología tradicional estadística o gestionarlo con los recursos computacionales básicos. A su vez, cabe mencionar que otros muchos términos lingüísticos surgieron con la explosión del “Big Data”. El detalle explicativo de los mismos lo obviamos en este discurso. Por citar algunos de esos términos, en los últimos años aparecen de forma abundante palabras como: “Machine Learning”, “Artificial Intelligence”, “Data Analytics” y “Data Science”.

En la actualidad las empresas cuentan con sistemas de lectura de datos en tiempo real y de forma continuada; las cotizaciones bursátiles se pueden hacer con datos de alta frecuencia; las diversas páginas web, las redes sociales o simplemente los usos de tarjetas de crédito, generan información masiva. En el ámbito de la energía eléctrica se tienen mediciones de alta frecuencia sobre consumo o demanda (Figura 19). En todos estos casos puede resultar muy importante ser capaces de procesar y controlar toda esta información.

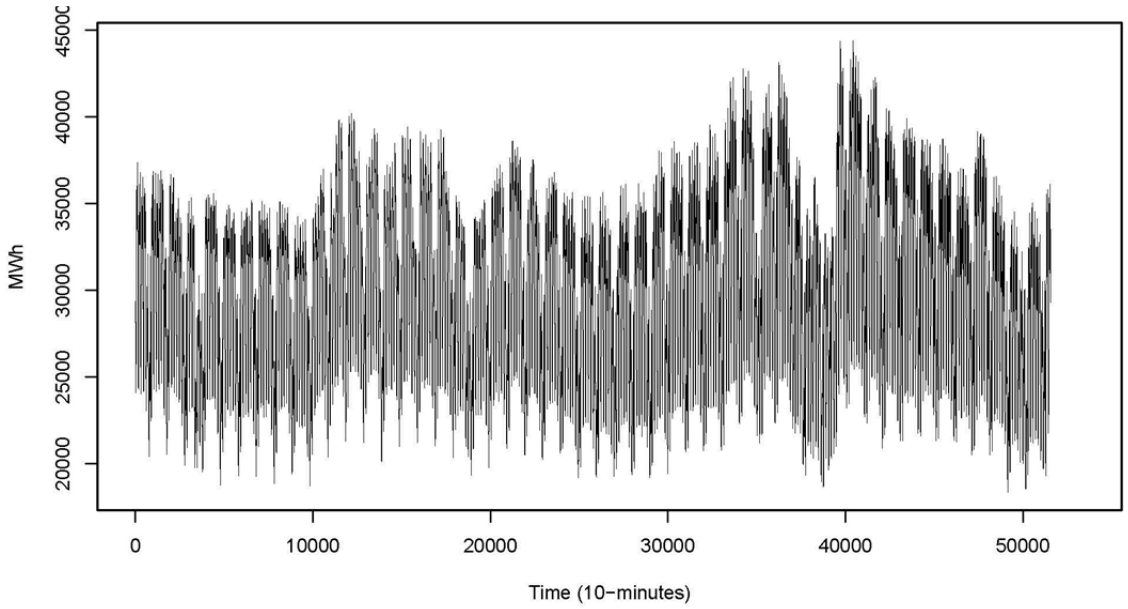


Figura 19: Mediciones, cada 10 minutos, de la demanda de energía eléctrica en España en el año 2008.

El enfoque estadístico tradicional estaba pensado para analizar de forma muy eficiente pequeñas muestras de datos medidos con bastante precisión y homogeneidad. La realidad actual nos lleva a bancos de datos masivos, con medidas heterogéneas, procedentes de distintas poblaciones y con formatos diversos que pueden incluir imágenes, gráficos, textos, arboles filogenéticos…etc. En general, el concepto de vector de valores numéricos de los datos de una muestra pasa a ser el de un conjunto, de lo

que podríamos llamar objetos estadísticos, y que hace de esta nueva situación un foco de investigación de gran proyección en la Estadística Matemática hacia el futuro.

Con el enfoque previo de los creadores del Análisis Multivariante, la información suministrada para el estudio posterior se reflejaba en una matriz de datos de dimensión "n*x*p", llamada de individuos por variables, en donde "n" representaba el tamaño de muestra o número de observaciones bajo estudio y "p" el número de las características observadas sobre los mencionados individuos. La mayor parte de desarrollos metodológicos realizados en la primera mitad del siglo XX fueron para variables de tipo Gaussiano, con "p" de una dimensión reducida y "n" bastante más grande que "p". De aquí salieron las distribuciones con nombre propio para estadísticos unidimensionales (p=1) o multidimensionales (p>1). Son ejemplos de ello, la distribución normal, la ji cuadrado de Pearson, la t de Student o la F de Snedecor entre las unidimensionales; y la distribución de Wishart, la de Wilks o la de Hotelling entre las multidimensionales más importantes. Los datos habituales eran biomédicos, de tipo experimental o de tipo socioeconómico.

