

Colaboración

Software

Actividades da SGAPEIO

Novas do IGE

Traballos de Estatística e IO

Coñecéndonos

Estatística recreativa

Editorial

ESTHER LÓPEZ VIZCAÍNO – PRESIDENTA DA SGAPEIO.

Estimados socios e socias,

Neste editorial toca a despedida de seis compañeiras e un compañeiro que durante os últimos catro anos formaron parte do Consello Executivo (CE) da SGAPEIO. Tamén é hora de dar a benvinda aos novos membros que se integraron en novembro de 2017. Como presidenta da SGAPEIO no período 2014-2017, só me vén á cabeza unha palabra: agradecemento. Agradezo o traballo de todos os membros do CE deste último período, que estiveron achegando ideas e traballando para levar a cabo todas as actividades que a SGAPEIO propuxo. Así que, dende aquí, moitas grazas pola vosa constante colaboración! Como en toda despedida, toca facer balance daquilo que se fixo, pero tamén, e moi importante, daquilo que, dende o meu punto de vista, queda por facer e que considero interesante para o futuro da nosa sociedade. Considero que nos últimos anos se fixo un importante esforzo para que os profesores de ESO/Bacharelato volvesen confiar na SGAPEIO e nas súas actividades. Existen datos concretos que o avalan, como pode ser a participación nos Concursos da *Incubadora de Sondaxes e Experimentos* e a asistencia ás distintas Xornadas de formación que se organizaron en colaboración con AGAPEMA. Estas son dúas actividades que se consolidaron nestes últimos anos e penso que deberían seguir manténdose no futuro.

Outro feito que merece a pena destacar é a colaboración, que parece xa consolidada, da SGAPEIO coa Sociedade Portuguesa de Estatística (SPE). Celebráronse tres Encontros de Biometría en colaboración coa SPE, o último deles en Santiago de Compostela no ano 2016 e agora estase a organizar un novo que terá lugar en Aveiro no ano 2018. Ademais, está en marcha a iniciativa de firmar un convenio de colaboración coa SPE, que formalice, entre outros aspectos, as actividades que realizamos conxuntamente.

Os tempos de cambio e a sociedade dixital na que estamos inmersos fixo, tamén, que a SGAPEIO incrementase a súa presenza nas redes, e así agora estamos en Twitter e LinkedIn e temos o boletín dixital INFORMEST, que publicamos trimestralmente. Animamos a todos os socios e socias a seguir participando nas redes sociais e a achegar colaboracións para o noso boletín.

Aproveitando que neste ano cumprimos 25 anos e que xa cumpría unha actualización, fixemos unha renovación do noso logotipo e da imaxe institucional. Trátase dunha nova imaxe e dun novo logo moderno e adaptado aos tempos que corren, aínda que conservando as nosas marcas de identidade como sociedade: $\#$ e as mesmas cores.

A pesar de todas estas actividades que se fixeron, notamos un certo estancamento nas incorporacións de xente nova, dato que nos provoca certo grao de preocupación. Considero fundamental revitalizar a participación da xente nova na Sociedade. Tamén penso que a SGAPEIO debería ser quen de manter unha oferta anual de un ou dous cursos nos que os seus socios e socias estivesen interesados e, tamén, aumentar as relacións co mundo empresarial.

Como se pode observar existen aínda cousas por desenvolver nos vindeiros anos e, de seguro, que estímulos e ganas para levalos a bo porto non han faltar. Doulles a benvinda aos compañeiros que formarán o novo CE, que, como xa saben, teñen a colaboración de todos nós para que a SGAPEIO siga sendo un punto de referencia na divulgación e na promoción da Estatística e a Investigación Operativa.

Dirección:

Paula Raña Míguez

Comunicación coa SGAPEIO:

www.sgapeio.es
secretaria@sgapeio.es
boletin@sgapeio.es
Twitter @sgapeio
LinkedIn Sgapeio

Depósito Legal:

LU-191-1995 - I.S.S.N.:1695-7083



Sociedade Galega para a Promoción da
Estatística e da Investigación de Operacións

Facultade de Matemáticas

Rúa Lope Gómez de Marzoa s/n Campus Vida-
15782 Santiago de Compostela (A Coruña)
<http://www.sgapeio.es/>



Estatística no cine

por José María Sorando Muzás

José María Sorando é Catedrático de Matemáticas de Educación Secundaria. Ademais da docencia, adícase á didáctica e divulgación das Matemáticas. Autor de máis de 85 publicacións, 6 exposicións, máis dun cento de conferencias en congresos e diferentes cursos e autor da web “Matemáticas en tu mundo”. Gañador, entre outros, do Premio Santillana de Experiencias docentes 2010.

A presenza das matemáticas no cine e nos seriais de televisión é desigual no tocante aos contidos. Adoita trátarse de simples cálculos aritméticos, sendo frecuentes os erros, como na vida mesma. Cando se quere mostrar o talento do protagonista, a miúdo soe ser presentado como unha “calculadora humana”, identificando de forma equívoca o pensamento matemático co cálculo mental. Para outros propósitos fóra do común recórrase aos números primos, á serie de Fibonacci ou ao número π , talvez por seren os conceptos matemáticos máis sofisticados ao alcance do cidadán medio; pode tratarse de claves esotéricas e incluso como medio de comunicación extraterrestre. O cálculo de probabilidades é citado en situacións de risco. Hai pouca xeometría, case sempre en pizarras escolares. O cálculo diferencial, tamén en pizarras, pero agora de científicos, como elemento exclusivo dunha elite intelectual. E a estatística no cine? Vexamos.

Onde menos se conta, salta a lebre

De forma inesperada, no serial infantil *Bob Esponxa* encontramos unha alusión aos tests de hipóteses. No episodio *Patricio intelixente* (8º – tempada 4. 2005) Bob Esponxa e Patricio, xogan nos campos de medusas, até que Patricio cae por un acantilado. Bob Esponxa baixa correndo, atopa a cabeza que se desprende do corpo do seu amigo e vólvela a colocar. Algo se activa no cerebro do simple Patricio, facendo que se volva intelixente e fale dunha forma descoñecida nel.

- *Veña Patricio, fagamos algo divertido!*
- *¡Oh! Que che parece se preparamos unhas estatísticas acerca da observación dos fenómenos e expresamos hipóteses sobre eles?*
- *Estaba pensando máis en saltar á corda, carreiras cunha soa perna ou xogar ao agáchate, agáchate!...*

Sirva este anecdótico fragmento para constatar un tópico e prexuízo moi común: as matemáticas como algo aburrido, que só lle pode interesar a alguén moi listo,

o que o fai estraño perante os demais. É tarefa de docentes e divulgadores espertar nos nosos concidadáns esa durmida “ollada matemática” que descubra como as matemáticas están na vida cotiá; como poden axudarnos a comprender mellor a realidade e a tomar decisións adecuadas; como, en definitiva, son necesarias para todos. A estatística debería ocupar un lugar de privilexio neste descubrimento, por ser a parte das matemáticas que traballa de forma máis directa con esa realidade.

Bromas e confusións

Hai no cine máis citas estatísticas espurias, buscando o humor ou como reflexo dunha ignorancia.

Broméase coa media aritmética en *Los Simpson*. No episodio *Jacques, El Rompecorazones* (nº 9 – tempada 1. 1990), Marge vai xogar aos birlos. Para xogar debe poñer o calzado axeitado. O encargado pregúntalle:

- *Que número calza, por favor?*
- *O 43.*
- *43? –di o encargado, que non encontra zapatos do número solicitado–. Apañárase cun 44 e un 42?*



En *La última noche* (Spike Lee. 2002), sorprendentemente fábase de percentís:

- *Sabes que? Estás no percentil 62, aí é onde estás.*
- *E iso que?*
- *O percentil 62? É a túa clasificación. Todos os solteiros de Nova York competimos polas mesmas mulleres, os heteros, claro.*
- *E eu estou no percentil 62? Son mellor que o 62 % dos solteiros de Nova York? Pero peor que o 38 %?*
- *Que o 37, non existe o 100.*
- *E como calculaches o 62?*
- *Entendelo é toda unha ciencia.*

- Entendelo ten algo de ciencia? Ben, parece moi científico... dime, cal é a túa clasificación?
- Queres sabelo? Estou xusto no percentil 99.

Tamén cre saber estatística Jean de Florette, un notario convertido en agricultor, protagonista de *El manantial de las colinas* (Claude Berry. 1986):

Jean: *Non temos, propiamente dita, unha estación de choivas. Segundo as estatísticas dos últimos 50 anos establecidas polo Observatorio de Marsella: abril, 5 días de choiva; maio, 6 días; xuño, 4 días; xullo, 2 días; agosto, 3 e setembro, 6... de media.*

Chega unha longa seca e Jean novamente dialoga sobre o tema:

Jean: *Creo que esta noite pode chover.*

Ugolin: *Por que? Nótao na súa reuma?*

Jean: *Non, afortunadamente non teño reuma, pero o ceo, que nos tiña que ter dado 6 días de choiva en maio, só nos deu 3. Desde primeiros de xuño tiña que ter chovido 2 veces. Débenos, polo tanto, 5 días de choiva. E esa débeda na contabilidade celeste, ten que ser pagada nas próximas 48 horas.*

Despois, as choivas seguen sen chegar e berra exaltado:

Jean: *Pero onde se viu unha seca ininterrompida de 36 días?... Aquí é matematicamente imposible!*

En *El Apartamento* (Billy Wilder. 1960), o seguinte diálogo ofrécenos unha das típicas chanzas a propósito da media:



- Estiven lendo unha estatística sobre accidentes e enfermidades. O cidadán neioirquino entre os 20 e os 50 ten dous arrefriados e medio por ano.
- Que gran responsabilidade a miña!

– Por que?

– Porque como eu non me arrefrí, para que non fallen as estatísticas outro infeliz ten que ter cinco arrefriados.

En serio e en broma, a mesma confusión: considerar a información estatística como leis deterministas de obrigado cumprimento, a curto prazo e pequena escala. A realidade non está lonxe. Abonda escoitar declaracións públicas e conversas privadas.

Citas con fundamento

Pero tamén hai producións, moi poucas, que introducen a estatística nas súas tramas, de forma correcta e con sentido. En teleseries, con amplísima vantaxe, destaca *Numb3rs*. Nos 119 episodios das súas seis tempadas (2005-2010) pode haber tantas matemáticas ou máis que no resto da ficción televisiva e, desde logo, máis solventes. Charlie, o protagonista matemático, axúdalle ao seu irmán, axente do FBI, na resolución de casos criminais. Recibe datos de cada caso. Para a súa interpretación a miúdo recorre ao estudo estatístico e soe concluír facendo predicións en termos de probabilidade. Vémoslle utilizar, por exemplo: a regresión á media, a minería de datos, a curiosa Lei de Benford, a análise estatística lingüística, etc. A única obxección, desde o noso punto de vista, é que eses recursos matemáticos, aínda que pertinentes en cada caso, son moi pouco explicados e talvez o telespectador medio os vexa novamente como fórmulas máxicas dun tipo estraño.

