INFORMEST
Boletín da SGAPEIO. Novembro de 2015. Número 41

Colaboración
Software
Actividades da SGAPEIO
Novas do IGE
Traballos de Estatística e IO
Coñecéndonos
Sabías que?

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

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

Facultade de Matemáticas
15706 Santiago de Compostela

Editorial

Tomás Cotos Yáñez – Vicepresidente da SGAPEIO

Estimados compañeiros e compañeiras, neste ano 2015 a Sociedade tivo e ten unha intensa actividade. Empezamos o ano, en xaneiro, con dúas conferencias e despois coa Asemblea Xeral da Sociedade. No mes de xuño a Sociedade participou nunha xornada sobre Big Data e Estatística na Facultade de Informática da Universidade da Coruña con sesións plenarias e un obradoiro. Ao mesmo tempo e durante o primeiro semestre do ano levouse a cabo a edición V do Concurso Incubadora de Sondaxes e Experimentos tendo o seu remate en xuño coa entrega de premios da fase local na Facultade de Matemáticas en Santiago de Compostela, e, en xullo, en Madrid organizado pola Facultade de Estudios Estatísticos da Universidade Complutense de Madrid e a Sociedade de Estatística e Investigación de Operacións, coa entrega de premios da fase nacional onde o representante de Galicia da Categoría ESO tivo unha mención de honra polo traballo "Reproducción de las tablas de crecimiento infantil que utilizan los pediatras".

Dende xullo ata setembro fíxose unha consulta virtual para a aprobación de cambios nos estatutos da Sociedade, consulta que foi aprobada.

Hai uns días, rematou en Lugo o XII Congreso da Sociedade dedicado ao sector agroalimentario como eixo principal, un éxito de organización e participación. O último día do congreso celebrouse a Asemblea Xeral da Sociedade onde se aprobaron as contas do 2015 e o orzamento para o ano 2016.

De aquí ata que remate o ano aínda teremos máis actividades, concretamente:

Unha exposición de Estatística "Explorística" no Museo Pedagóxico de Galicia (MUPEGA) en Santiago de Compostela, onde se ilustran conceptos de estatística e probabilidade dun xeito aplicado aos alumnos de ESO e Bacharelato, iniciativa da Sociedade Portuguesa de Estatística.

Tamén en novembro teremos unha nova Xornada de formación para elaborar proxectos de Estatística coorganizada pola Sociedade e AGAPEMA dirixida a profesores de ESO e Bacharelato.

Como vedes, unha actividade frenética!

Para rematar, dende o Comité Executivo queremos agradecer a colaboración prestada por todas aquelas persoas e institucións que fixeron posible todas estas actividades. Tamén, solicitar de todas aquelas persoas que crean poder aportar contidos de interese para o noso boletín que se poñan en contacto coa directora de Informest ou con calquera outro integrante do Consello Executivo. A colaboración de todos os membros da SGAPEIO é indispensable para manter o interese de INFORMEST.

Recibide un cordial saúdo.

Da estrutura económica á renda dos fogares: trazos da economía galega

por Raúl Jácome Rodríguez

Raúl Jácome Rodríguez é economista e traballa no IGE dende o ano 2003, actualmente como Xefe do Servizo de Contas Económicas.

O Dicionario da Real Academia Galega define o termo "estrutura" como a "*maneira de estaren dispostas e organizadas as diferentes partes dun todo, sendo esas partes dependentes do todo, así como solidarias e dependentes entre si*". Se esas partes son os distintos sectores que conforman un todo que é a economía no seu conxunto, falariamos da estrutura económica.

Se queremos botar unha ollada á estrutura económica de Galicia, a mellor ferramenta dispoñible é o sistema de contas económicas que elabora o Instituto Galego de Estatística. As operacións estatísticas de síntese que forman este sistema pretenden dar unha imaxe fiel da economía no seu conxunto, das relacións entre os sectores e entre estes e o sector exterior. *Grosso modo* diremos que estas operacións cumpren dous tipos de obxectivos básicos: descritivos, centrados en proporcionar unha fotografía detallada da economía nun momento dado, e analíticos, degrañando unha realidade complexa como é a economía coa finalidade de realizar previsións económicas ou simulacións de política económica, por exemplo.

O crecemento económico, a distribución da actividade no territorio, a desigualdade deben partir do coñecemento da realidade económica galega, posto que o coñecemento da nosa estrutura económica contribúe a interpretar ou entender problemas máis complexos como a baixa produtividade do traballo, os salarios baixos ou o despoboamento de diversas zonas da nosa comunidade.

Nesta contribución farei un percorrido por algúns trazos da nosa estrutura produtiva, centrándome no enfoque da oferta, isto é, qué sectores caracterizan o tecido produtivo de Galicia, e como esta caracterización contribúe a explicar que os fogares galegos teñan unha renda dispoñible bruta por habitante que só chega ao 91 % da media española[1].

Unha visión xeral da nosa estrutura económica

A economía galega non foi allea ao proceso de tercialización das economías occidentais nas últimas décadas. O sector servizos é o que achega máis valor engadido ao produto interior bruto (PIB) de Galicia: o 62,4 % no ano 2014. Pese á tradición rural e mariñeira da nosa comunidade o sector primario non alcanza o 5 % do PIB, achega xa superada polo sector enerxético, a construción e, por suposto, a industria manufactureira[2].

Se comparamos estes resultados con España ou coa Unión Europea teremos unha idea máis precisa do que é Galicia hoxe, fronte ás economías da nosa contorna. A táboa 1 é esclarecedora. O peso do sector servizos é inferior en Galicia, a industria ten unha achega superior e o sector primario, pese a representar menos dun 5 % no PIB, ten unha maior presenza nosa rexión.

Agora ben, convén debullar máis estes sectores para completar a imaxe da economía galega e das súas compoñentes e contrastalas coa estrutura das economías española e europea.