En los últimos años está situación cambió drásticamente con la multitud de ejemplos que surgen, en los que "p" puede ser mucho más grande que "n" (frecuente, por ejemplo, en datos de carácter genómico) o incluso en aquellas otras situaciones en donde el tamaño de muestra "n" se hace gigantesco, como por ejemplo en los datos de alta frecuencia para el ya comentado mercado eléctrico. Estos dos hechos, que se conocen como de matrices muy altas o muy anchas, según sean muy grandes los valores de "p" o "n", hacen necesario que los procedimientos creados hasta hace aproximadamente tres o cuatro décadas necesiten un re-escalamiento para la aplicación práctica. Esto no siempre es posible, como bien comentan Daniel Peña (2014) (46) o Ricardo Cao (2017) (47) en sus magníficos documentos publicados en los boletines de las Sociedad de Estadística e Investigación Operativa.

También, como consecuencia de los problemas en alta dimensión se hace muy necesario el desarrollo de nuevos procedimientos basados en proyecciones de los datos a una dimensión inferior. Téngase en cuenta, que cuando la dimensión "p" crece notablemente la noción de vecindad entre los datos desaparece. Si simulamos, por ejemplo, 1000 observaciones en el cubo unidad de dimensión "p", se puede ver (Figura 20) que el histograma de "distancias entre puntos" en el espacio p dimensional crece notablemente, mostrándose empíricamente que los puntos se van aislando. El llamado fenómeno del "desastre de la dimensionalidad".

También es muy útil, para este contexto de la alta dimensión, la construcción de nuevos mecanismos de selección de variables. El desarrollo de estos procedimientos nos lleva a su vez a otro problema difícil de abordar, los contrastes de significación múltiples cuando el número de hipótesis a contrastar es también de alta dimensión. Téngase en cuenta que, paradójicamente, en "Big Data" las variables estadísticas

suelen ser significativas debido a la cantidad masiva de datos, produciéndose numerosas correlaciones significativas de carácter artificial o espurio. Ejercicios de simulación nos indican que variables aleatorias, incorreladas en su estructura poblacional, muestran correlaciones muestrales significativas cuando aumenta la dimensión p, (Figura21).

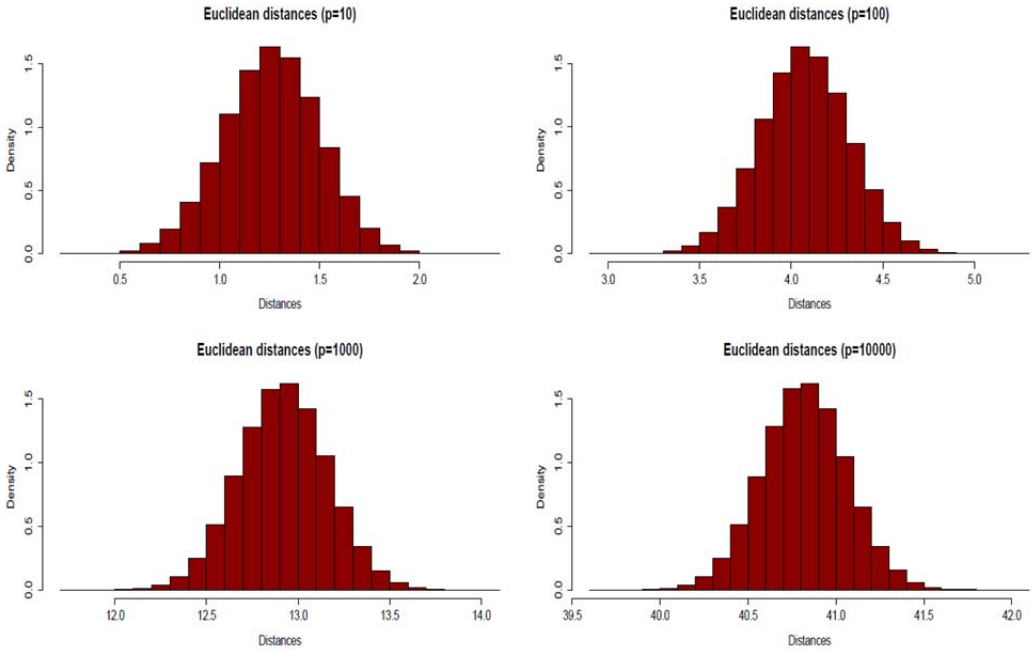


Figura 20: Histogramas de las distancias entre los puntos simulados en un cubo unidad de dimensión p= 10, 100, 1000 y 10000 con una muestra de tamaño 1000. Gráficos extraídos de la presentación del profesor Pedro Galeano de la Universidad Carlos III con motivo de su seminario en "Big Data" en Santiago en 2015.