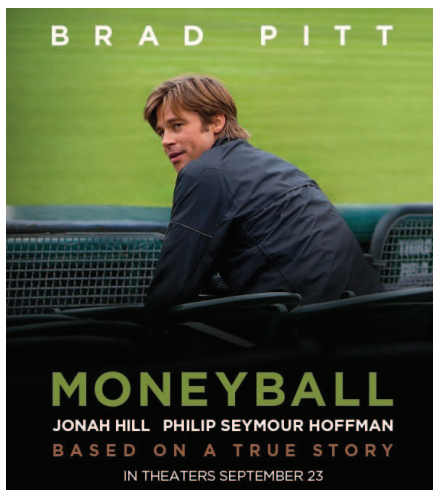
No cine, tan só coñezo dous títulos estimables: *Moneyball* e *Ciudad Mágica*.

Estatística descriptiva

Moneyball: Rompiendo las reglas (Bernett Miller. 2011) conta a historia real de Billy Beane, *manager* dos *Athletics* de Oakland, os “A’s”, un equipo de béisbol modesto. Beane rebélase contra a desigualdade económica que contamina a competición deportiva. Coñece a Peter Brand, un economista que desenvolveu un minucioso método para a análise de xogadores.

Os clásicos observadores do béisbol recomendan fichaxes en función do estilo ou calidade individual, pero tamén do palmarés e da imaxe, con termos tan subxectivos como “virtuosismo”, “ter boa estrela” ou “estar de sorte”. Brand obvia eses aspectos persoais para, no canto deles, cuantificar a aportación de cada xogador ao seu equipo, elaborando un número índice que ten en conta todas as súas estatísticas de pasadas temporadas. Propón “non comprar xogadores, senón vitorias”. Para el, cada xogador vale “non o seu prezo no mercado, senón a súa aportación ao xogo”. E aínda engade: “Valóranse aparencias e prexuízos. As matemáticas están por riba de todo iso”.

Beane aplica os criterios de Brand, para escándalo xeral e elabora un padrón repleto de xogadores infravalorados segundo o índice de Brand, descartados por outros equipos. Como di, “un corral de patíños feos”. Na temporada 2002 os “A’s” encadean unha xeira de 20 vitorias consecutivas, batendo o récord histórico, e chega aos *play-offs* polo título. Conseguirano dous anos máis tarde.



Entre os eloxios recibidos por esta película escoitei: “*Moneyball* ten o mérito de ser unha película sobre o menos cinematográfico, sobre estatísticas” (Cadea Ser 04-02-12). Quen así se expresou non apreciaba o fondo épico da historia: o intento de vencerlle ao diñeiro aplicando a intelixencia. Nesta nova versión do combate de David contra Goliat, as armas do débil son matemáticas.

E non obstante, a estatística xa estaba presente no béisbol de forma máis exhaustiva que en ningún outro deporte, con rexistros completos temporada a temporada. O béisbol préstase especialmente a isto dadas as súas especiais características. Cada xogador cumpre unha función moi específica no seu posto e non cabe a versatilidade doutros deportes de equipo onde os xogadores se dispersan polo terreo e poden cumprir funcións varias. O novo era pasar de considerar a calidade individual a valorar o potencial de eficiencia no equipo.

Esas técnicas reciben o nome de *sabermetría*. Toma o seu nome da SABR ou Sociedade para a Investigación do Béisbol Americano e cítase tamén en *Numb3rs* e en *Los Simpson*, onde vemos a Lisa, adestradora de béisbol, rodeada de libros de estatística.

Hoxe en día, co *big data* e a aplicación de modelos matemáticos, xa é usual falar no deporte de planificación de índices de eficiencia e de optimización de recursos.

Estatística inferencial

A trama de *Ciudad Mágica* (William A. Wellman. 1947) xira arredor da representatividade das mostras.

Rip Smith dirixe unha empresa que se dedica aos sondeos de opinión. Ao bordo da bancarrota, descobre casualmente que os resultados que se obteñen na pequena cidade de Grandview son unha réplica exacta dos obtidos para todos os Estados Unidos. É o que repetidamente chama o “milagre matemático”, a salvación para a súa empresa, que lle permitirá ofrecer resultados válidos cun custo e rapidez moi vantaxosos fronte á competencia. Rip planea trasladarse cos seus colaboradores a Grandview e establecerse baixo unha falsa identidade de vendedores de seguros, para así conseguir información de forma permanente sen que os veciños sospeiten cal é a súa finalidade.

Rip intégrase rapidamente na vida da cidade, chegando a ser moi apreciado e namórase dunha xornalista. Esta descobre o propósito de Rip e publícao no diario local. A nova corre por todo o país e Grandview pasa a ser coñecida como a cidade modelo da nación. Nada volverá ser o mesmo. A cobiza duns e outros traerán o rápido ascenso e a estrepitosa caída da cidade. Como non podía faltar nun filme de Hollywood en posguerra, hai un final de esperanza. É esta “unha película das de antes”: en branco e negro, onde os personaxes son elegantes e cortesés, a honradez triunfa e o amor redime. Na película dise que en Grandview as porcentaxes de homes e mulleres, de brancos e negros, de republicanos e demócratas, etc. son réplicas exactas das que se dan para o conxunto da nación. Desta forma xustifícase aceptablemente a súa fiel representatividade como mostra.

Cando a cámara se detén na relación de datos podemos comprobar que as porcentaxes que atenden a cada criterio suman 100 (que menos!... anque no cine non sempre se coidan eses detalles numéricos). Os cálculos son realizados á man, sobre un gran cuadrante, por un contable á vella usanza, con manguitos e viseira. Pensemos que en 1947 aínda faltaban algúns anos para a extensión das calculadoras mecánicas de oficina e 25 anos para as calculadoras electrónicas de peto.

As predicións da película obtéñense como porcentaxes fixas (estimadores puntuais) e non mediante intervalos de confianza, como é habitual nos estudos demoscópicos.

Por último, expónse como a consciencia de estar sendo observados, nese caso cunha presión mediática excesiva, altera o comportamento dos cidadáns de Grandview e o “milagre matemático” comeza a fallar.

Usos e abusos políticos

Moitas situacións dramáticas que reflicte o cine parten de conflitos e dilemas éticos. Algúns deles teñen que ver coa utilización das estatísticas polo poder e a súa repercusión social. Unha das máis vulgares manipulacións prodúcese cando o cuestionario é indutor das respostas. Na serie da BBC *Yes, Prime Minister* (1986-1987) dialogan así dous asesores do primeiro ministro:

- O Partido realizou unha enquisa e vese que os votantes están a favor de restablecer o Servizo Militar.
- Temos outra enquisa que demostra que os votantes están en contra do Servizo.
- Non poden estar a favor e en contra.
- Claro que si, é que non che fixeron nunca enquisas?
- Si, pero non unha enquisa política. Non sabería que contestar.
- Ti xa sabes como funciona iso. Visítate unha señorita moi guapa e queres quedar ben, non queres facer o ridículo.
- Non.
- E comeza a facerche preguntas: Preocúpalle o gran número de mozos que están sen traballo?
- Si.
- E preocúpalle tamén o aumento da delincuencia xuvenil?
- Si.
- Parécelle que falta disciplina nas escolas públicas?
- Si.
- Pensa que os mozos aceptarían un maior nivel de liderazgo e de autoridade? Que son capaces de aceptar un reto?
- Si.
- Estaría a favor de restablecer o Servizo Militar?
- Si.
- Está claro que si, despois do que me dixeches non podes contestar que non. Eles non publican as primeiras preguntas e só mencionan a derradeira.
- É iso o que fan realmente?
- Home, os que son honrados talvez non, pero non hai moitos deses. E a señorita tería conseguido o resultado oposto se tivese querido.
- Como?
- Preocúpalle o perigo dunha guerra?
- Si.
- Preocúpalle a carreira de armamentos?
- Si.
- Atopa perigoso darlles armas aos mozos e aprenderlles a matar?
- Si.
- Parécelle mal obrigarlle á xente a usar armas se eles non queren?
- Si.
- Vostede oporíase ao restablecemento do Servizo Militar?
- Si.

- Velaí, é así como se encargan as enquisas para o Ministerio de Defensa.

O anterior diálogo sería xocoso se non fose polo seu parecido coas enquisas que realizan algúns medios de comunicación politicamente belixerantes entre os seus propios lectores (neste caso engádese a falta de representatividade da mostra), concluíndo cunha alegre xeneralización dos resultados.

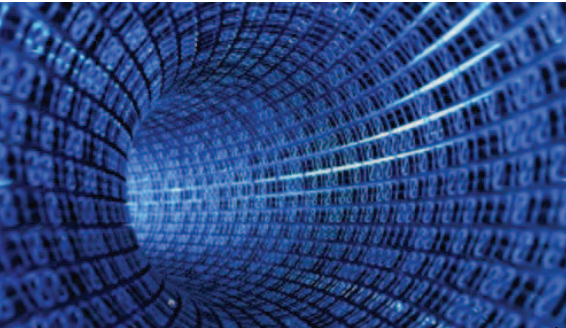
Menos evidente, pero por isto máis efectiva, é a argucia política de actuar preferentemente sobre os aspectos que máis inflúen nas estatísticas oficiais, en detrimento doutros talvez máis importantes. Na aclamada serie *The Wire* (2002-2008), as autoridades locais de Baltimore necesitan estatísticas que reflectan unha redución da delincuencia nas rúas. Tamén necesitan estatísticas que reflectan melloras nos resultados das probas estatais realizadas nas escolas públicas e no absentismo escolar. Unhas e outras serán argumentos ao seu favor na próxima campaña electoral.

As autoridades policiais e escolares tamén necesitan esas melloras estatísticas, por imperativo dos seus superiores políticos e por conveniencia: para seguir a súa carreira persoal no corpo, nun caso, e para asegurar a subvención escolar, noutro.

Eses intereses comúns levan a que se emitan ordes perversas, tanto no mundo escolar como no policial. Na escola, as clases dedícanse a que os alumnos memoricen as preguntas e respostas do exame, unha vez asegurado que serán as mesmas do ano anterior; e son utilizados conserxes para recoller nas rúas aos alumnos absentistas e levalos á clase só até que se teña alcanzado a data de cómputo de ausencias, a partir da cal se lles permite ausentarse para que non molesten. Na policía, interrómponse investigacións de longo percorrido que levarían até os narcos xefes para dedicarse ás redadas cos pequenos traficantes e así aumentar o número de detencións. Incluso se retrasa o descubrimento dun bo número de cadáveres, vítimas dos narcos, ao feito de que computen como asasinatos no mandato dun alcalde ou outro.