	Galicia	**España**	**UE-28**
Primario	4,5	2,3	1,4
Industria	17,3	15,5	16,9
Construción	6,3	4,9	4,8
Servizos	62,4	68,4	71
Impostos netos sobre os produtos	9,5	8,9	5,9
PIB	**100**	**100**	**100**

Táboa 1: Distribución do PIB pola vía da oferta. Ano 2014. Datos en porcentaxes.

Especialización produtiva fronte a España

A oficina estatística da Comisión Europea (EUROSTAT) obriga aos Estados membros a remitir datos desagregados do PIB a 64 ramas de actividade, e no caso de Galicia incluso se dispón dunha desagregación máis ambiciosa (73 ramas de actividade) no sistema de contas. Estas desagregacións realízanse conforme a unha clasificación de actividade estándar o que permite realizar comparacións homoxéneas, e coñecer, por exemplo, cómo de diferentes son a economía galega e española.

Se definimos uns índices de especialización produtiva como o cociente entre o peso dun sector na economía galega e o peso dese mesmo sector pero na economía española, teremos un cociente que tomará valor igual a 1 se a achega porcentual do sector é a mesma en ambas as economías. Un maior valor deste índice nun determinado sector implica unha especialización maior.

Este modo de analizar a estrutura produtiva resalta aqueles sectores que nos caracterizan fronte a España, pero non necesariamente aqueles sectores que achegan máis valor engadido á nosa economía. Por exemplo, o comercio polo miúdo achega o 4,5 % en 2013, pero este sector ten unha achega similar en España, polo que non podemos dicir que é característico da nosa economía.

	CNAE-09	Índice de especialización	Peso sobre o PIB (%)
Pesca e acuicultura	03	10,3	1,2
Silvicultura e explotación forestal	02	6,3	0,5
Industria da madeira e da cortiza	16	3,2	0,6
Fabricación de vehículos de motor e compoñentes	29	2,7	2,2
Outras actividades profesionais científicas e técnicas	74-75	1,9	0,7

Táboa 2: Especialización produtiva na economía galega. Promedio 2010-2012.

Na táboa 2 presentamos os cinco sectores cun maior índice de especialización produtiva. O sector da pesca é o que ten un índice de especialización máis elevado con respecto a España (10,3), porén o seu peso na nosa economía[3] sitúase nun 1,2.

A continuación sitúase o sector da silvicultura e explotación forestal e a industria da madeira e da cortiza, primeiros elos dunha cadea produtiva (a cadea forestal-madeira) que se completa coa industria do papel e a fabricación de mobles, ambas cun coeficiente de especialización moi inferior (0,7 e 0,8 respectivamente)[4].

A fabricación de vehículos de motor e compoñentes ten un elevado peso na economía galega e un índice de especialización de 2,7. Dentro da industria manufactureira este é o sector co maior valor engadido[5], moi por riba de sectores senlleiros na nosa economía como a industria alimentaria, a metalurxia ou o téxtil.

Por último, aparece a rama de actividade das outras actividades profesionais científicas e técnicas, que medrou considerablemente nos últimos anos froito da externalización de parte dos servizos de deseño industrial realizados previamente no sector industrial.

Produtividade e salarios: Somos menos produtivos?

É de utilidade confrontar os ditos índices con outros valores dos sectores como pode ser a produtividade do factor traballo, o que nos permitirá coñecer se aqueles sectores nos que estamos especializados son sectores de produtividade alta ou baixa. Unha economía especializada en sectores de produtividade alta terá, previsiblemente, uns salarios máis elevados[6] e unha renda dos fogares superior a unha economía especializada en sectores de baixa produtividade.

As tres ramas de actividade con maior índice de especialización teñen tanto unha produtividade media inferior á galega como unha ganancia salarial media tamén inferior (ver táboa 3, columna "a" e "b"). A especialización en sectores cunha produtividade baixa transmítese á renda dos traballadores a través de ganancias salariais tamén baixas.

Porén estes sectores son competitivos no contexto español xa que presentan produtividades superiores ás estatais, pero esta maior competitividade non se reflicte nuns salarios máis elevados. Nos cinco sectores a produtividade por hora traballada é superior á española en todos eles, porén esta alta produtividade non se traslada a unha ganancia salarial maior (ver táboa 3, columnas "c" e "d"), agás no caso da silvicultura, se ben esta rama ten pouco emprego asalariado.

Este é un fenómeno xeneralizado na nosa economía. Se clasificamos cada rama de actividade da economía galega en función de se a súa produtividade e ganancia salarial está por riba da media española no sector, e vemos os postos de traballo en cada unha delas, temos unha distribución dos ocupados en Galicia en función das características da rama de actividade na que traballan (táboa 4).

		Con respecto á economía galega		Con respecto ao mesmo sector en España	
	CNAE-09	Produtividade (a)	Ganancia (b)	Produtividade (c)	Ganancia (d)
Pesca e acuicultura	03	67,8	67,5	103,3	82,6
Silvicultura e explotación forestal	02	95,9	76,2	100,1	144,8
Industria da madeira e da cortiza	16	78,4	91,5	116,3	95,1
Fabricación de vehículos de motor e compoñentes	29	153,5	106,2	101,7	67,6
Outras actividades profesionais científicas e técnicas	74-75	113,0	98,7	156,0	93,8

Táboa 3: Produtividade e ganancia salarial por hora traballada. Promedio 2010-2012.

	Produtividade inferior á media española	Produtividade superior á media española	TOTAL
Ganancia inferior á media española	68,3	17,0	85,3
Ganancia superior á media española	10,5	4,2	14,7
TOTAL	78,8	21,2	100,0

Táboa 4: Distribución dos postos de traballo en función da produtividade e ganancia salarial. Galicia. Ano 2013

O 78,8 % do emprego en Galicia está concentrado en ramas que teñen unha produtividade inferior á media española. Isto é, o emprego en Galicia está concentrado en actividades de baixa produtividade. Ademais o 17 % dos postos de traballo no ano 2013 estaban en ramas que tiñan unha produtividade superior á media estatal pero unha ganancia inferior. Neste caso teríamos que falar da incapacidade de trasladar á renda do factor traballo esa produtividade superior. De feito só o 4,2 % do emprego estaría en ramas que teñen unha produtividade e ganancia superior á media nacional. Como resultado temos un 85,3 % do emprego en ramas cunha ganancia salarial inferior á media nacional.