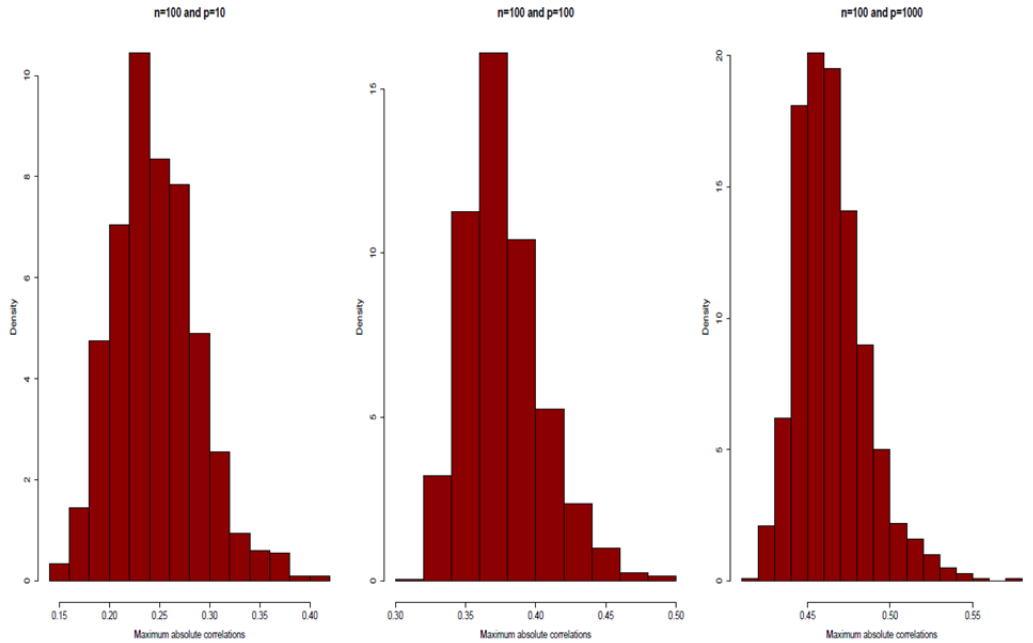


Figura 21: Histogramas construidos con las 1000 mayores correlaciones muestrales obtenidas con 1000 conjuntos de datos vectoriales, simulados de una normal multidimensional estándar, para los ejemplos (n,p)= (100,10), (100,100) y (100, 1000). Gráficos extraídos de la presentación del profesor Pedro Galeano de la Universidad Carlos III con motivo de su seminario en "Big Data" en Santiago en 2015.

Una de las aspiraciones tradicionales en el desarrollo de la metodología estadística es la obtención de métodos óptimos resultantes de la minimización del riesgo asociado a las estimaciones. No obstante, dadas las características del nuevo paradigma que genera el "Big Data", los procedimientos eficientes desde el punto de vista estadístico pueden resultar de difícil computación en tiempo reducido en muchos problemas. Se hace, por tanto, necesaria la incorporación simultánea de la eficiencia computacional en los nuevos procedimientos a considerar.

Está claro que la era del "Big Data" ofrece grandes oportunidades para el desarrollo de la Inferencia Estadística. Son varios los volúmenes especiales de revistas de prestigio que se dedican monográficamente a esta temática en los últimos años. En especial menciono el artículo de Franke y otros (2016) (48), escrito por 14 coautores de relevancia mundial y en el que se tratan conclusiones del encuentro monográfico que organizó el "Canadian Statistical Sciences Institute" entre enero y junio del 2015.

La dificultad de transformar el "Big Data" en conocimiento se debe en gran medida a las "Vs", en particular volumen, velocidad, variedad y veracidad.

- En cuanto al volumen, desde el punto de vista de la computación se persiguen infraestructuras especiales, como clústeres distribuidos Hadoop que combinan la computación en paralelo y distribuida, permitiendo el desarrollo de tareas intensivas de computación, dividiéndolas en pequeñas partes y distribuyéndolas en un conjunto tan grande como se quiera de máquinas. La tecnología Hadoop introduce nuevas técnicas de programación más accesibles para el almacenamiento de datos masivos estructurados o no. Sin embargo, desde el punto de vista de la Estadística, con la nueva metodología se pretende una potenciación del análisis exploratorio de datos que combine todos los avances computacionales con la inteligencia metodológica del matemático.
  Un ejemplo muy claro, el que menciona Ricardo Cao (2017) (47), con el argumento de "divide y vencerás". En particular lo sugiere para la construcción de nuevos estadísticos. Así, con sus propias palabras, "la construcción de una simple función de distribución empírica aplicada a una muestra del orden de 10 elevado a 10 datos, para un ordenador que realizase 100 millones de operaciones por segundo, le llevaría sobre 10 horas". Similares problemas de tiempo de computación, tendría el cálculo de una mediana, ya que habría que hacer uso de la ordenación de este tamaño masivo de datos.
  Para reducir el tiempo de computación, Ricardo Cao (2017) (47) sugiere la construcción de submuestras de tamaño "m", la construcción de los estadísticos sobre esas submuestras y la combinación de estos

estadísticos valorados en las submuestras para la obtención del resultado final. Un ejercicio matemático, cuya solución se puede beneficiar, sin duda, de la mencionada computación en paralelo. Esta estrategia, de división en submuestras, puede incluso ser, a veces, eficiente desde el punto de vista de la precisión estadística para ciertas combinaciones de valores “m” y “n”.