As estatísticas reflecten a realidade para permitírnos, coñecéndoa, actuar sobre ela. En *The Wire*, é sobre as estatísticas que se actúa, con desprezo da realidade que se vive e da ética. Os números son neutrais, pero o uso que deles se fai non o é. Por iso, a formación estatística e lóxica na etapa de educación obrigatoria é unha necesidade imprescindible para o libre exercicio da cidadanía. Cando se carece desa formación, como escribira Darrell Huff: “a maxia dos números anula o sentido común”.

Unha vez máis, paradoxicamente, o cine é unha grande ficción que nos permite reflexionar sobre a realidade.



Xornadas de R

M^a José Ginzo Villamayor

IV Xornada de Usuarios de R en Galicia

O día 19 de outubro, tivo lugar na facultade de **Matemáticas** da **Universidade de Santiago de Compostela** (USC) a IV Xornada de Usuarios de R en Galicia. Estivo organizada pola **Oficina de Software Libre** do **CIXUG**, e patrocinada pola **AMTEGA** (**Xunta de Galicia**). Co obxectivo de promocionar e difundir o coñecemento libre da linguaxe estatística R e mostrar as súas aplicacións, leváronse a cabo relatorios durante todo o día e mais dous talleres. A xornada pretende ser un punto de encontro para todas aquelas persoas interesadas en intercambiar as súas experiencias e atopar colaboracións do resto da comunidade. Pódese dicir que esta IV Xornada, é como se fose a V, xa que nos anos 2013 e 2015, celebráronse a *I e II Xornada de Usuarios de R en Galicia* e no ano 2014, organizáronse as *VI Jornadas de usuarios de R* (a nivel nacional).

O comité organizador estivo formado por Roberto Martín Souto coordinador da Oficina de Software Libre do CIXUG (consorcio das tres Universidades Galegas), Rafael Rodríguez Gayoso, Concello de Santiago de Compostela e M^a José Ginzo Villamayor do Departamento de Estatística, Análise Matemática e Optimización da USC. O comité científico estivo constituído por M^a José Ginzo Villamayor e Miguel Ángel Rodríguez Muíños (da Dirección Xeral de Saúde Pública da Consellería de Sanidade). A mesa de apertura estivo formada por M^a Elena Vázquez Cendón (Decana da Facultade de Matemáticas - USC), Rosa M^a Crujeiras Casais (Vicerreitora de Comunicación e Coordinación da USC), Miguel Ángel Rodríguez Muíños e Roberto Martín Souto.

O programa da Xornada contou con doce relatorios seguido por dous talleres paralelos.

- “*Utilización do software R na unión de rexistros administrativos*” impartida por Noa Veiguela Fernández do Instituto Galego de Estatística (IGE).

A finalidade desta charla é presentar o procedemento do que se serve o IGE para lograr a unión dos rexistros administrativos cando non poden contar co DNI da persoa. Nesas casos empregan variables comúns en dous rexistros, coma o nome e apelidos da persoa, o seu sexo, data de nacemento e concello de residencia, que soen figurar en case todas as bases de datos que se manexan no IGE. Para tratar de casar variables, que poden diferir para unha mesma persoa dun rexistro a outro, empregáronse as librerías *stringdist* e *fuzzyjoin* do software libre de programación R.

- “*Cálculo da adxudicación dos premios de fin de carreira (2017)*” impartida por Marcos Fernández Arias, Pablo Espido Noya (coautor), ambos da

Xunta de Galicia.

Mostrou brevemente como se programa e calcula mediante R a adxudicación dos premios fin de carreira universitarios da Comunidade Autónoma de Galicia.

- “*Xeración de música determinista con R*” impartida por Miguel Ángel Rodríguez Muíños (Consellería de Sanidade), Alejandro Rodríguez Antolín (coautor).

Explicación do proxecto **DETMUS** e comentario de aspectos técnicos de programación con R e as súas utilidades fóra dos “usos tradicionais” do mencionado software.

- “*Novo entorno para Big Data con R: sparklyr*” impartida por Aurora Baluja González do Complejo Hospitalario Universitario de Santiago (CHUS), Javier López Cacheiro (coautor) do Centro de Supercomputación de Galicia (CESGA).

A tecnoloxía que permite a lectura e escritura de macrodatos (“*Big Data*”) experimentou un espectacular avance nos últimos anos, coa aparición de plataformas *open source*. O paquete *sparklyr*, lanzado en 2016 por RStudio, permite unha comunicación directa coa API de Spark para “dataframes”, empregando a sintaxe de R.

- “*R en paralelo e HPC*” impartida por Diego Mainera Díaz do Cesga, Andrés Gómez (coautor - CESGA), Aurelio Rodríguez (coautor - CESGA), Santiago Cerviño (coautor - Instituto Español de Oceanografía), Javier López Cacheiro (coautor - CESGA).

O emprego de R en paralelo permite unha redución considerable do tempo de execución da maioría dos programas. Se combinamos isto cos recursos dun centro HPC, podemos obter un gran rendemento, tanto na execución de *scripts* propios, como no emprego de certas aplicacións que fan uso de R.

- “*R como pedra angular de proxectos de Data Science*” impartida por Daniel Prieto Rodríguez (Minsait by Indra).

Os proxectos de *Data Science* divídense en varias fases que van desde os estádios iniciais de deseño da modelización ata a posta en produción. Neste relatorio mostraron como R convive co resto de elementos do ecosistema deste tipo de proxectos en cada unha das fases. Noutras palabras, amosouse unha pequena aproximación ao uso aplicado de R dentro do ecosistema de ferramentas, na realidade empresarial dunha industria de distribución.

- “*Predición de Series Temporais en Datasets Multidimensionais de Situacións de Negocio Mediante Paralelización Agradable en R Server*” impartida por Antonio Soto Rodríguez de SolidQ.

Aplicación de distintos modelos predictivos de series temporais para predecir a un horizonte finito un determinado valor facendo n-particións incoexas altamente paralelizables do conxunto de datos dispoñible co uso de R Server, e xerando unha serie temporal única para cada partición. Especificamente predise o número de recollidas de usuarios de taxi na cidade de Nova York, por orixe, destino, tipo de pago, duración do traxecto e outras dimensións, pero o problema é facilmente trasladado a outras situacións de negocio.

- “*Análise da incidencia da leucemia granulocítica empregando a estimación non paramétrica de conxuntos de nivel*” impartida por Paula Saavedra Nieves da Universidade de Vigo (UVigo).

Un problema fundamental en epidemioloxía é determinar se as zonas nas que se concentran os casos dunha enfermidade se corresponden coas rexións máis poboadas ou, polo contrario, existen áreas de risco nas que a súa incidencia é maior. A estimación non paramétrica de conxuntos de nivel é unha ferramenta estatística útil para abordar este tipo de cuestións. A partir dun conxunto de datos reais, analizaron a incidencia da leucemia granulocítica en Lancashire e Greater Manchester estimando conxuntos de nivel con R.

- “*RStudio como ferramenta para desenvolvemento de material docente interactivo e multimedia*” impartida por Alejandro Quintela del Río da Universidade da Coruña (UDC).

Ademais de servir para a realización de programas e material científico reproducible, mostrárase a versatilidade de Rstudio para a creación de material docente (clases e apuntamentos). Neles poden incluírse elementos multimedia (vídeos, animacións, táboas e gráficos interactivos), código html, código latex, e facilitar o proceso de aprendizaxe creando tutoriais e preguntas. Ao poder converterse directamente nunha páxina web, os apuntamentos serán accesibles a través de internet aos alumnos, ademais de poder publicarse en *pdf* ou *word*.

- “*Debuxando curvas ROC en R*” impartida por Arís Fanjul Hevia (USC).

O obxectivo desta charla foi explicar brevemente que librerías existen para tratar as curvas ROC e

cales son as súas principais funcións. Falouse dos distintos métodos que hai para estimar estas curvas, como atopar os puntos de corte óptimos para a toma de decisións, como calcular rexións de confianza e como comparar dúas ou máis destas curvas. Ademais, expuxéronse problemas con datos reais como exemplo de aplicación destas técnicas.

- “*A ecoloxía na súa revolución da cantidade masiva de datos*” impartida por Miguel Branco (UVigo).

Os estudos de biodiversidade a escala global e os cambios nas guías editoriais e planos gobernamentais, que obrigan á publicación de datos, xeraron na última década grandes fontes de información ligada á ecoloxía. Para o seu aproveitamento, creáronse librerías en R que cobren tanto os pasos de obtención, integración e meta-análise dos datos como a súa publicación. Así, librerías como “taxize” e “phyloseq” permiten a obtención de censos de organismos; outras, como “rfigshare”, a publicación de estudos de campo e outras, como “aRxiv”, a publicación de artigos, froito dos estudos.

- “*Web Scraping*” impartida por José Luis Juncal Pérez.

Extracción de datos a un ficheiro CSV, con formato definido polo cliente, de diversas webs abertas ao público.

Para rematar houbo dous talleres, un, para aqueles que se inician con R e o outro, para usuarios máis avanzados: “*Taller práctico e básico de R*” impartido por Arís Fanjul Hevia, M^a José Ginzo Villamayor (coautor) ambas da USC e o “*Taller Web Scraping*” impartido por José Luis Juncal Pérez. Os talleres, ademais de actividades da xornada, enmárcanse dentro das actividades da Aula de Matemáticas da Facultade de Matemáticas, coa que colabora estreitamente a CIXUG.

Pódese dicir que a Xornada acadou unha alta participación, con 102 persoas inscritas, así mesmo, os talleres de R e *Web Scraping*, de case 2 horas de duración, contaron con 39 e 29 asistentes, respectivamente.

Quero rematar agradecendo, no nome do Comité Organizador e Científico, a colaboración de todas as persoas (poñentes, moderadores, asistentes á xornada, *twiteros*...) que fixeron posible esta IV Xornada, e as contribucións dos patrocinadores e colaboradores. Grazas tamén á OSL pola iniciativa de organizala.



Información Adicional

Para máis información sobre a IV Xornada de Usuarios de R en Galicia véxase a páxina web: <https://www.r-users.gal/>, onde se poderán atopar as presentacións e vídeos da xornada.



Actividades da SGapeio

**XIII Congreso Galego de Estatística e Investigación Operativa.
Ferrol, 26-28 de outubro de 2017.**

por Salvador Naya e Javier Tarrío

Durante o 26, 27 e 28 celebrouse no Campus de Esteiro de Ferrol o XIII congreso da Sociedade, cun total de 110 asistentes de diferentes países como a República Dominicana, Ecuador, Colombia ou USA, e doutras comunidades como a andaluza, País Vasco, Asturias, Castela ou Valencia.