Algunhas reflexións finais

Nos últimos anos asentouse unha corrente que pretende ir alén do PIB[7] como medida de desenvolvemento económico dun espazo. Esta iniciativa defende que este agregado económico é unha medida insuficiente e que debe ser complementada con outras que midan o desenvolvemento sustentable, a renda dos fogares ou a calidade de vida.

A renda dos fogares en Galicia ten un nivel baixo en relación a España, como consecuencia dun PIB que se xera en actividades caracterizadas por unha baixa produtividade, pero tamén noutras nas que, pese a ter unha produtividade elevada, non se consegue trasladar á remuneración do factor traballo. Os procesos de redistribución de renda a través de impostos, transferencias, prestacións e cotizacións sociais contribúen a suavizar esta situación, pero a remuneración do factor traballo é o principal compoñente da renda bruta dispoñible dos fogares, polo que baixos niveis salariais implican menor renda bruta dispoñible. Non só iso. Unha remuneración media baixa implica unha prestación social futura baixa o que limitará o efecto redistribuidor do Estado na renda dos fogares galegos.

Volvendo á definición coa que comecei esta colaboración, son as diferentes partes dun todo as que condicionan a situación actual da renda en Galicia, sendo estas partes dependentes entre si e dependentes do todo. Mudar esta situación non é traballo sinxelo, pero todo indica que posibilitar o incremento da produtividade do traballo na nosa rexión é o primeiro paso.

Un sistema de información económica coherente, integrado e cun nivel de desagregación ambicioso como o que elabora o IGE constitúe unha ferramenta especialmente útil para analizar este e outros problemas, máis cando temos unha estrutura económica diferenciada da estatal.

Notas

[1] Ver INE. Contabilidad Regional de España. Base 2010. Cuenta de Renta de los Hogares.
[2] No ano 1980 o sector primario achegaba preto do 14 % do PIB mentres os servizos, representaban algo máis da metade da actividade económica total. Véxase. IGE (2010): "A economía galega a través do Marco-Input Output de Galicia 2005", Xunta de Galicia, Consellería de Facenda, páx 10.
[3] A pesca, acuicultura e a industria transformadora do pescado representan en 19 concellos galegos máis do 10 % do PIB municipal. Véxase IGE (2015): "Análise do sector da Pesca". Xunta de Galicia. Consellería de Facenda. Páxinas 13 e seguintes.
[4] Unha das características do sector forestal galego é o elevado peso das ramas máis vencelladas ao territorio (a primeira transformación da madeira) e con menor capacidade para crear valor engadido. Véxase IGE (2013): "Análise da cadea forestal-madeira". Xunta de Galicia. Consellería de Facenda.
[5] É un sector moi vinculado ao exterior pero cunha dependencia moi alta dos inputs importados. Véxase IGE (2014): "Análise do sector da automoción". Xunta de Galicia. Consellería de Facenda.
[6] A produtividade definímola como o cociente entre o valor engadido e as horas traballadas e a ganancia salarial como o cociente entre a remuneración de asalariados e as horas traballadas asalariadas. Esta é unha limitación no enfoque desta colaboración xa que o traballo autónomo (non asalariado) está contribuíndo á produtividade do sector pero non está incluído nesta ganancia salarial.
[7] O coñecido como Informe Stiglitz/Sen/Fitoussi é o xerme da iniciativa "Beyond GDP" promovida por EUROSTAT

Táboas

1. Fonte: Elaboración propia a partir de IGE. Contas Económicas Trimestrais de Galicia. Base 2010; INE. Contabilidad Nacional de España; EUROSTAT. National Accounts

2, 4 e 5. Fonte: Elaboración propia a partir de IGE. Contas Económicas Anuais; INE. Contabilidad Nacional de España

3. Fonte: Elaboración propia a partir de IGE. Contas Económicas Anuais

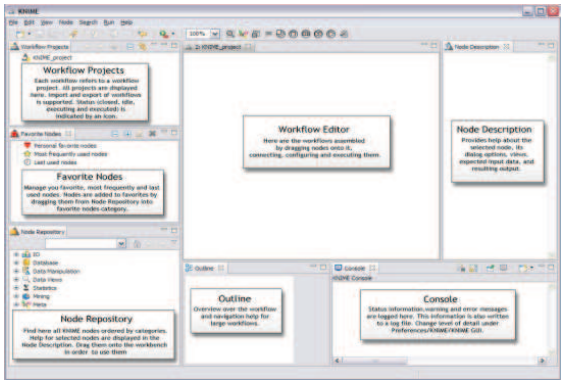
O programa KNIME

por Irene Castro Conde e David Lozano Núñez

KNIME (Konstanz Information Miner)

KNIME é un entorno gratuíto para o desenvolvemento e execución de técnicas de Minería de Datos baseado no deseño dun fluxo de execución que plasme as etapas dun proxecto nun entorno visual. Este programa está concibido como unha ferramenta gráfica que dispón dunha serie de nodos (que ofertan distintos tipos de algoritmos) e frechas (que representan fluxos de datos), que se despregan e combinan de maneira interactiva permitindo a montaxe fácil e rápida de fluxos para o pre-procesamento, o modelado e a visualización de datos. Este software está construído baixo a plataforma Eclipse e programado, esencialmente, en Java.

KNIME foi desenvolto orixinariamente no departamento de Bioinformática e Minería de Datos da Universidade de Constanza, Alemaña, baixo a supervisión do profesor Michael Berthold. Dende o ano 2006 ten sido empregado na investigación farmacéutica, en intelixencia de negocio e na análise de datos financeiros, entre outros campos. KNIME distribúese baixo a Licenza Pública Xeral (GPL) de GNU que garante aos usuarios finais a liberdade de empregar, compartir e modificar o *software*. O carácter aberto desta ferramenta fai posible a súa extensión mediante a creación de novos nodos que apliquen algoritmos á medida do usuario. Ademais, existe a posibilidade de incorporar de maneira sinxela código creado nas linguaxes R ou Python, así coma o uso doutras ferramentas de Minería de Datos (Weka) ou a importación de bases de datos e a súa exploración con consultas SQL.