- En relación a la velocidad, la capacidad computacional permite procesar información “on line” muy rápidamente. Ello hace que desde el punto de vista de la Estadística debamos adaptar los algoritmos diseñados para muestras recibidas de una sola vez para aquellos casos en que el suministro de información sea secuencial. Para algunas aplicaciones, como en la Física de altas energías, la recepción de datos es tan rápida que a veces es imposible el análisis de toda la información y se requiere la colaboración de los expertos del campo para asegurar que información relevante debe ser retenida para el objeto en estudio.
  Por otro lado, para la agilización del cálculo, también es necesario sacar partido de los esfuerzos que se están haciendo con la paralelización a todos los niveles. Plataformas como “Amazon Web Services” o “Google” ya ofrecen procesado en la nube en contraposición a los cálculos realizados en un cluster de computación físicamente localizado. El entorno de computación Apache Spark también está ganando popularidad y sustituyendo a Hadoop para ciertas tareas. Por ejemplo, un elemento positivo a favor de Spark es el “interface” de programación para lenguajes como Python o R, de gran interés en “Machine Learning” y en Estadística Computacional.

- En lo que concierne a la variedad, desde el punto de vista de la computación nos encontramos con la necesidad de procesar información diversa, como imágenes, redes sociales, grabaciones de audio, videos, redes de sensores,…etc, mientras que desde la perspectiva de la Estadística Matemática necesitamos modelar, en vez de datos numéricos, objetos estadísticos o datos complejos. Esto hace necesario crear nuevas estructuras de tipo topológico que permitan cuantificar la proximidad entre la variedad de los diversos objetos estadísticos y generalizar con otro tipo de métricas a las técnicas tradicionales, como las componentes principales o el análisis cluster de entre las diversas metodologías desarrolladas en la primera mitad del siglo XX.
  Como comentaba el propio profesor Steve Marron, en el pasado mes de junio en Madrid, en la conferencia sobre “Big Data” organizada por la

fundación Gadea, el llamado “Object Oriented Data Analysis” cambia los átomos estadísticos, que tradicionalmente eran datos escalares o vectoriales, por datos funcionales, mediciones espectrales, imágenes, formas, arboles, vídeos,…etc., que tienen que ser dotados de nuevas estructuras topológicas.

- Finalmente, en relación a la veracidad, cabe citar que desde el punto de vista computacional, con los argumentos antes detallados de las altas prestaciones computacionales, somos capaces de almacenar una fuente masiva de información. No obstante, desde la óptica de la Estadística, debemos ser muy cuidadosos con la información suministrada. Muchas veces dicha información es el resultado de la fusión de varios conjuntos de datos coleccionados de forma anárquica y sin contemplar los objetivos de estudio a llevar a cabo. Por citar un ejemplo, la calidad y veracidad de los datos que se requieren en las fuentes propias de la Estadística Oficial, es un objetivo primordial que puede chocar claramente con la lluvia de información de carácter sesgado o interesado de algunas fuentes informativas. Diseñar metodología que permita la limpieza de esa información es también un objetivo de investigación de la Estadística en los años venideros.

Más allá de esas Vs, u otras, el “Big Data” tiene que ver con la naturaleza humana y por tanto con los aspectos éticos y de privacidad de la información suministrada y en movimiento. De forma especial, los algoritmos predictivos sobre la evolución de una enfermedad o sobre las posibilidades que podría tener una persona de contraerla, contemplando información genética, deberían tener un carácter confidencial. El conocimiento de tal información por parte de una aseguradora podría cambiar completamente las reglas del juego así como el importe del precio a pagar por un supuesto seguro de vida asociado. El encriptado de la información y los aspectos jurídicos asociados a la misma también van a formar parte de esta revolución que representa el Big Data. Al igual que la energía nuclear puede ser beneficiosa o nociva según sus fines, lo mismo sucede con el “Big Data”, que puede ayudar o destruir nuestra sociedad.

Aunque la Estadística es una rama rigurosa de las Matemáticas, repetidas veces fue objeto de comentarios jocosos. Como este del dramaturgo, crítico y polemista irlandés y premio nobel de literatura George Bernard Shaw. “La Estadística es una ciencia que demuestra, que si mi vecino tiene dos coches y yo ninguno, los dos tenemos uno”.

No obstante, en esta era del “Big Data”, los procedimientos estadísticos adaptados y diseñados para analizar las nuevas cantidades masivas de datos, están muy lejos de comentarios como el anterior y tienen que desempeñar, con la ayuda de las Ciencias de la Computación, un papel fundamental. Especialmente si tenemos en cuenta que el

90 por ciento de la información disponible actualmente se ha generado en los últimos tres años.
Adjunto a continuación algunos ejemplos, en diversos contextos de interés actual y futuro, en donde la información que se adquiere tiene o puede tener la naturaleza de "Big Data" requiriéndose del uso de técnicas sofisticadas de la Estadística Matemática o de Ciencias de la Computación.

1) La creación de mapas de la distribución de aves en tiempo real a nivel europeo en base a la información suministrada.

2) El control en continuo que se genera de la información sensorial en los sistemas de producción de la Industria 4.0.

3) El uso de las llamadas técnicas de "deep learning" en la gran masa de datos que se genera del campo de la Astrofísica.