O congreso, organizado polo Departamento de Matemáticas da Universidade de A Coruña, por encargo da SGapeio, contou co apoio de diferentes grupos de investigación e outras entidades, como os grupos de investigación MODES e CITIC da Universidade de A Coruña, o Concello de Ferrol, a Deputación de A Coruña e de empresas de Ferrolterra como Navantia, Cándido Hermida e Intaf. Ademais do grande apoio recibido pola Armada para as visitas culturais á Exponav, ao Arsenal de Ferrol e a unha das fragatas F100.

No contexto científico o congreso tivo un total de catro conferencias plenarias completadas con sesións de comunicacións orais e pósteres e cunha mesa sobre control dimensional. Ademais, tamén se ofreceu a posibilidade de facer dous cursos. O primeiro, sobre “R markdown” impartido por Manuel Oviedo de la Fuente (ITMATI) e Rubén Fernández Casal (UDC), e un segundo curso, destinado a profesores de ensino medio, coordinado pola profesora Covadonga Rodríguez Moldes (IES Mugardos). Este segundo curso contou, como relatores, cos profesores Manuel Alfredo Mosquera Rodríguez e Tomás R. Cotos Yáñez, ambos da Universidade de Vigo, con Dolores Pilar García Agra (IES de Ordes) e as profesoras gañadoras da última convocatoria do concurso nacional de *Incubadora de Sondeos e Experimentos*, da SEIO, Milagros Diéguez Taboada (CPI As Revoltas, Cabana de Bergantiños) e Paula Blanco Mosquera (CPI de San Vicente. A Baña).

A primeira das conferencias, impartida polo profesor Luís Escobar, da Louisiana State University, tratou sobre fiabilidade. Nela analizáronse diferentes problemas na estimación de distribucións de tempos de vida dos compoñentes dun sistema, poñendo exemplos de aplicación a casos reais.

A segunda conferencia, impartida polo profesor da Universidade da Coruña, Ricardo Cao, estivo centrada nas aplicacións estatísticas no contexto da “Industria 4.0”, concepto que se utiliza para aludir á dixitalización das cadeas de valor industriais. Neste paradigma da chamada cuarta revolución industrial, a Estatística (e tamén a Investigación de Operacións!) está chamada a xogar un papel clave, como ferramenta indispensable para converter datos en información útil para a mellora dos procesos produtivos e para engadir valor aos produtos. Nesta conferencia percorréronse algúns problemas reais existentes en empresas do ámbito industrial, comercial e dos servizos, que foron abordados nos últimos anos polos membros do grupo de investigación MODES mediante o uso de modelos e técnicas estatísticas axeitados.

Na conferencia impartida polo profesor da Universidade de Sevilla, Emilio Carrizosa, fíxose unha revisión dalgúns dos campos das Análises de Datos e da súa visualización nos que o tradicional papel da Optimización ofrece estimulantes solucións. O conferenciante fixo fincapé no espectacular avance nos métodos da Optimización Matemática (en particular, da Programación Non Lineal en Números Enteiros) que permiten abordar exitosamente novos problemas, ou abordar problemas tradicionais de modo moito máis eficiente que hai unha década.

A conferencia de peche do congreso foi impartida pola profesora da Universidade de A Coruña, Amparo Alonso, que nos ofreceu unha visión doutros algoritmos con alta aplicabilidade no entorno do *BigData* e que van máis aló dos clásicos problemas de regresión e clasificación, ofrecendo unha perspectiva dos novos retos con datos de alta dimensión, pero desde o contexto da intelixencia artificial. Dentro destas sesións tivo lugar unha especial dedicada á estatística pública, organizada coa colaboración do IGE, que contou coa asistencia de responsables doutros institutos de estatística como o vasco ou o asturiano. Esta sesión, aproveitando que o congreso se celebrou na semana do día europeo da estatística, foi denominada “Sesión do *European Statistics Day*: A estatística pública en Galicia”.

Entre as sesións de comunicacións cómpre destacar as que presentaron os 7 investigadores mozos que expuxeron os traballos para optar aos “premios SGAPEIO ás mellores comunicacións presentadas por investigadores novos”. Os membros do xurado, formado por Juan Manuel Vilar Fernández (UDC), Jacobo de Uña Álvarez (UVigo) e José M^a Alonso Meijide (USC), tendo en conta os traballos enviados e as presentacións realizadas durante o congreso, decidiron por unanimidade conceder o premio á mellor comunicación teórica ao traballo titulado “*Characterization of point processes with covariates through the first-order intensity function*”, presentada por María Isabel Borrajo García e o premio á mellor comunicación aplicada ao traballo titulado “Un modelo de programación matemática para la selección y asignación temporal de recursos aéreos destinados a la contención de un incendio forestal”, presentada por Jorge Rodríguez Veiga. O xurado destacou a gran calidade de todos os traballos presentados.

As sesións científicas remataron coa mesa redonda sobre control dimensional na que participaron Álvaro Brage, responsable da Sección de Control Dimensional de Na-

vantia; Salvador Fábregas responsable de calidade do grupo Intaf e Irene Barros do departamento de calidade da empresa Cándido Hermida. A mesa estivo coordinada polo profesor da UDC, Javier Tarrío Saavedra.

Actos sociais do congreso

Os actos sociais consistiron nun concerto a cargo do músico Pedro Lamas acompañado por outros dous músicos da súa banda, que tivo lugar o xoves 26 ao remate das sesións orais. Completouse cunha visita ao Arsenal Militar de Ferrol, onde os congresistas puidemos visitar a Fragata Almirante Juan de Borbón, que forma parte da frota das fragatas F100 fabricadas pola empresa Navantia de Ferrol. Logo dunha comida no propio Arsenal puidemos percorrer Exponav, un dos museos máis importantes da construción naval a nivel europeo, contando con pezas únicas como o *pecio* da fragata Magdalena.

O congreso rematou cunha comida de confraternidade no restaurante do Hotel Almirante. Ao final da comida tivo lugar a actuación do actor Vicente Mohedano, que presentou un espectáculo sobre a historia dos números.





INSTITUTO GALEGO DE
ESTADÍSTICA

Novas do IGE

Novas do IGE

Novidades na información estatística publicada

ESPERANZA DE VIDA SEN LIMITACIÓN

O obxectivo desta actividade estatística é coñecer a esperanza de vida sen limitación e sen limitación grave da poboación galega, é dicir, estimar o número medio de anos que pode esperar vivir unha persoa sen limitación e sen limitación grave das actividades que realiza na vida cotiá. Os indicadores sintéticos que se calculan nesta actividade, máis alá de reflectir o impacto da mortalidade nunha poboación, recollen tamén as condicións de saúde presentes nela. No ano 2014, a esperanza de vida da poboación galega situouse en 83 anos; non obstante, a esperanza de vida sen limitación só alcanza os 61,2 anos. Ademais, tanto a esperanza de vida convencional como a esperanza de vida sen limitación grave sitúanse en valores lixeiramente máis altos para as mulleres que para os homes. No caso da esperanza de vida sen limitación, os valores en homes e en mulleres son moi similares. Máis información en: [web](#)

ENQUISA ESTRUCTURAL A FOGARES. CONDICIÓN NO TRABALLO. 2016

Esta operación estatística ten por finalidade recoller información sobre as condicións no traballo dos residentes en Galicia. Entre outros aspectos, estúdanse as características do posto de traballo, os aspectos relacionados coa seguridade e a prevención de accidentes no traballo, grao en que os ocupados poden conciliar traballo e vida familiar, a traxectoria profesional, nivel de formación e o grao de satisfacción co traballo percibido polos ocupados e ocupadas en Galicia. Desta operación estatística pódese concluír que o 68,26 % das persoas ocupadas en Galicia traballa cinco días á semana; o 20,12 %, seis días e o 6,35 %, os sete días da semana. A porcentaxe de persoas ocupadas que traballa menos de cinco días sitúase, en 2016, no 5,28 %. Por outra parte, tamén se obtén que ao 7,19 % das persoas ocupadas en Galicia resúltalles moi doado compaxinar a vida familiar e laboral; ao 55,10 % resúltalles doado; o 29,38 % recoñecen que lles é difícil compaxinar ambos os dous aspectos da vida e ao 8,33 % resúltalles moi difícil. Máis información en: [web](#)

EXPLOTACIÓN DA MOSTRA CONTINUA DE VIDAS LABORAIS. 2015

A [Mostra continua de vidas laborais \(MCVL\)](#) é un conxunto de microdatos anonimizados procedentes de diversos rexistros administrativos: da Seguridade Social, do Padrón municipal de habitantes e da Axencia Tributaria.

A MCVL constitúe unha mostra representativa de todas as persoas que durante o ano tiveron relación coa Seguridade Social, ben porque estiveron afiliadas en situación de alta laboral, ben porque percibiron unha prestación contributiva ou subsidio por desemprego, ou ben porque percibiron algún tipo de pensión contributiva da Seguridade Social. Desta actividade estatística pódese deducir que no ano 2015, 1.112.650 galegos e galegas en idade laboral (de 16 ou máis anos) mantiveron algunha relación de afiliación en alta laboral coa Seguridade Social, un 1,9 % máis que no ano 2014. Por outra parte, o 22 % dos galegos e das galegas que en 2015 estiveron afiliadas en alta laboral por conta allea percibiu un salario total menor ou igual aos 6.000 euros anuais. Entre a poboación traballadora que só estivo afiliada en alta laboral parte do ano, esta porcentaxe sobe ata o 48,5 %, mentres que para os que estiveron afiliados o ano completo é do 3,6 %. Máis información en: [web](#)

Outras novidades

XORNADA DE XEORREFERENCIACIÓN DA INFORMACIÓN ESTATÍSTICA EN GALICIA: CASOS PRÁCTICOS

O día 25/10/2017 celebrouse no Salón de Actos da Consellería de Sanidade en Santiago de Compostela unha xornada de traballo sobre a xeorreferenciación da información estatística en Galicia, organizada polo IGE e o grupo GAME-IDEGA da Universidade de Santiago de Compostela. Nesta xornada presentáronse exemplos da integración da información estatística e xeográfica, co fin de mostrar a importancia e o potencial da xeorreferenciación. A localización de unidades estatísticas permite unha avaliación máis precisa das zonas de interese e, polo tanto, unha mellor toma de decisións por parte da Administración e da empresa privada. Na xornada participaron persoal do IGE, do Instituto de Estudos do Territorio, CESGA, USC, CITIUS, Axencia de Turismo de Galicia, AMTEGA, ABANCA e a Universidade Complutense de Madrid. Máis información en: [web](#)

OBSERVATORIO GALEGO DE DINAMIZACIÓN DEMOGRÁFICA

O Observatorio Galego de Dinamización Demográfica creouse polo Decreto 104/2016, do 28 de xullo (DOG núm. 152 do 11 de agosto) modificado polo Decreto 68/2017, do 13 de xullo (DOG núm. 143, do 28 de xullo) e ten como obxectivos o coñecemento exhaustivo e permanentemente actualizado da realidade demográfica de Galicia e o asesoramento ás administracións públicas da Comunidade Autónoma de Galicia na incorporación da perspectiva demográfica. Neste Observatorio está presente o IGE cun vogal titular.