Na imaxe anterior amósase unha vista xeral da interface de KNIME, a cal se divide en varias ventás:

- *Explorador de proxectos*: ventá onde se mostran todos os proxectos ou fluxos dispoñibles.
- *Editor de fluxos*: ventá onde se crea o fluxo de traballo arrastrando, conectando, configurando e executando os nodos.
- *Repositorio de nodos*: contén todos os nodos ordenados por categorías.
- *Descrición dos nodos*: proporciona axuda sobre o nodo seleccionado.
- *Nodos favoritos*: mostra os nodos favoritos ou que se usan máis frecuentemente.
- *Esquema*: vista xeral sobre o fluxo de traballo que serve de moita axuda se este é moi grande.
- *Consola*: amosa as mensaxes de advertencia e información.

Os nodos dispoñibles en KNIME, que se poden ampliar instalando novas extensións do repositorio, levan a cabo distintos tipos de accións que se poden executar sobre unha táboa de datos. Estes son algúns exemplos:

- *Manipulación*: traspoñer matrices, ordenar filas, filtrar filas ou columnas, combinar ou dividir columnas, converter e substituír valores da columna, etc. Todo isto permitiranos facer mostraxes, transformacións, agrupacións,...
- *Visualización*: histogramas, diagramas de caixa, gráficos de dispersión, táboas interactivas,...
- *Estatísticas*: mínimo, máximo, media, desviación, varianza, correlación, contrastes de hipóteses,...
- *Minería*: creación de modelos estatísticos, de Aprendizaxe Automático e de Minería de Datos (análises de compoñentes principais, regresión, análise cluster,...). Validación dos modelos, aplicación destes sobre conxuntos novos de datos, cálculo da matriz de confusión e da curva ROC,...

Coa combinación destes nodos pódense crear tamén informes a medida grazas á súa integración con *BIRT* (Business Intelligence and Reporting Tools).

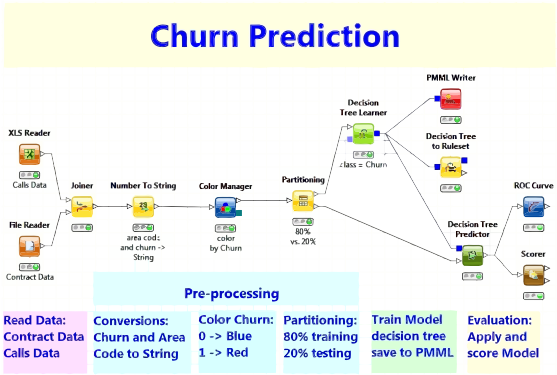
Unha das vantaxe da ferramenta KNIME é que, unha vez executado un nodo ou un fluxo de traballo, non se precisa volver a executalo xa que o resultado obtido en cada nodo queda gardado. É dicir, KNIME permite executar de forma selectiva algúns dos pasos creados, así como executar todo o fluxo creado. Trala execución, os resultados poden ser explorados mediante vistas interactivas tanto dos datos coma dos modelos. Esta característica, xunto coa súa sinxeleza e a posibilidade de integración con outras ferramentas (p. ex. R e Hadoop), fan que KNIME sexa moi útil na exploración, preparación, análise e modelado de grandes volumes de datos.

Exemplo: Predición do *churn* (taxa de abandono)

Como vimos de comentar, podemos empregar KNIME para dar solución a problemas de Minería de Datos nos que esteamos interesados. En concreto, veremos como exemplo un dos problemas que resulta de maior interese dentro do sector das Telecomunicacións, a predición do *churn*.

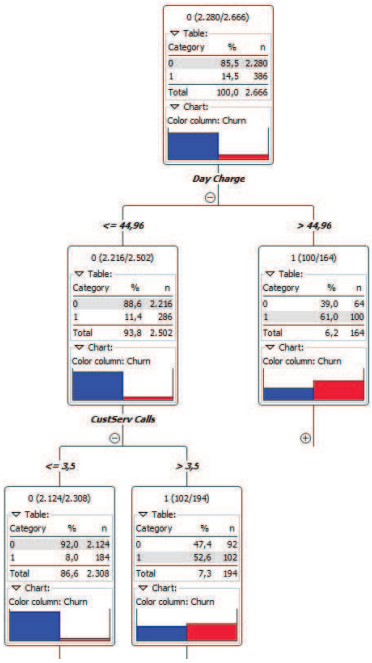
O anglicismo *churn* utilízase para describir a taxa de abandono dos clientes, é dicir, aqueles clientes que deixan a compañía ou o provedor dun servizo durante un período de tempo determinado. É coñecido que resulta entre cinco e quince veces máis caro captar novos clientes que reter aos actuais. Isto fai que a predición do *churn* sexa unha arma moi potente para coñecer de forma temperá que clientes abandonarán a compañía co fin de aumentar o grao de retención, construír unha sólida estratexia e aforrar custos de adquisición para ser más eficientes e competitivos.

A continuación achégase un fluxo básico, creado en KNIME, para predicir as baixas. Este fluxo componse dunha serie de nodos que realizan as seguintes accións: lectura de dúas bases de datos, unión destas, preprocesado dos datos, separación en adestramento e test, creación dun modelo (do que se extraen e gardan as regras) e a súa avaliación (mediante a curva ROC e a matriz de confusión).



Esta análise poderíase levar a cabo con diversas técnicas de clasificación implementadas nos nodos de KNIME como poden ser: Naive Bayes, Árbores de Decisión, Random Forest, KNN, Redes Neuronais,...