4) El diseño de técnicas de clasificación para el entendimiento con los datos del Eurobarómetro en los problemas de emigración.

5) El uso de nuevos procedimientos estadísticos en la Quimiometría para el análisis de datos, de nueva naturaleza, como son los datos ómicos.

6) La aplicación de modelos de predicción y de "Machine Learning" a datos geo-localizados para el estudio de la eficiencia energética.

7) El uso de procedimientos de predicción a tiempo real de aparición de bancos de algas en ubicaciones de producción en acuicultura utilizando matrices históricas de información.

8) Algoritmos predictivos a tiempo real en base a información masiva del seguimiento de indicadores epidemiológicos.

9) Diseños de procedimientos de clasificación y filtrado de las llamadas "fake news" en las fuentes de información de las redes sociales.

10) Diseños de técnicas de clasificación de reconocimiento de caras de las personas en sistemas de seguridad utilizando la Estadística orientada a objetos.

11) Modelización predictiva de los costes de la energía utilizando información diversa estructurada o no estructurada.

12) Estudios de la integración de las grandes bases de datos informativas para decisiones clínicas de mucha relevancia, como por ejemplo los trasplantes de órganos en base a las características, moleculares u otras, asociadas al enfermo potencial

¿Alguien tiene alguna duda de la utilidad de la Estadística?

La Estadística Matemática gallega en la era del “Big Data”.

Para mí es un placer y enorme satisfacción vivir este momento con energía positiva y con la experiencia de un cierto recorrido de participación en la actividad investigadora de Galicia. A la investigación le dediqué mucho tiempo en los mejores años de mi vida, y en este momento ésta me corresponde con toda una generación maravillosa de investigadores repartidos por las tres universidades gallegas, además de aquellos otros, que dirigieron su actividad profesional a otros entornos académicos o al sector privado.

De la generación de mis 32 doctores se derivaron otras generaciones, que podríamos llamar de nietos o bisnietos científicos, que en su conjunto podrían representar una masa crítica de investigadores en “Inferencia no Paramétrica” o temas relacionados, de más de 50 personas actualmente en la comunidad gallega. Soy conocedor de 4 proyectos competitivos de investigación, del plan nacional, vigentes en este momento en esta temática y de otros proyectos o redes que conectan con nuestros grupos de investigación a nivel internacional y con ámbitos del sector productivo.

Nuestros investigadores desarrollan actividad científica en institutos de investigación, como el de Matemáticas en la Universidad de Santiago y participan en actividades de transferencia en entidades como el Instituto Tecnológico de Matemática Industrial (ITMATI), o en otros centros tecnológicos, en los que se presentan temas relacionados con la estadística oficial, la estadística computacional, el ámbito empresarial, el de las finanzas y en los nodos de asesoramiento estadístico para la salud.

Ya hace dos años que, con motivo del homenaje que recibí por mi 60 cumpleaños, pude darme cuenta con gran detalle de la gran cantidad de investigadores que conforman el mapa científico actual de la Estadística en Galicia con un potencial de relevancia competitiva a nivel internacional (Figura 22 y Figura 23).