Traballos de Estatística e Investigación Operativa no SUG

Teses

Novas aportacións estatísticas ao estudo da fiabilidade de materiais

Antonio Meneses.

Directores: Salvador Naya e Javier Tarrío.

Data da defensa: 8 de setembro de 2017. UDC.

O obxectivo da presente tese de doutoramento é a aportación de novas técnicas e metodoloxías estatísticas para o estudo da fiabilidade de materiais sometidos a esforzos térmicos e de fatiga mecánica. Concretamente, introducíronse dous procedementos estatísticos de aplicación inmediata nas áreas da física aplicada e enxeñaría de materiais. O primeiro consiste nunha alternativa para realizar estudos de superposición tempo/temperatura. O obxectivo é a predición das propiedades mecánicas de materiais viscoelásticos, a unha temperatura dada, a tempos e frecuencias máis aló do rango experimental mediante o desprazamento da primeira derivada das curvas de propiedades mecánicas obtidas a outras temperaturas (probos aceleradas), definindo para isto, por vez primeira, o principio TTS desde o paradigma estatístico. O segundo procedemento representa un novo método flexible, baseado no axuste de regresión lineal de efectos mixtos, que ten como fin estimar a distribución de tempos de fallo de materiais (principalmente metais) sometidos a esforzos de fatiga mecánica. Ambas as metodoloxías representan aportacións de utilidade e competitivas para a resolución de problemas de caracterización de materiais e predición das súas propiedades físicas durante a súa vida útil (fiabilidade), sempre a partir da modelización de datos de degradación obtidos por técnicas experimentais de análise térmico e mecánico.

Bargaining Models with Asymmetric Agents

Alfredo Valencia Toledo.

Director: Juan José Vidal Puga.

Data da defensa: 26 de outubro de 2017. UVigo.

No presente traballo, estudamos modelos da teoría de xogos con axentes asimétricos, sendo esta simetría definida dende dúas diferentes perspectivas. Por unha parte, a asimetría pola propia natureza dos axentes. En particular, consideramos un axente particular diferente do resto, un único inquilino fronte a un grupo máis numeroso de arrendadores de terreos. Por outra parte, a asimetría debida a unha compoñente temporal. En particular, algúns axentes poden actuar e/ou desenvolverse de forma simultánea mentres que outros o fan de forma secuencial, como é o caso das actividades dentro dun proxecto de grande envergadura.

As características coas que se define a asimetría dos axentes, en ambas as perspectivas, xogan un papel importante no momento de definir os modelos. Consideramos como motivación dúas situacións reais: os conflitos de terras e os atrasos de actividades.

Na primeira situación tomamos como exemplo os casos de adquisición de terreos para a explotación ou extracción de minerais por unha empresa mineira en Perú e para a localización da base militar da BRILAT (Brigada Lixeira AeroTransportada) en Vilaboa (Pontevedra). Nestes casos pódese observar un proceso de negociación entre unha empresa ou institución e un grupo de propietarios ou comunidades. Estas situacións estúdanse como un problema de conflito de intereses desde a perspectiva dos xogos cooperativos e os non-cooperativos.

Na segunda situación consideramos como exemplo o custo que producen os atrasos das actividades que forman parte de proxectos de grande envergadura. Esta situación estúdase como un problema de asignación de custos entre os axentes responsables.

No enfoque cooperativo buscamos regras de reparto que verifiquen propiedades tales como consistencia, monotónía, estabilidade, dualidade e estándar para dous. No enfoque non-cooperativo, estudamos como sería o reparto final baixo unhas condicións de negociación razoables.

Traballos Fin de Máster

Traballos do Máster Interuniversitario en Técnicas Estatísticas.

Detección de anomalías en espazos de alta dimensión

Víctor Alonso Macías.

Director: Javier Roca Pardiñas.

Titor: Bruno Fernández Castro, Gradient.

USC, modalidade prácticas.

Neste traballo trátase de comprender a problemática da alta dimensión nos métodos de detección de anomalías co obxectivo de realizar unha comparativa entre varios algoritmos. Esta levarase a cabo investigando as súas prestacións, características, limitacións e comportamento sobre diferentes conxuntos de datos de proba, logrando, desta maneira, validar un método para a súa aplicación directa á análise de datos xerados por máquinas.

Optimización baixo incerteza en redes de gas

Ana Belén Buide Carballosa.

Director: Julio González Díaz.

Titor: Diego Rodríguez Martínez, ITMATI.

USC, modalidade prácticas.

Este traballo xorde pola necesidade de validar un algoritmo de programación estocástica empregado por unha empresa de transporte e regasificación de gas para a resolución de problemas de optimización no contexto das redes de gas. No primeiro capítulo lévese a cabo unha revisión das nocións teóricas da programación matemática. No segundo capítulo introdúcese o problema de optimización determinista en redes de gas e o algoritmo de programación linear sucesiva de dúas etapas implementado dende o ITMATI. Logo, introdúcese a programación estocástica e o problema de programación estocástica no contexto das redes de transporte de gas, os cales xorden na toma de decisión relativa á planificación da construción de novas infraestruturas en redes de gas. A formulación deste problema é o punto de partida para a validación do algoritmo implementado sobre unha batería de problemas asociados ás redes de gas galega e nacional. Finalmente, amósanse os resultados da validación realizada durante as prácticas desenvolvidas nas instalacións do ITMATI e que dan mostra da utilidade do algoritmo implementado na práctica.

Text Analytics para Procesado Semántico

María Calvo Torres.

Director: Rubén Fernández Casal.

Titor: Juan Ramón González, Oesía Networks.

UVigo, modalidade prácticas.

Este traballo trata de introducir as principais técnicas da minería de textos e as súas diversas aplicacións. Para isto, no primeiro capítulo, comézase a explicar que é a minería de textos e cales son os seus principais usos actualmente. Deseguido, explícanse varios xeitos de como obter os datos e logo ilústrase o preprocesado dos mesmos cun exemplo. No terceiro capítulo, faise unha análise exploratoria dos datos e, a continuación, explícanse os métodos máis utilizados para a clasificación de documentos, adéstranse os modelos e válídanse. Por último, amósase unha aplicación *Shiny* do clasificador de textos que se construíu seguindo paso a paso as explicacións do traballo. A súa finalidade é unha mellor comprensión do tema coa axuda dun método visual interactivo. Para a realización deste traballo, recibíuse a axuda da empresa Oesía, experta no ámbito da minería de textos.

Estudo de asociación de variantes xenéticas coa miocardiopatía hipertrófica

Sandra Carballido Regueiro.

Director: Ignacio López de Ullibarri Galparsoro.

Titor: Lorenzo Monserrat Iglesias, Health in Code.

UDC, modalidade prácticas.

A busca de asociacións entre variantes xenéticas e certas enfermidades é un tema de grande importancia na actualidade, sobre todo, a medida que se avanza no coñecemento de técnicas de secuenciación do ADN. Este tipo de técnicas proporcionan unha gran cantidade de información que precisa ser analizada con metodoloxías estatísticas especialmente adaptadas ao volume de datos co que se traballa. Con tal propósito, é habitual considerar estudos de tipo caso-control que, mediante o uso dos modelos de regresión apropiados, permitan identificar que variantes aumentan o risco de padecer a enfermidade considerada. O obxectivo do proxecto é describir como funcionan estas técnicas estatísticas e amosar a súa aplicación en datos reais no ámbito da cardiología. Os datos utilizados proceden de pacientes diagnosticados de diversas cardiopatías conxénitas, entre elas, a miocardiopatía hipertrófica, enfermidade na que se centran os resultados desta análise. Esta patoloxía é unha das doenzas cardíacas máis comúns e é, na maioría de ocasións, hereditaria, polo que resulta interesante tratar de identificar a compoñente xenética que desencadea a súa manifestación.

Problemas de rutas de vehículos estocásticos

Juan Carlos Gonçalves Dosantos.
Directora: Balbina Casas Méndez.
USC, modalidade académica.

Na actualidade, existe unha mellora significativa na recollida de todo tipo de datos. Isto provocou que a maioría dos problemas tratados na investigación operativa pasen a ser máis realistas e dinámicos. Entre eles destacamos os problemas de rutas de vehículos, que incorporaron novas restricións aplicadas a datos en tempo real e parámetros estocásticos. Un caso concreto desta situación é considerar as demandas dos clientes como elementos aleatorios. É por isto que, para comprender o modelado deste tipo de problemas, é necesaria unha Introducción á Programación Estocástica, así como, algunha linguaxe de programación que a soporte. Nós faremos mención do formato SMPS, *Stochastic Mathematical Programming System*, para modelar un problema de programación. Para así despois resolver estes problemas mediante algún *solver* dispoñible nun conocido servidor de optimización: NEOS. Pero, dado o elevado custo computacional que requiren este tipo de problemas, a nosa intención é utilizar algún método heurístico que nos axude a obter solucións factibles para os nosos problemas en tempos razoables. Para isto, imos considerar unha adaptación da heurística do caso determinista, o coñecido algoritmo de aforros de Clarke e Wright, ao noso contexto estocástico. Amosaremos o seu funcionamento nun problema real tomado do ámbito da loxística agrícola e programamos en R ambos os algoritmos (determinista e estocástico). Finalmente, deseñamos unha interface gráfica en R que facilite a introdución de datos, a chamada de algoritmos, a visualización de resultados e a realización de análise de post-optimalidade de maneira rápida.

Regresión segmentada

Víctor González Carro.
Directora: Rosa María Crujeiras Casais.
USC, modalidade académica.

Este traballo analiza en profundidade os modelos de regresión segmentada no caso en que os dous anacos nos que se divide a función de regresión sexan rectas. Concretamente, adaptaranse os métodos clásicos de estimación en regresión lineal simple e compararanse nun estudo de simulación e unha aplicación a datos reais no ámbito da ecoloxía. A maiores, engádese un capítulo de ampliación do estudo do modelo no que se presentan técnicas de inferencia baseadas no modelo de regresión segmentada e outras alternativas paramétricas a este modelo.

Optimización baixo incerteza en plantas de proceso industrial

Daniel López García.
Director: Julio González Díaz.
Titor: Patricio Reyes Valenzuela, ITMATI.
USC, modalidade prácticas.

O obxectivo deste traballo é presentar os conceptos básicos de optimización con incerteza e a construción do algoritmo *progressive hedging* que está pensado de forma específica para esta clase de problemas. Este algoritmo baséase en movementos na dirección do gradiente do dual, por este motivo ao inicio do traballo presentamos conceptos básicos de programación matemática como as condicións de KKT e o dual lagrangiano. No *progressive hedging* hai dous parámetros que ten que fixar o usuario. Son o paso da actualización das variables do dual (que denotamos por π) e o peso da penalización cuadrática (que denotamos por r). Ao final do traballo facemos un breve estudo do calibrado destes parámetros. Este calibrado foi o obxectivo das prácticas asociadas a este Traballo Fin de Mestrado que se realizaron na Unidade Mixta de Investigación Repsol-Itmati.