Neste caso o modelo seleccionado para realizar a predición é unha Árbore de Decisión (nodos "Decision Tree Learner" e "Decision Tree Predictor") que nos devolve a predición binaria (baixa ou non baixa), a probabilidade asociada e a regra de clasificación de cada cliente. Estas regras resultan moi útiles para a compañía xa que delas se extrae información adicional aportando os motivos polos que cada cliente decide darse de baixa, o que permitirá levar a cabo accións de retención e fidelización personalizadas (o nodo "PMML Writer" permite gardar as regras). Por outra banda, o nodo "Scorer" calcula a matriz de confusión do modelo aplicado aos datos de test que xunto coa Curva ROC (nodo "ROC Curve") permite cuantificar como de bo é o modelo.



Na imaxe anterior amósanse as primeiras pólas da Árbore de Decisión obtida ao aplicar este modelo aos datos no fluxo de exemplo. A maneira de interpretar esta árbore sería a seguinte: se a variable "Day Charge" é menor ou igual que $44{,}96$ entón a porcentaxe de clientes que abandonan a compañía é do $11{,}4\,\%$, pero se ademais a variable "CustServ Calls" é maior que $3{,}5$ entón esta porcentaxe sobe ao $52{,}6\,\%$.

Debido a que a clasificación se realiza tomando como limiar a probabilidade $0{,}5$, os clientes con probabilidades de baixa maiores ou iguais a este valor serán clasificados como *churners*. Polo tanto, considérase que os clientes que cumpran a regra anterior ("Day Charge"$\leq 44{,}96$ e "CustServ Calls" $> 3{,}5$) daranse de baixa.

Información Adicional

Para máis información sobre KNIME véxase a páxina web: https://www.knime.org, onde se atopan dispoñibles diversos recursos para empezar a manexar esta ferramenta como documentación de axuda, vídeos e un foro onde facer consultas.

Actividades da SGAPEIO

O XII CONGRESO SGAPEIO DE PRINCIPIO A FIN

Luis Alberto Ramil e José Manuel Colmenero

O pasado sábado 24 de outubro rematou o XII Congreso SGAPEIO organizado polo Departamento de Estatística e Investigación Operativa da USC en colaboración coa SGAPEIO. Foron tres días intensos de congreso, con variadas actividades científicas pensadas para dar a coñecer os avances na investigación teórica e aplicada, e con numerosas actividades sociais programadas para facilitar os contactos entre os congresistas e, ao mesmo tempo, dar a coñecer o patrimonio cultural da cidade sede do congreso e da súa provincia. Mais o congreso non se pode entender sen todo o traballo previo do comité organizador para acadar unha programación de interese, e dos congresistas que, coas súas comunicacións, contribúen a dar un bo nivel científico ao congreso. E destes traballos previos é do que imos falar.

As primeiras decisións que afrontamos no comité organizador foron escoller a sede do congreso e maila elección da temática principal da programación oficial. Hai xa ano e medio que acudimos á Deputación de Lugo a solicitar a celebración do congreso na súa sede, e o apoio desta institución foi total desde o primeiro momento. Non só puidemos celebrar o congreso no magnífico Pazo de San Marcos, senón que ademais contamos coa valiosa axuda das áreas de Presidencia, Economía e Cultura e os servizos de Protocolo, Reprografía, Novas Tecnoloxías e Publicacións. Á Deputación sumáronse con apoios destacables o Concello de Lugo, o Centro Tecnolóxico Agroalimentario de Lugo (CETAL), o Grupo TRAGSA e o grupo de investigación MODESTYA da USC.

Como eixo principal da programación oficial eliximos dar a coñecer as aplicacións da Estatística e da Investigación de Operacións no sector agroalimentario, área de especialización do campus universitario de Lugo, o Campus Terra da USC, e sector de gran relevancia na economía galega e mesmo española. O xoves 22 de outubro as conferencias plenarias correron a cargo de especialistas moi vencellados a empresas galegas. Jorge López e Julio González amosaron aos asistentes o importante papel que poden xogar a Estatística na industria alimentaria e a Investigación de Operacións na toma de decisións para optimizar a planificación de grandes proxectos industriais. O venres 23 as conferencias plenarias correron a cargo de grandes especialistas como Philippe Vieu e Mónica Bécue, que amosaron aos congresistas técnicas para extraer información relevante das mostras de datos funcionais, moi comúns nas análises quimiométricas dos alimentos, e a metodoloxía para a análise estatística de datos sensoriais obtidos mediante paneis de catas de alimentos, realizadas con técnicas vangardistas. A introdución a estas técnicas facilitárona o taller de R para datos funcionais, impartido por Manuel Oviedo, e a cata de viños e o taller de FactoMineR coordinados por Josep A. Sánchez e Mónica Bécue. O xoves e o sábado celebráronse dúas mesas redondas, na primeira das cales interviñeron representantes de importantes empresas galegas do sector da alimentación (Estrella Galicia e GADISA) e de organismos moi relevantes no ámbito da saúde pública en Galicia (Hospital Clínico Universitario de Santiago de Compostela e Laboratorio de Saúde Pública de Galicia). Na segunda, interviñeron membros de servizos de asesoramento estatístico de Galicia, tanto da empresa (Biostatech) coma do sector público, do ámbito universitario (Servizo de Consultoría Estatística da USC) e do ámbito da saúde pública (Hospital Clínico Universitario de Santiago de Compostela e Complexo Hospitalario Universitario A Coruña). Ao mesmo tempo que acontecían estas e outras actividades diversos grupos de estudantes de centros de ensino medio, acompañados polos seus profesores, visitaron a exposición *Explorística* de xogos educativos de temática estatística. A presentación oficial desta exposición correu a cargo de Conceição O. Nunes Rocha en representación da Sociedade Portuguesa de Estatística.