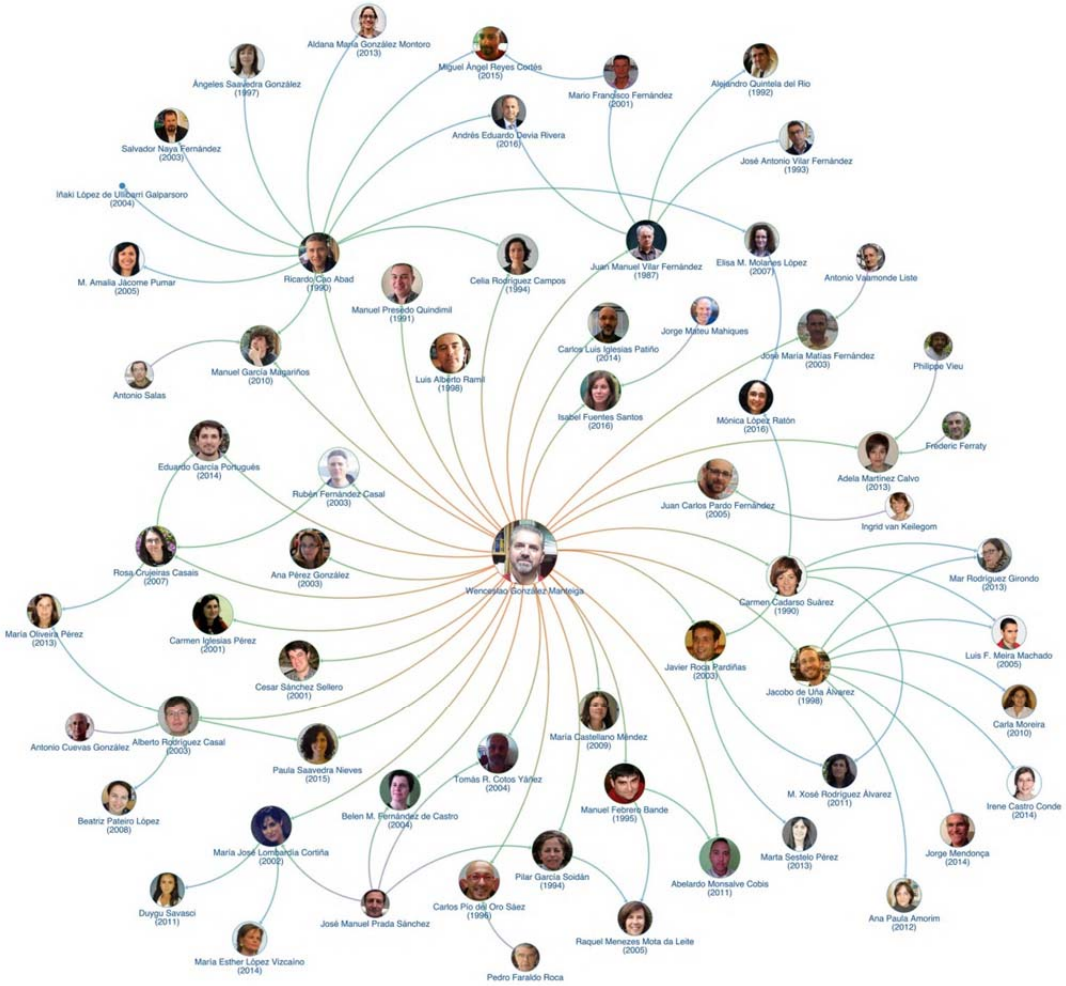


Figura 22: Póster recibido durante el Galician Seminar of Nonparametric Statistical Inference (GSNSI) los días 8 y 9 de junio 2016. A fecha de hoy, se completa la lista con: Statistical inference in quantile regression models - 2017 - Mercedes Conde Amboage y Nonparametric inference on point processes with covariates - 2018 - María Isabel Borrajo García.

Figura 23: Galician Seminar of Nonparametric Statistical Inference (GSNSI) Junio 2016

Este fenómeno de crecimiento de la investigación en Estadística, que se ha dado en Galicia, es también una tendencia concordante y común con lo que ocurre actualmente a nivel mundial. Se están creando nuevas titulaciones en "Big Data" o en "Data Science" en las Universidades públicas y/o privadas de nuestro país o en las empresas del sector privado. A nivel mundial se ha creado, muy recientemente, el "International Data Science in Schools Project (IDSSP)" (http://www.idssp.org/), con la colaboración de numerosas sociedades científicas de mucha relevancia y relacionadas con la Estadística y la Computación, con el objeto de ofrecer cursos a estudiantes y a profesores que quieran perfeccionar sus habilidades en la enseñanza del análisis de datos.

La irrupción del "Big Data" está generando un gran debate a nivel mundial, sobre qué papel deben desempeñar la Estadística y las Ciencias de la Computación en la enseñanza para la formación de profesionales del futuro en ciencia de datos. En la web mencionada, se puede encontrar una referencia de obligada lectura sobre estos aspectos en un provocativo artículo publicado por el reciente premio "Carl Friedrich Gauss", en la edición del 2018, David Donoho. Dicha publicación aparece en el "Journal of Computational and Graphical Statistics", el año pasado, y recoge el motivo de su discurso "50 years of Data Science" en la conferencia "John Tukey centennial Conference" (49).

He experimentado una gran satisfacción al recibir el pasado año la distinción de “IMS Fellow” por parte del “Institute of Mathematical Statistics”. Soy consciente de la dificultad que entraña la obtención de tal distinción para un estadístico europeo y especialmente para un español. No obstante, he de decir que para mí lo más importante fueron las palabras asociadas a tal distinción, palabras que se mencionan en el documento del que se me hizo entrega, y que tienen mucho que ver con el servicio que he prestado a la Comunidad Gallega, (Figura 24). Siempre me he sentido especialmente orgulloso de llevar a Galicia a todos los diversos foros en los que he participado.

August 11, 2017 erg@imstat.org
Wenceslao González-Manteiga Named IMS Fellow

Wenceslao González-Manteiga, Professor of Statistics and Operation Research, University of Santiago de Compostela, Spain, has been named Fellow of the Institute of Mathematical Statistics (IMS). An induction ceremony took place July 31 at the Joint Statistics Meeting in Baltimore, Maryland.

Dr. González-Manteiga received the award for influential contributions to the theory of nonparametric inference, the bootstrap and functional data analysis, and for outstanding service to the community, especially in Galicia and Spain.

Figura 24: Jon Wellner, presidente de la IMS entregando la placa, como fellow de la IMS, a Wenceslao González-Manteiga en el evento del “Joint Meeting of Statistics” celebrado en Baltimore en julio de 2017.

Una breve conclusión.

El fenómeno del “Big Data” representa para todos los estadísticos, y para los matemáticos en general, una gran oportunidad para posicionar la Estadística como una herramienta potente dirigida a la adquisición del conocimiento que se puede derivar de la inmensa cantidad de información disponible en los problemas de la vida real.

Los estadísticos debemos colaborar con los profesionales de la Computación o relacionados y con los usuarios de esa información masiva. Todos somos estrictamente necesarios.