Modelos cuantís autorregresivos

Silvia Novo Díaz.
Director: César Sánchez Sello.
USC, modalidade académica.

As vantaxes que o enfoque cuantil aporta na regresión, pola análise máis robusta e completa que proporciona, levaron a distintos autores a estender esta metodoloxía ao contexto dos modelos de series temporais, como é o caso dos procesos autorregresivos. Neste traballo revisamos distintos procedementos para a obtención de intervalos de predición no contexto do modelo autorregresivo clásico, e incorporaremos unha proposta de novo método para a obtención destes intervalos realizando a estimación mediante mínimas desviacións absolutas e mediante *bootstrap*. Ademais, para aproveitar as vantaxes que a estimación cuantil aporta baixo heterocedasticidade, estudamos un proceso autorregresivo estacionario máis xeral que permite a dependencia dos coeficientes respecto do cuantil, constituíndo unha alternativa na análise de series temporais con dinámicas asimétricas ou con persistencias locais. Para este modelo estúdase un modo de identificar a súa orde, así como para estimar os seus parámetros. Finalmente, ilústranse estes procedementos mediante a súa aplicación á análise da serie da taxa de variación trimestral do desemprego en España, cuxo estudo queda incompleto empregando as técnicas clásicas.

O Problema de Coloración de Grafos

Sergio Pena Seijas.

Directora: Silvia Lorenzo Freire.

USC, modalidade académica.

Dentro do amplo campo da Teoría de Grafos, neste traballo realizamos unha revisión do Problema de Coloración de Grafos. En concreto, centrámonos na coloración dos vértices dun grafo. Por iso, presentamos no capítulo inicial unha breve introdución ao problema dende a súa aparición a través do famoso Teorema das Catro Cores, e revisando os avances e desvantaxes que xurdiron a medida que se profundou neste problema. No segundo capítulo, facemos unha revisión daqueles algoritmos máis relevantes asociados ao Problema de Coloración de Grafos, analizando cales son os pros e contras de cada un deles. Finalmente, introducimos un conxunto de aplicacións derivadas da análise do Problema de Coloración de Grafos, presentes no noso día a día e que fan ver dito problema dende un punto de vista máis práctico.

Analise de datos dos consumos de computación e os seus resultados empregando plataformas de *Big Data*

Elena Pernas Goy.

Director: Manuel Febrero Bande.

Titor: Andrés Gómez Tato, CESGA.

USC, modalidade prácticas.

Esta memoria recolle as tarefas realizadas durante as prácticas no CESGA. As bases de datos empregadas nas mesmas pertencen ao sistema *Finisterrae* e son consideradas *Big Data*, polo que se empregaron ferramentas adecuadas, neste caso *Apache Spark*. *Spark* é unha plataforma que permite implementar código en distintas linguaxes de programación das cales se escolleu *Python*. Nun primeiro instante tómase contacto co programa realizando tarefas sinxelas de filtrado e cálculo de proporcións e medidas estatísticas para, posteriormente, pasar á parte central das prácticas que consiste na análise do sistema de colas. A análise do sistema de colas divídese en dúas partes. Nunha delas, trátase de predicir os tempos de espera das colas mediante series de tempo. Na outra, trátase de realizar unha simulación do propio sistema para poder obter máis información del e do comportamento das colas. Para finalizar, propóñense diferentes camiños polos que se podería continuar traballando.

Os valores esquecidos na mostraxe de poboacións finitas. Técnicas de imputación

Gabriel Román Radío.

Director: Antonio Vaamonde Liste.

UVigo, modalidade académica.

Os valores esquecidos supoñen un obstáculo - actualmente non ben resolto - cando é necesaria a resposta completa - porque o require a estimación a facer ou porque o esixen as técnicas estatísticas a empregar na análise. As técnicas de imputación permiten asignar valores aos casos e variables con valores omitidos. Non obstante, as técnicas de imputación non son neutrais en relación cos resultados obtidos e a elección inadecuada da mesma podería ocasionar alteracións na distribución dos datos. Este traballo expón as diversas técnicas de imputación existentes, a avaliación das súas vantaxes e inconvenientes e a súa aplicación con distintos paquetes da linguaxe de programación estatística R.

Traballos Fin de Grao

Traballos do Grao en Matemáticas.

Flexibilizando a regresión lineal: regresión local lineal

María Alonso Pena.

Directora: Rosa M. Crujeiras Casais.

Os modelos de regresión paramétricos lineais poden resultar insuficientes para explicar a relación entre variables de xeito efectivo. Non obstante, para intentar explicar relacións non lineais de forma máis simple, pódense utilizar modelos non paramétricos que, en lugar de supoñer que o modelo ten unha forma determinada, só asumen unha hipótese de suavidade sobre a función de regresión. En concreto, neste traballo preséntase o método de regresión local lineal, que consiste en axustar modelos lineais unicamente nas observacións que estean dentro dun entorno dun punto fixado. Ao realizar este axuste en cada punto, obtense a estimación da función de regresión. Ademais de recoller a definición formal do estimador e as súas propiedades, no traballo móstranse distintos métodos de selección do parámetro de suavizado. O funcionamento do estimador analízase mediante un estudo de simulación no que se comparan algúns dos máis importantes métodos de selección de ventá. Por último, ilústrase o contido do traballo con datos reais, pertencendo estes ao ámbito da análise térmica.

Modelización estatística espacial: kriging

Almudena Barreiro Carrillo.

Director: Manuel Febrero Bande.

Ao longo deste traballo descubrimos os aspectos fundamentais dunha das ramas da estatística espacial: a Xeoestatística. Nos primeiros parágrafos facemos unha introdución amigable deste concepto. Tamén definimos dous exemplos que nos acompañarán durante todo o percorrido. O primeiro exemplo consiste nunha mostra de miligramos de níquel que foron medidos en certos puntos de Galicia. O segundo exemplo radica en datos recollidos durante as eleccións do 20 de decembro do 2015. O segundo tema que tratamos son os *variogramas*, sen dúbida, peza fundamental en xeoestatística. Descubriremos como modelalos e axustalos e cal é o seu papel á hora de definir unha estrutura espacial. En terceiro e último lugar aparecen as predicións. Explicaremos en que consiste a predición coñecida co nome de *kriging* e as súas tres variantes: *kriging* simple, *kriging* ordinario e *kriging* universal. Trataremos de facer a mellor predición posible

cos datos dos dous exemplos antes mencionados. Tamén abordaremos brevemente algúns dos comandos do programa R que empregamos con frecuencia nos nosos exemplos e que son comúns no estudo da xeoestatística.

Técnicas de clasificación

Fernando Castro Prado.

Director: Wenceslao González Manteiga.

Este traballo ten como obxectivo presentar detalladamente algunhas técnicas relacionadas coa análise discriminante. O seu contido distribúese en seis capítulos. No primeiro introdúcense algúns conceptos xerais relativos á optimidade, para logo ilustralos, no caso de poboacións gaussianas, presentando as denominadas regras lineares e cuadráticas no segundo capítulo. No terceiro, tamén baixo a hipótese de normalidade, explícase como contrastar de maneira satisfactoria a igualdade de matrices de covarianzas e vectores de medias entre dous grupos, rutinas previas á clasificación. Nos capítulos 4 e 5 amósase como, cando a dimensión do vector poboacional é alta en relación ao tamaño mostral, a maioría dos procedementos da análise multivariante clásica xa non son aplicables. Neste caso cómpre utilizar unha metodoloxía máis elaborada, tanto para os tests previos como para a análise discriminante linear. No sexto e último capítulo aplícanse estas técnicas a bases de datos de expresión xénica para distintos tipos de cancro e un exemplo simulado, demostrando a súa utilidade práctica. Finalmente, inclúese un apéndice con todo o código empregado.

Un modelo de Teoría de Colas: as redes abertas de Jackson

Laura Davila Pena.

Directora: Balbina Casas Méndez.

A *teoría de colas* é unha disciplina que se engloba dentro da Investigación Operativa e cuxo obxectivo é o estudo e análise de situacións nas que un cliente demanda un servizo, de tal forma que dito servizo non pode ser satisfeito instantaneamente, polo cal se provocan esperas. Os problemas de colas aparecen en moitas situacións da vida real. Poden xerarse, por exemplo, a partir das chamadas efectuadas a unha central telefónica, polas mostras recibidas nun laboratorio clínico para seren analizadas, polos vehículos que circulan polas diferentes glorietas dunha cidade, ou polos clientes que empregan os probadores dunha tenda de roupa e posteriormente se dirixen á cai-

xa para realizaren o pago. No presente traballo comézase revisando algunhas nocións básicas sobre os procesos estocásticos, necesarios para comprender a *teoría de colas*, que abarca todo o segundo capítulo. Despois de expoñer os aspectos máis relevantes dun sistema de colas e explicar algúns modelos sinxelos, chegamos ao caso particular das *redes de colas*, centrando a nosa atención nas redes abertas de Jackson. Preséntase neste momento un dos resultados máis importantes das redes: o *teorema de Jackson*. Como complemento ao marco teórico, finalízase o traballo cunha sección de tres aplicacións reais, incluíndo un caso aínda en fase de estudo do Complexo Hospitalario Universitario de Santiago (CHUS). Introdúcese o código de R correspondente á base de datos para esta última aplicación.

cadrados. Ademais expónse algúns dos problemas que xorden coas estimacións dos coeficientes do modelo en presenza, por exemplo, de multicolinealidad. As técnicas de regularización teñen por obxectivo dar solución a estes problemas a través da penalización da verosimilitude. Aínda que existen outras técnicas de regresión regularizada, a memoria centrárase na regularización *Ridge*, que usa unha penalización L2. No segundo capítulo da memoria, preséntase o método e descríbense algunhas das súas propiedades. Ademais tamén se propoñen diversos procedementos para elixir o parámetro óptimo para realizar a regresión regularizada. Finalmente, nun terceiro e último capítulo, ilústrase un exemplo cuns datos sobre o prezo medio da vivenda nos distintos barrios de Boston usando a regresión *Ridge*.

Técnicas de agrupamento de datos

Andrea Fernández López.
Directora: Rosa M. Crujeiras Casais.