Quédanos falar das contribucións dos congresistas, que supoñen un traballo previo moi importante, polo seu número e a súa calidade. Nesta edición deuse difusión ao

congreso nos tres idiomas principais do congreso: galego, castelán e portugués. O período de inscrición dividiuse en tres tramos con cotas crecentes. No primeiro tramo, ata finais de xullo, animábase aos interesados a adiantar os seus resultados e ofrecíaselles a posibilidade de realizar un dos dous cursos non presenciais previos, de R e EPIDAT. Ao remate deste período xa estaban inscritos 74 congresistas. No segundo tramo, ata principios de outubro, foron abundantes as comunicacións recibidas de fóra de Galicia e Portugal, e abríronse as portas a un recoñecemento da Consellería de Educación das xornadas do congreso como actividade de formación do profesorado de ensino medio. Finalmente o éxito acadado foi importante: 126 congresistas que presentaron 72 comunicacións. Cabe destacar a presenza dun número importante de investigadores estudantes. Os traballos dos congresistas foron organizados en seis sesións de comunicacións orais de estatística, dúas de investigación operativa, unha sesión de comunicacións para profesores de ensino medio, unha sesión de comunicacións póster e unha sesión especialmente destacable, a sesión de comunicacións de novos investigadores, na que nesta ocasión houbo dous primeiros premios outorgados pola SGAPEIO, resultando gañadores Paula Raña e Ángel M. González. Os nosos parabéns para eles. O labor dos congresistas foi ben merecidamente recollido nun número especial de INFORMEST, editado por María José Ginzo e Paula Raña, e nas actas dixitais do congreso, publicadas polo Servizo de Publicacións da Deputación de Lugo en colaboración coa SGAPEIO.

Para complementar as actividades científicas, ao longo dos tres días do congreso os congresistas e os seus acompañantes puideron desfrutar á súa elección de interesantes e divertidas actividades sociais e culturais. O xoves houbo unha recepción no Concello de Lugo, coa presenza da alcaldesa da cidade, e unha visita ao museo da Domus de Mitreo que rematou cun concerto de piano, "Resonancias aritméticas", a cargo de Juan Ignacio Martínez, no auditorio da Casa do Saber. O venres houbo unha excursión á Ribeira Sacra, patrocinada pola Deputación Provincial e na que colaboraron o Consello Regulador da D.O. Ribeira Sacra e o Consorcio de Turismo Ribeira Sacra. Nesta viaxe os asistentes puideron apreciar os catro grandes valores desta comarca candidata a Patrimonio da Humanidade: o viño, coa visita á bodega Regina Viarum; os canóns do Sil, percorridos en catamarán; a cerámica de Gundivós, asistindo a unha exhibición da súa elaboración por parte do oleiro Elías; e o patrimonio románico, coa visita á igrexa de Proendos.

Xa para despedir o congreso, moitos congresistas asistiron ao xantar oficial que se celebrou no restaurante Manuel Manuel, quen xa se encargara dos cafés-descanso do congreso. A guinda final puxérona dúas actuacións para rematar a comida-festa: o contacontos Celso F. Sanmartín e o mago Rafa fixéronnos rir abondo e sorprendéronnos coas súas extraordinarias habilidades.

Só nos queda recoñecer que o bo tempo que fixo durante os días do congreso tamén axudou. Como mostra podedes ver a foto de grupo tomada na excursión á Ribeira Sacra.

Moitas grazas a todos os participantes e, ata a próxima!

EXPLORÍSTICA: AVENTURAS NA ESTATÍSTICA

Ana Bouzas

Entre os días 21 e 24 de outubro, o XII Congreso SGAPEIO acolleu a Exposición Explorística: "Aventuras na Estatística", procedente de Portugal e avalada pola Agência Nacional para a Cultura Científica e Tecnolóxica Portuguesa (Ciência Viva). Esta exposición foi deseñada fundamentalmente para ilustrar experimentalmente algúns conceptos de Estatística e Probabilidade a alumnos de ESO e Bacharelato.

A exposición está formada por seis módulos interactivos que integran diversas actividades relacionadas coa estatística e adaptadas aos niveis formativos dos alumnos aos que vai dirixida esta acción. Son seis módulos independentes nos que o estudantado, distribuído en grupos reducidos, fai fronte aos retos que se lles propón en cada fase: Recepción de participantes, tiro con arco, submarino, quizz de censos, datos viciados e Sondaxe.

Explorística foi visitada por cerca de 200 persoas, das cales 125 eran alumnos de ESO e Bacharelato pertencentes a 5 Centros Escolares de Galicia: IES Francisco Daviña Rey (Monforte de Lemos), Colexio Fingoy (Lugo), IES de Arzúa, IES Perdouro (Burela) e C.P.I Dr. López Suárez (Friol).

Esta actividade divulgativa estivo cofinanciada pola SGAPEIO, pola Rede Inbioest (Consellería de Educación e Ordenación Universitaria da Xunta de Galicia), así coma polo proxecto "Bioestadística para Vivir y cómo Vivir de la Estadística" apoiado pola Fundación para a Ciencia e Tecnoloxía (FECYT) "Ministerio de Economía y Competitividad". Dito proxecto está promovido pola USC, a través do grupo Interdisciplinar de Estadística, Computación, Medicina e Bioloxía (GRIDECMB), xunto co Instituto Galego de Estatística (IGE), co apoio da Rede Nacional de Bioestatística (BIOSTATNET). O seu obxectivo é amosar a importancia que, a estatística en xeral e a bioestatística en particular, teñen para a sociedade ao permitir aportar solucións a problemas reais.

Explorística enmarcouse dentro duns Talleres titulados "Quieres Jugar a ser Estadístico" que teñen por finalidade a de transmitir aos alumnos por qué e cómo a estatística, xoga un papel tan importante na nosa sociedade, así como dar a coñecer entre os xoves unha vocación profesional, actualmente en crecente demanda.