La necesidad de la manipulación estadística de la información generará en el futuro próximo una gran demanda de profesionales que necesitarán una potente especialización en análisis de datos, una gran habilidad computacional y un conocimiento añadido de otras disciplinas matemáticas como por ejemplo, la optimización.

Necesitaremos, por tanto, que en las Universidades, las nuevas generaciones de personal docente e investigador se conviertan en buenos profesores con gran capacidad en esos ámbitos y que sean capaces de transmitir conocimiento para la generación de los profesionales del futuro. Las posibilidades son enormes pero los riesgos también. De nuevo recomiendo la lectura del informe del profesor Donoho (49).

En este discurso, un estadístico matemático ha estado hablando del mucho tiempo y trabajo que costó pasar de una situación de aislamiento científico, en la Galicia de los años 70, en lo que a la Estadística se refiere, a la situación actual en la que contamos con una gran masa crítica de estadísticos de gran prestigio internacional en las tres universidades de nuestra comunidad.

Construir lleva mucho tiempo, destruir puede ser una tarea de mucha menos duración. Los políticos y gestores de los recursos para la investigación están obligados a cuidar los logros conseguidos por los científicos y deben incrementar los recursos económicos en la investigación, para que de esta forma podamos estar a la altura de la gran exigencia del conocimiento que se requiere al tratar de abordar los nuevos retos en el futuro, y en especial en el campo del “Big Data”.

Finalizo reiterando mi agradecimiento a la Academia. Espero estar a la altura científica que la posición de académico en la RAGC exige e igualmente me gustaría que ésta pudiera contar con toda mi colaboración futura.

Muchas gracias.

Referencias bibliográficas.

(1) Scott, D.W., Tapia, R.A. y Thompson, J.R. (1977). Kernel density estimation. J. Nonlinear Analysis, Theory, Methods and Applications 1. 339-372.

(2) González Manteiga, W. (1982). Construcción axiomática, consistencias y distribuciones asintóticas de estimadores no paramétricos para funciones de densidad y de regresión. Tesis Doctoral. Universidad de Santiago de Compostela.

(3) Cristóbal, J.A., Faraldo Roca, P. y González Manteiga, W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. Annals of Statistics. 15(2). 603-609.

(4) Gómez Villegas, M.A. y De Mora Charles, M.S. (2018). Historia de la Probabilidad y de la Estadística. Madrid. UNED.

(5) Efron, B. (1979). Boostrap Methods: Another Look at the Jacknife. Annals of Statistics. 7. 1-26.

(6) Diaconis, P. y Efron, B. (1983). Métodos estadísticos intensivos por ordenador. Investigación y Ciencia. Julio 1983. Nº 82.

(7) Hardle, W. (1990). Applied Nonparametric Regression. Econometric Society Monographs.

(8) Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman Hall.

(9) Kaplan, E.L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association. 53. 457-481.

(10) Cox, D.R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B. 34(2). 187-220.

(11) Hastie, T.J. y Tibshirani, R.J. (1990). Generalized Additive Models. Chapman Hall.

(12) Vilar Fernández, J.M. (1987). Estimación no paramétrica de la función de densidad y la función de predicción en series de tiempo. Tesis doctoral. Universidad de Santiago de Compostela.

(13) Cadarso Suárez, C.M. (1990). Nuevos aportes a la regresión paramétrica y no paramétrica con datos censurados. Tesis doctoral. Universidad de Santiago de Compostela.

(14) Iglesias Pérez, M.C. (2001). Estimación de la función de distribución condicional con censura y truncamiento. Tesis doctoral. Universidad de Santiago de Compostela.

(15) Sánchez Sellero, C.A. (2001). Inferencia estadística con datos con censura y/o truncamiento. Tesis doctoral. Universidad de Santiago de Compostela.

(16) De Uña Álvarez, J. (1998). Inferencia estadística en modelos de censura proporcional. Tesis doctoral. Universidad de Santiago de Compostela.

(17) Cao Abad, R. (1990). Aplicaciones y nuevos resultados del método Bootstrap en la estimación no paramétrica de curvas. Tesis doctoral. Universidad de Santiago de Compostela.

(18) García Soidan, M.P. (1994). Generalización de los desarrollos de Edgeworth. Aplicación a la estimación de la función de distribución. Tesis doctoral. Universidad de Santiago de Compostela.

(19) Rodríguez Campos, M.C. (1994). Inferencia estadística en modelos condicionales con respuesta de tipo discreto usando estimación no paramétrica de curvas. Tesis doctoral. Universidad de Santiago de Compostela.

(20) Presedo Quindimil, M.A. (1991). Un análisis mediante la función de influencia de la estimación no paramétrica de la función de regresión. Tesis doctoral. Universidad de Santiago de Compostela.

(21) Del Oro Sáez, C.P. (1996). Análisis de observaciones influyentes en regresión no paramétrica. Tesis doctoral. Universidad de Santiago de Compostela.

(22) Fernández Casal, R. (2003). Geoestadística espacio-temporal. Modelos flexibles de variogramas anisotrópicos no separables. Tesis doctoral. Universidad de Santiago de Compostela.