As técnicas de agrupamento (ou técnicas *cluster*) son un conxunto de ferramentas multivariantes utilizadas para clasificar un conxunto de individuos segundo os valores que determinan certas variables sobre eles, de xeito que os individuos pertencentes a un mesmo grupo sexan o máis similares entre si posible. Existen dous tipos de agrupamento, xerárquico, cuxos algoritmos máis importantes son o de vinculación simple, o de vinculación completa e o método de *Ward*, e o non xerárquico, cuxo algoritmo máis importante é o de *k-medias*. Ao longo deste traballo, recordaremos algunhas nocións básicas estudadas en diferentes materias do Grao e que son necesarias para a comprensión do mesmo, como a distancia, a media, a varianza... e explicaremos os dous tipos de agrupamento, así como os algoritmos mencionados. Ademais, aplicaremos todo isto sobre uns datos reais procedentes da Área de Calidade e Mellora de Procedementos (ACMP) da Universidade de Santiago de Compostela para ver como se agruparían os individuos segundo os distintos algoritmos.

Métodos de regresión lineal regularizada: Regresión lineal Ridge

Alejandro Ferreiro Méndez.
Directora: Beatriz Pateiro López.

Neste traballo realízase unha revisión do modelo de *regresión lineal regularizada Ridge*, co que se pretende solucionar algúns problemas que aparecen na regresión lineal xeneral con estimación por mínimos cadrados. No primeiro capítulo, preséntase o modelo de regresión lineal xeral coa súa estimación usual polo método de mínimos

Os segredos da normal

Diego Frade Amil.
Director: Alberto Rodríguez Casal.

Ao longo do Grao úsase continuamente a distribución normal, pero son moitos os aspectos desta que non chegan a verse ou non con suficiente detalle e que, polo tanto, se descoñecen. Neste traballo, dividido en tres partes, trátase de dar resposta a outras tantas preguntas: como xurdiu a normal, como se comproba se un conxunto de datos é normal ou non e como se constrúen datos normais. Así, no primeiro capítulo abórdanse os aspectos históricos da normal, desde os seus antecedentes ata os desenvolvementos que realizaron Abraham de Moivre e Carl Friedrich Gauss (e simultaneamente a este último, Robert Adrain), así como as aportacións de Pierre Simon de Laplace. Ademais, neste capítulo trátase o TCL e a normal multivariante. Saber se un conxunto de datos segue unha distribución normal ou non é unha cuestión interesante, que constitúe o tema central da segunda parte. Preséntanse varios conxuntos de datos univariantes e analízanse dous tests de normalidade: *Shapiro-Wilk* e *Kolmogorov-Smirnov-Lilliefors*. Realízase un estudo de simulación dos mesmos (calibrado e análise de potencia) e aplícanse aos datos. Ademais, trátase o caso bivariente cun par de conxuntos de datos e os correspondentes tests de normalidade para este caso. E no derradeiro apartado deste traballo, bótase man da simulación para dar catro métodos de xeración de variables normais a partir de variables uniformes, de sinxela programación no software estatístico R.

Introdución á programación non lineal

Brais González Rodríguez.
Director: Julio González Díaz.

Esta memoria comeza cun breve repaso dos conceptos básicos da programación matemática, co fin de axudar ao lector que non estea familiarizado con ditos conceptos. Posteriormente, trátanse con profundidade os problemas de programación non lineal, nos cales se buscan condicións suficientes e/ou necesarias que permitan garantir optimalidade local e, baixo certos supostos, optimalidade global. O anterior ilústrase con diversos exemplos. Tamén se explica brevemente a dualidade lagrangiana e a súa utilidade práctica. Por último, desenvólvese un algoritmo específico (chamado RLT) para problemas de programación polinómica. En primeiro lugar, preséntanse as súas bases teóricas, probando resultados de converxencia e ilustrando paso a paso o algoritmo con exemplos concretos. Posteriormente, utilízase unha implementación do algoritmo para obter resultados que nos permitan elaborar gráficas, co fin de comprobar a efectividade do algoritmo.

Análise Discriminante Non Paramétrica

Pilar Marful Rocha.
Director: Wenceslao González Manteiga.

Este traballo establece regras discriminantes que permiten clasificar un individuo en dous ou máis grupos ben diferenciados. Dentro do contexto paramétrico, explícase o procedemento para obter una das regras máis clásicas, que é a *regra discriminante lineal de Fisher*. A pesar de que nos atopemos nun contexto paramétrico, pode darse que algún parámetro sexa descoñecido polo que o temos que obter en base a mostras dos diferentes grupos, o que leva á construción dunha regra de clasificación estimada. Dentro do contexto non paramétrico, explícase algunha das regras que se poden obter, entre as cales se pode distinguir entre regras mediante estimación da función de densidade e estimación da función de regresión. Destácase a regra dos veciños máis próximos e o *regresograma*. Nos dous contextos presentados, cuantifícase o erro cometido dado que o obxectivo da análise discriminante é a clasificación dun individuo nun grupo, co menor erro posible. Por último, faise un breve estudo no software R, aplicado a unha base de datos proposta no ano 1978 por D. Harrison e D. L. Rubinfield, "Aloxamento en Boston", que se usará, esencialmente, para obter o prezo que tería que pagar un novo individuo por unha vivenda en Boston. Baseándose no erro de clasificación, chegaremos á conclusión de que a mellor regra para facer a clasificación é unha regra non paramétrica,

é dicir, unha regra que non fai suposicións erróneas dos datos, senón que se obtén en base aos propios datos.

Introdución á construción de intervalos de confianza *bootstrap*

Marta Martínez Romay.
Director: Jose Manuel Prada Sánchez.

Este traballo céntrase na teoría *bootstrap* e nos diferentes tipos de intervalos de confianza construídos mediante a mesma. O método *bootstrap* é unha das técnicas de remuestreo máis utilizadas na actualidade grazas ao avance da tecnoloxía computacional nas últimas dúas décadas. Este método consiste en xerar un elevado número de mostras a partir da miña mostra inicial coa finalidade de obter información acerca dun parámetro poboacional que nos interese. Explícase a construción dos intervalos de confianza tanto clásicos como *bootstrap* e considérase o caso particular da media poboacional obtendo os intervalos para ela. Tamén se realiza un estudo de simulación para analizar o comportamento dos distintos tipos de intervalos baixo certos escenarios controlados e, por último, dáse unha aplicación a un caso real implementando o código en R.

Teoría de xogos: do Hex ao Teorema de punto fixo de Brouwer

Damián Pallas Carrillo.
Directores: Julio González Díaz e Ángel Manuel González Rueda.

John Nash, célebre matemático especializado na Teoría de xogos, presentou en 1951 o concepto de equilibrio de Nash para xogos non cooperativos e tamén o Teorema para a súa existencia (que máis adiante recibiría o seu nome e polo cal gañou o Premio Nobel de Economía en 1994). Para probalo, precisa dalgún teorema de punto fixo como pode ser o Teorema do punto fixo de Brouwer. Independentemente disto, John Nash tamén inventou o *xogo do Hex*, un xogo de taboleiro que non pode acabar nunca en empate. Sorprendentemente, este feito é equivalente ao Teorema do punto fixo de Brouwer, o pilar fundamental no que se apoia a demostración do Teorema de Nash. Por dicilo dalgún xeito, Nash, sen pretendelo, foi quen de enunciar o seu Teorema e de proporcionar as ferramentas necesarias para a súa demostración. Neste traballo comezamos facendo unha breve introdución á Teoría de xogos non cooperativos, resaltando a importancia dos equilibrios de Nash. Logo, coa finalidade de probar o Teorema de Nash, enunciámos e probamos resultados necesarios como o Lema de Sperner ou o Teorema do punto fixo de Brouwer.

A continuación, presentamos o xogo do Hex orixinal (de 2 xogadores), demostramos que non pode rematar en empate e probamos a súa equivalencia co Teorema de Brouwer. Finalmente, como curiosidade, analizamos o xogo do Hex n -dimensional e obtemos del un algoritmo para a aproximación de puntos fixos.

Problemas de decisión multiobxectivo con varios axentes

Alba Pérez Fernández.

Directora: Balbina Casas Méndez.

En moitos ámbitos da vida real xorden certos problemas á hora de decidir entre varias alternativas. Se a iso lle engadimos a posibilidade de que sexan varios os axentes que deben realizar tal decisión, influíndose así uns e outros no resultado da súa elección, e que ditas decisións se tomen en base a diferentes obxectivos que se desexa conseguir simultaneamente, estamos ante o que se coñece como un “xogo multiobxectivo”. É a teoría de xogos multiobxectivo quen proporciona as solucións a tales xogos. Segundo o marco no que nos atopemos, faise unha distinción entre xogos cooperativos multiobxectivo e xogos non cooperativos multiobxectivo, en función de se os xogadores poden ou non establecer acordos vinculantes co propósito de mellorar os seus beneficios. Neste traballo faise unha pequena introdución aos xogos multiobxectivo, mostrando os principais conceptos de solución e algúns resultados básicos, onde é necesario o dominio de ferramentas informáticas como LINDO ou R, e plásmanse varias aplicacións da teoría de xogos multiobxectivo na vida real, sobre todo, no ámbito empresarial.

Estimación do erro típico en contrastes de hipóteses

Natalia Pérez Veiga.

Director: César Sánchez Sellero.

Neste traballo fin de grao estúdanse dúas formas básicas de estimar o erro típico de fronte ao contraste de hipótese: utilizando as suposicións da hipótese nula ou sen utilizar a devandita información. Compáranse os procedementos baseados nas dúas estratexias mediante argumentos teóricos e estudos de simulación. O estudo máis detallado céntrase no contraste da media e no contraste dunha e dúas proporcións. Atopamos que no caso da media, como cabía esperar, é preferible empregar o estatístico T de Student, onde o erro típico se estima baixo a alternativa. Con todo, se se emprega a aproximación normal, é preferible estimar o erro típico baixo a hipótese nula. No contraste dunha proporción, os nosos resultados mostran que a recomendación que figura nos libros de texto a favor de estimar o erro típico utilizando a

hipótese nula é plausible, a corrección por continuidade non mellora o funcionamento do test, senón que mesmo o empeora, e, en calquera caso, o test exacto debería ser recomendado como procedemento xeral. Para o contraste de dúas proporcións, de novo resulta preferible empregar a hipótese nula para a estimación do erro típico. Tamén se esboza o problema no contexto máis xeral de contraste dun parámetro baixo estimación de máxima verosimilitude, o cal permite abordar outros exemplos notables como contrastes sobre o parámetro dunha distribución de Poisson, dunha distribución exponencial ou os contrastes de significación dos coeficientes de regresión.

Metodos estadísticos en clasificación de imaxes

Ricardo Recarey Fernández.

Directora: Beatriz Pateiro López.