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA
UNIÓN EUROPEA
FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL "Unha maneira de facer Europa"
XUNTA DE GALICIA
CONSELLERÍA DE EDUCACIÓN E ORDENACIÓN UNIVERSITARIA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
INSTITUTO GALEGO DE ESTATÍSTICA
SGAPEIO
Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións

Novas do IGE

Últimas novidades en publicacións e documentos do IGE

ENQUISA ESTRUTURAL A FOGARES. Ocio e hábitos culturais. Edición 2014

A Enquisa estrutural a fogares é unha actividade estatística anual dirixida aos fogares galegos co obxecto de obter información sobre as súas características socioeconómicas. Consta de dúas partes: unha primeira de carácter xeral que se repite todos os anos, deseñada para recoller información sobre variables básicas e unha segunda de carácter específico. O módulo específico de 2014 está dedicado ao ocio e hábitos culturais e ten como obxectivo coñecer os hábitos culturais e de lecer da poboación galega. Desta enquisa pódese extraer que a porcentaxe de poboación galega de 5 ou máis anos que asiste ao cine polo menos unha vez ao mes é do 8,3 %. Esta porcentaxe caeu 4,9 puntos respecto ao ano 2006. No que respecta á asistencia a espectáculos deportivos, aumentou a porcentaxe de galegos que asiste de xeito habitual (polo menos unha vez no mes) a este tipo de eventos: do 10,0 % en 2006 ata o 13,8 % en 2014. Pódese obter máis información en: web

MARCO INPUT-OUTPUT DE GALICIA

O Marco Input-Output de Galicia (MIOGAL) é unha meta de información recollida no Plan Galego de Estatística que ten por obxectivo a descrición do proceso produtivo dentro da Comunidade Autónoma galega e os fluxos de bens e servizos da economía. A este obxectivo, de carácter moi xenérico, pódese engadir outros máis concretos: afianzar o sistema de contas económicas de Galicia, consolidarse como principal ferramenta para o cambio de base contable e facer un balance das fontes dispoñibles no sistema estatístico. O IGE publicou o Marco Input-Output de Galicia correspondente ao ano 2011. Pode consultarse máis información en: web

ESCENARIO MACROECONÓMICO

O IGE difunde por primeira vez o escenario macroeconómico para Galicia con horizonte bianual. O escenario macroeconómico refírese á evolución prevista do Produto Interior Bruto (PIB) e dos seus compoñentes desde a óptica da demanda: consumo privado, consumo público, formación bruta de capital, exportacións e importacións de bens e servizos. O cadro complétase coas previsións do emprego: postos de traballo equivalentes a tempo completo e taxa de paro. Segundo este escenario marcroecónico, o PIB de Galicia para o ano 2015 medrará un 2 % en taxa de variación interanual. Máis información en: web

CONFERENCIA DO DÍA MUNDIAL DA ESTATÍSTICA

Con motivo do Día Mundial da Estatística que se celebra cada 5 anos, o IGE organiza o día 20 de outubro ás 12:00 horas, no salón de actos da Consellería de Sanidade, unha conferencia a cargo do Subdirector Xeral de Presupostación e Seguimento dos ingresos Tributarios da Axencia Estatal da Administración Tributaria (AEAT), Rafael Frutos Vivar, co título " vella estatística económica nos tempos do *big data*: a renda dos fogares na Contabilidade Nacional e outras historias exemplares".

CATÁLOGO DE SERVIZOS

En aplicación da Lei 14/2003, do 26 de decembro, de racionalización do sector público autonómico, en setembro de 2014 o Consello da Xunta aprobou o catálogo de servizos, que foi publicado no DOG nº7, do 13 de xaneiro de 2015, como Resolución do 17 de novembro de 2014, da Secretaría Xeral Técnica da Consellería de Presidencia, Administracións Públicas e Xustiza.

O catálogo inclúe servizos de asistencia técnica e xurídica, funcionamento, control financeiro, auditoría, avaliación de resultados, saúde laboral, formación, administración electrónica, innovación, desenvolvemento tecnolóxico e estatística. Estes servizos só se poderán contratar externamente cando a súa autoprovisión sexa inviable. O IGE ofrece servizos de formación, elaboración de diagnósticos e previsións, recolla de información e análise de datos.

Traballos de Estatística e Investigación Operativa no SUG

Teses

Valores coalicionais en xogos cooperativos con utilidade transferible

Julián Costa Bouzas.
Directores: José María Alonso Meijide e Ignacio García Jurado.
Data da defensa: Maio, 2015. UDC.

Nesta memoria abórdanse diversas cuestións relativas á teoría do valor en xogos cooperativos con utilidade transferible nos que existe unha estrutura coalicional que condiciona a negociación entre os xogadores.
Dentro do contexto dos problemas de asignación de costes, faise un recorrido por valores coalicionais de Shapley eficientes, e estúdase a súa aplicación a unha subclase destes problemas, os xogos do aeroporto. Analízase a utilización do valor de Owen nos xogos de asignación de custos de mantemento en presenza dunha estrutura coalicional e proponse unha expresión fácil e rápida para o seu cálculo.
Estúdanse os valores coalicionais mínimo cuadráticos. Exponse, e resólvese, un problema de optimización cuadrática con restricións lineais que xorde de forma natural ao aproximar un xogo xenérico mediante xogos inesenciais. Obtéñense, como solución particular, o valor de Owen e o valor coalicional simétrico de Banzhaf.
Por último, introdúcese o valor particional proporcional de Shapley, un novo valor con dominio nos xogos monótonos cunha estrutura coalicional que, aínda que segue o enfoque do valor de Aumann-Drèze, incorpora as opcións externas dos xogadores. Estúdanse as súas propiedades, comparándoas coas do valor de Aumann-Drèze, analízanse as súas respectivas caracterizacións axiomáticas e preséntanse potenciais campos de aplicación.

Traballos fin de Máster

O problema do carteiro chinés

Jesús Yordá Pérez.
Directora: Silvia Mª Lorenzo Freire.
Data de defensa: Setembro, 2015. USC.