(23) Febrero Bande, M. (1995). Modelización para la predicción en series de tiempo: Aspectos computacionales con nuevas aportaciones y aplicaciones. Tesis doctoral. Universidad de Santiago de Compostela.

(24) Roca Pardiñas, J. (2003). Aportaciones a la Inferencia no Paramétrica en modelos aditivos generalizados y extensiones. Aplicaciones en medioambiente y salud. Tesis doctoral. Universidad de Santiago de Compostela.

(25) Cotos Yañez, T. (2004). Predicción puntual y por intervalos con modelos parcialmente lineales y aditivos de regresión. Aspectos computacionales y aplicación a datos medioambientales. Tesis doctoral. Universidad de Santiago de Compostela.

(26) Fernández de Castro, B.M. (2004). Modelos de predicción con redes neuronales y modelos funcionales: Una aplicación a un problema medioambiental. Tesis doctoral. Universidad de Santiago de Compostela.

(27) García Jurado, I., González Manteiga, W., Prada Sánchez, J.M., Febrero Bande, M. y Cao Abad, R. (1995). Predicting Using Box-Jenkins, Nonparametric and Bootstrap Techniques. Technometrics. 37 (3), 303-310.

(28) Ramil Novo, L.A. (1998). Aportaciones a la teoría de contrastes de bondad de ajuste para modelos de regresión polinómicos. Tesis doctoral. Universidad de Santiago de Compostela.

(29) Pérez González, A. (2003). Aportaciones a la inferencia no paramétrica de modelos de regresión con respuesta faltante. Tesis doctoral. Universidad de Santiago de Compostela.

(30) Pardo Fernández, J.C. (2005). Tests in nonparametric regression based on the error distribution. Tesis doctoral. Universidad de Santiago de Compostela

(31) Conde Amboage, M. (2017). Statistical inference in quantile regression models. Tesis doctoral. Universidad de Santiago de Compostela.

(32) Lombardía Cortiña, M.J. (2002). El Bootstrap en la estimación de la función de distribución en poblaciones finitas. Tesis doctoral. Universidad de Santiago de Compostela.

(33) Iglesias Patiño, C.L. (2014). Estimación no paramétrica de la regresión en el contexto de poblaciones finitas. Tesis doctoral. Universidad de Santiago de Compostela.

(34) García Magariños, M. (2009). Nonparametric inference for classification with high dimensional genetic data. Tesis doctoral. Universidad de Santiago de Compostela.

(35) Matías Fernández, J.M. (2003). Redes neuronales para regresión y clasificación. Nuevos algoritmos y aplicaciones. Tesis doctoral. Universidad de Santiago de Compostela.

(36) Monsalve Cobis, A.E. (2011). Tests de bondad de ajuste para modelos de tipo de interés. Tesis doctoral. Universidad de Santiago de Compostela.

(37) Castellano Méndez, M. (2010). Modelización estadística con redes neuronales. Aplicación a la hidrología, aerobiología y modelización de procesos. Tesis doctoral. Universidad de A Coruña.

(38) Martínez Calvo, A. (2013). Estimates and Bootstrap calibration for functional regression with scalar response. Tesis doctoral. Universidad de Santiago de Compostela.

(39) Crujeiras Casais, R.M. (2007). Contribution to spectral spatial statistics. Tesis doctoral. Universidad de Santiago de Compostela.

(40) Fuentes Santos, I. (2015). Nonparametric inference for first order characteristics od spatial and spatio-temporal point processes. Tesis doctoral. Universidad de Santiago de Compostela.

(41) Borrajo García, M.I. (2018). Nonparametric inference on point processes with covariates. Tesis doctoral. Universidad de Santiago de Compostela.

(42) Rodríguez Casal, A. (2003). Estimación de conjuntos y sus fronteras: un enfoque geométrico. Tesis doctoral. Universidad de Santiago de Compostela.

(43) Saavedra Nieves, P. (2015). Nonparametric data-driven methods for set estimation. Tesis doctoral. Universidad de Santiago de Compostela.

(44) García Portugués, E. (2014). Nonparametric inference with directional and linear data. Tesis doctoral. Universidad de Santiago de Compostela.

(45) Gasser, T., Muller, H.G., Kohler, W., Molinari, L. y Prader, A. (1984). Nonparametric regression analysis of growth curves. Annals of Statistics. 12 (1). 210-229

(46) Peña, D. (2014). Big Data and Statistics: Trend or Change. Boletín de Estadística e Investigación Operativa. 30. 313-324.

(47) Cao Abad, R. (2017). Ingenuas reflexiones de un estadístico en la era del Big Data. Boletín de Estadística e Investigación Operativa.33 (3). 295-321.

(48) Franke, B., Plante, J.F., Roscher, R., Lee, E.A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M.M., Grosse, R. Hendricks, D. y Reid, N. (2016). Statistical Inference, Learning and Models in Big Data. International Statistical Review. 84 (3). 371-389.

(49) Donoho, D. (2017). 50 Years of Data Science. Journal of Computational and Graphical Statistics. 26 (4). 745-766.