A aprendizaxe estatística é unha rama da estatística que, a pesar de existir desde a metade do século pasado, se atopa nun momento de rigurosa actualidade. Esta defínese como o conxunto de técnicas para construír funcións predictivas baseándose nos datos da mostra. Estas técnicas, segundo a información da que se dispoña dos datos, poden ser supervisadas ou non supervisadas. Neste traballo centraremos no caso supervisado, que ten como obxectivo a regresión e a clasificación. Concretamente, o noso problema a tratar será a clasificación de imaxes, onde o noso conxunto de adestramento serán imaxes correctamente clasificadas, sendo o obxectivo final a construción dun clasificador que, dada unha nova imaxe, lle asigne a clase que lle corresponda. Así, ao longo deste TFG farase unha revisión, tanto práctica como teórica, das técnicas máis relevantes deste campo, dende as máis simples e intuitivas, como os K -veciños máis próximos, ata as técnicas máis complexas, como as máquinas de vectores de soporte. Comezase facendo unha breve contextualización do problema, para logo introducir os conceptos clave que irán aparecendo ao longo do texto. Posteriormente presentarase as técnicas a tratar definindo cada unha delas formalmente, estudando as súas propiedades matemáticas e computacionais, xunto coas súas aplicacións e usos. Unha vez analizados os métodos, introduciremos ferramentas para medir o rendemento dos mesmos, para estudar así cales son os máis axeitados para abordar un problema concreto. Para finalizar, analizaranse os resultados obtidos por distintas redes neuronais en problemas de clasificación de imaxes levado a cabo no supercomputador *Finisterrae II* no CESGA.

C_{urvas} ROC

Xaime Suárez Blanco.
Director: Wenceslao González Manteiga.

Neste traballo expónse unha introdución ao concepto de curvas ROC (*Receiver Operating Characteristic*) en tests con resultados continuos, a cal nos proporciona unha ferramenta estatística para a avaliación da precisión dos tests e por tanto son útiles na comparación destes. No primeiro capítulo introdúcense os diferentes tipos de test e os conceptos de sensibilidade e especificidade, os cales son básicos no desenvolvemento teórico da curva ROC, ademais do concepto de curva ROC e os principais indicadores de resumo. No segundo e terceiro capítulo abórdanse as principais estimacións para test con resultados continuos e as propiedades de cada unha destas. Veremos ademais como afectan na curvas ROC as covariables, tanto sobre as distribucións dos resultados do test, como sobre a curva ROC. Todos estes procedementos serán ilustrados no último capítulo a partir dunha mostra de datos biomédicos reais mediante o software libre R.

M_{odelización} estatística con Random Forests

Manuel Vaamonde Rivas.
Director: Manuel Febrero Bande.

A finais do século XX, naceu a teoría de árbores de regresión e clasificación. Co fin de mellorar esta teoría e adaptarse a problemas altamente dimensionados, naceron métodos coma o método de *Random Forests*. Este método sèrvese dunha colección aleatoria de árbores, e utiliza a información contida nelas para proporcionar predicións e detectar variables importantes. O obxectivo do traballo é ofrecer una descrición e análise do método de *Random Forests*. A primeira metade do traballo céntrase nas cuestións teóricas acerca do método. En primeiro lugar, ofrecendo unha introdución ás árbores de regresión e clasificación (a partir das cales se crean os *Random Forests*). Posteriormente, explicando o método de Random Forests, así como as características deste e detalles relacionados co seu uso. Na segunda metade examínase o funcionamento do método. Para isto o traballo conta con dous exemplos prácticos do uso dos *Random Forests*. O primeiro é unha análise dunha posición de xadrez; o segundo, un estudo acerca dos crimes violentos e das súas causas. Nesta segunda metade, é utilizado o programa informático R, do que se inclúe fragmentos de código.

M_{ovemento} browniano

Andrea Vilar Álvarez.
Director: Alberto Rodríguez Casal.

Neste traballo preténdese profundizar no concepto de movemento *browniano*. Comézase facendo unha pequena revisión histórica, que permite seguir de preto a evolución do concepto, dende a súa primeira utilización para definir o movemento aleatorio de partículas en suspensión nun fluído, ata a súa definición actual, máis rigorosa e matemática, que o sitúa como un dos procesos estocásticos máis estudados. A continuación, preséntanse algunhas nocións básicas de teoría da probabilidade e procesos estocásticos que nos permiten construír un exemplo de paseo aleatorio para obter, de xeito natural, o movemento *browniano*. Unha vez definido matematicamente, expóñense algunhas características curiosas da súa traxectoria como a continuidade, a non diferenciabilidade ou o seu carácter fractal. Finalmente, o traballo remata cun breve capítulo dedicado á ecuación de *Black-Scholes*, un método matemático de valoración de activos financeiros que permite predicir a evolución dos prezos dun produto financeiro ao longo do tempo, facilitando así a toma de decisións económicas en mercados de todo o mundo.

Coñecéndonos

Coñecendo a unha socia da SGAPEIO: Amalia Jácome Pumar.

Amalia Jácome é profesora do Departamento de Matemáticas na Facultade de Ciencias (UDC) dende 2005, e antes no Departamento de Estatística e Investigación Operativa (UVIGO) no campus de Ourense, de 2000 a 2005. Socia da SGAPEIO dende 1999, membro do Consello Executivo en 2005-2013 e co-editora de *Informest* en 2009-2013, non sabe se lle gusta mais a docencia ou a investigación.



Cantos anos levas traballando en estatística ou IO?

Toda unha vida, dende que comecei a tese doutoral no 2000.

Descríbenos o teu traballo

Imparto Estatística nos Graos de Bioloxía e Química na UDC, e en dous mestrados. Como investigadora do grupo MODES, traballo en Análise de Supervivencia e en aplicacións ás Ciencias da Saúde. Actualmente co-dirixio tres teses de doutoramento.

Dende cando es socio da SGAPEIO?

Dende 1999. Foi o primeiro que fixen ao rematar a licenciatura en Matemáticas.

Cal foi o 1º Congreso da SGAPEIO ao que asistiches?

O IV Congreso da SGAPEIO (1999), en Santiago de Compostela. Levo asistindo dende aquela a todos os congresos da SGAPEIO.

O tema do teu próximo proxecto/artigo?

Estou traballando en contrastes de significación non paramétricos en modelos de curación con A. López-

Cheda e R. Cao, e en análise de datos de metilación do ADN en relación co cancro (co grupo Oncomet) entre outros temas.

Un curso que che gustaría facer?

Calquera de programación con *Big Data* en R.

Que libro de estatística ou I.O. recomendarías?

Calquera de J.A. Paulos, como “El hombre anumérico” ou “Un matemático lee el periódico”, nos que se ilustra de forma moi amena as consecuencias de non entender as teorías da probabilidade e a estatística, e, en xeral, do anumerismo matemático tan tristemente estendido.

Se non foses estatístico, que che gustaría ser?

Teño moita sorte porque non me imaxino facendo outra cousa. Pero, de renunciar á estatística, imaxínome como profesora de calquera materia relacionada coas ciencias.

Se oes “estatística”, que é o primeiro que che vén á cabeza?

A cita “Hai tres clases de mentiras: as mentiras, grandes mentiras e as estatísticas” de Mark Twain. As estatísticas forman parte do día a día, pero o descoñecemento fomenta os malentendidos e a manipulación. Por iso é indispensable a formación en matemáticas e estatística, sobre todo, ante o actual desafío de xestionar e analizar enormes volumes de datos (*Big Data*).

Que é o que máis valoras do labor da SGAPEIO?

A promoción da Estatística, no só a nivel universitario (cursos, xornadas, premios) senón tamén nos institutos (xornadas de formación, Incubadora de sondaxes) e divulgación en xeral.

En que cres que deberíamos mellorar?

Promover a Estatística entre profesionais tan diferentes non é doado. Felicito á SGAPEIO porque penso que está a facer un labor magnífico. Poucas suxestións teño. De dicir algunha, sería tentar atraer socios do mundo empresarial se a SGAPEIO proporcionase (ou fose un enlace para) un servizo de asesoramento estatístico. E, en educación, poderíanse promover convenios para visitas a centros onde a Estatística é fundamental (o IGE, a Unidade de Observación e Predición Meteorolóxica-MeteoGalicia, etc.)



Estatística recreativa

Gran Serpe por Antonio Vaamonde Liste

Setembro do ano 1000. Olaf, rei de Suecia, e mais Svein, rei de Dinamarca, xuntan as súas forzas contra Olaf Triggvesson, rei de Noruega. A súa frota conxunta de máis de setenta barcos agarda oculta tras a illa de Svolder o paso do lexendario *drakkar* “Gran Serpe” do seu inimigo. Un traidor, Sigvalde, tenta levar a frota norueguesa de once barcos a través dun estreito paso cara á emboscada preparada polos seus inimigos.

Gran Serpe é o barco máis grande de Triggvesson e sería doadamente recoñecido entre o conxunto da frota, pero os seus inimigos non o viran antes, e os 11 barcos van pasar o canal, un por un, en orde aleatoria, sen que sexa

posible ver máis naves que a que vai saíndo do estreito, e tan alonxadas entre si que se a coalición ataca un barco que non sexa a nave real, alertará aos seguintes e o rei fuxirá.

“Svein, alí vén o primeiro, e parece grande... Será o do rei...? Ataquemos...!”, “Non, Olaf... non sabemos se é o rei. Quizais pase en derradeiro lugar, ou no medio... elixiremos ao chou un número entre un e once, e ese será a nosa presa, non temos alternativa mellor”.

Hai alternativa? Tes algunha estratexia que permita a Olaf e a Svein elixir mellor a súa presa?

A solución estará dispoñible, proximamente, neste [enlace](#). Permanece atent@s!

Dirección:

Paula Raña Míguez

Equipo editorial:

Esther Calvo Ocampo, M^a Jesús Casado Barrio, Balbina Casas Méndez, Tomás Cotos Yáñez, Rosa Crujeiras Casais, M^a Esther López Vizcaíno, Silvia M^a Lorenzo Freire, Javier Roca Pardiñas e Rebeca Villaverde López.

Revisión lingüística:

María Miragaya Pereira

Coa colaboración de:

José María Sorando Muzás, M^a José Ginzo Villamayor, Salvador Naya Fernández, Javier Tarrío Saavedra, Amalia Jácome Pumar e Antonio Vaamonde Liste.

COLABORA!!!

Informest é unha publicación cuatrimestral da SGAPEIO que podes atopar na web da sociedade, www.sgapeio.es/, no apartado PUBLICACIÓN.

Se queres participar cun artigo de divulgación para a sección “Colaboracións” ou cunha curiosidade estatística para o “Sabías que?”, contacta con nós no correo boletin@sgapeio.es onde te informaremos das normas de edición.

Anímate!!

Comunicación coa SGAPEIO:

Facultade de Matemáticas, Campus Vida, 15706-Santiago de Compostela, A Coruña.

www.sgapeio.es

secretaria@sgapeio.es

boletin@sgapeio.es

Twitter @sgapeio

LinkedIn SGAPEIO

Depósito Legal: LU-191-1995 - I.S.S.N.:1695-7083

#SGAPEIO
Sociedade Galega para a Promoción da
Estatística e da Investigación de Operacións