A Investigación de Operacións ou Investigación Operativa fai uso de métodos cuantitativos como ferramenta de apoio para o proceso de toma de decisións. Moitos dos problemas de investigación operativa poden presentarse como problemas de programación lineal, non obstante, pese a que a formulación é correcta, cando o tamaño do problema crece, o tempo computacional vólvese excesivo. Por isto, os matemáticos intentan desenvolver algoritmos, unhas veces exactos, e outras veces heurísticos, que resolvan estes problemas nun tempo razoable. Nesta memoria introducimos brevemente o problema do carteiro chinés coas súas principais variantes, xunto cos principais algoritmos para resolvelo. Por outra parte presentaremos un paquete de R capaz de resolver o problema do carteiro chinés e poñerémolo a proba con datos reais cedidos pola empresa Ecourense.

Clasificación supervisada en datos multivariantes, de alta dimensión e funcionais.

Xián López Álvarez.
Director: Manuel Febrero Bande.
Data de defensa: Setembro, 2015. USC.

Este traballo revisa varias técnicas da clasificación estatística supervisada, detallando a teoría sobre a que se sustentan. Tamén se explica a representación de datos funcionais en bases de *B-splines* e compoñentes principais, e como os coeficientes nestas poden ser empregados para a clasificación. Escóllense unha serie de bases de datos, obtidas no repositorio da UCI e en certas librerías de R, e aplícanse os modelos seguindo un determinado procedemento de sintonización de parámetros e estimación da precisión. Finalmente amósanse os resultados e compáranse os diferentes clasificadores, analizando cuestións como o tipo de variables ou o tamaño dos grupos. Tamén se estuda a dependencia, no caso dos datos funcionais, fronte ao número de elementos das bases.

Análise das liñas de ómnibus urbanos da Coruña

Juliana Díaz López.
Director: Rubén Fernández Casal.
Titor: Pablo J. Vázquez.
Data de defensa: Setembro, 2015. UDC.

O presente traballo de fin de mestrado é de carácter público aplicado e foi tutelado no Concello da Coruña, que se mostrou interesado en realizar unha análise estatística, sobre as liñas de ómnibus urbanos da Coruña. O traballo divídese en dúas partes. En primeiro lugar, preséntanse as diferentes ferramentas utilizadas e, de pasaxe, os pequenos resumos que foron feitos de como utilizar algunhas delas. A seguir, na segunda parte, realizouse unha análise descritiva clásica dos datos, unha análise exploratoria descritiva, considerando que eses datos forman unha rede, e xa para finalizar, representáronse graficamente datos das paraxes de ómnibus e variábeis de tramos entre dúas paraxes, como por exemplo, a velocidade media.

Estimación presuavizada para o problema de truncamento

Ana Nicieza Iglesias.
Director: César Andrés Sánchez Sellero.
Data de defensa: Setembro, 2015. USC.

Neste traballo estúdase o problema que sofre o estimador límite-produto baixo truncamento aleatorio, polo perigo de que os conxuntos a risco sexan pequenos, ou incluso que conteñan un único individuo. Un conxunto a risco unitario dá lugar a resultados anómalos do estimador límite-produto, mentres que conxuntos a risco con poucos individuos elevan considerablemente a varianza do estimador.
Posteriormente proponse un método de corrección do estimador límite-produto para superar estes inconvenientes, que consiste en fixar un límite mínimo para os conxuntos a risco. Para a selección do límite proponse un método *bootstrap*. Estúdanse mediante simulacións as propiedades de estimadores corrixidos mediante límites fixos, así como do selector *bootstrap*. Obtense que a corrección por un límite mínimo proporciona estimacións máis eficientes, en termos do erro cuadrático medio, que o estimador límite-produto ordinario.

Modelización conxunta de datos lonxitudinais e de supervivencia. Aplicacións a datos biomédicos

Arís Fanjul Hevia.
Directora: Carmen María Cadarso Suárez.
Data de defensa: Xullo, 2015. USC.

O modelado conxunto de análises de supervivencia con datos lonxitudinais recibiu moita atención nos últimos anos. O seu uso foise estendendo cada vez máis en estudos clínicos de seguimento, xa que neles adoitamos encontrar datos lonxitudinais para os cales as técnicas habituais de análise de supervivencia non sempre resultan adecuadas. Por outra parte, neste tipo de estudos tamén se pode dar máis dun evento de fallo, o que conleva a necesidade de utilizar riscos competitivos á hora de analizar a supervivencia. A intención desde proxecto é ilustrar como se poden combinar as análises de supervivencia, os datos lonxitudinais e os riscos competitivos, así como as vantaxes que aporta este modelado conxunto. Ademais, as técnicas estudadas aplícanse a uns datos procedentes dun programa de diálise peritoneal, realizado pola Unidad de Dialysis Peritoneal do Hospital Geral de Santo António (Portugal).

Traballos fin de Grao en Matemáticas na área de Estatística

Os traballos presentados nesta sección foron defendidos ao longo do curso 2014–2015 na facultade de Matemáticas da USC como Traballos Fin de Grao dirixidos dende o Departamento de Estatística e Investigación de Operacións.

Modelización de series de tempo financeiras

Laura Borrajo López.
Director: Manuel Febrero Bande.

A modelización de series de tempo financeiras ten como obxectivo principal describir a estrutura dinámica subxacente nos valores observados co fin de establecer unha pauta de actuación coa que poder predicir valores futuros. As series financeiras posúen certas peculiaridades como son a presenza de heterocedasticidade condicional, que orixina a aparición dunha estrutura dinámica na varianza, ou o agrupamento da volatilidade, particularidades que ocasionan que a metodoloxía *ARIMA* empregada na modelización de series de tempo sexa insuficiente

para analizar series temporais financeiras. Co fin de modelizar a varianza, xorden os procesos *GARCH*. Neste proxecto preséntanse as ferramentas teóricas precisas para a comprensión destes modelos e emprégase o software estatístico R para exemplificar o funcionamento de ambos procesos. No caso da modelización *Box-Jenkins*, axústase un modelo *ARIMA* á serie que recolle as cantidades mensuais de pasaxeiros en liñas aéreas comerciais en EE.UU., ampliando o estudo coa modelización de atípicos e a predición final a horizonte 19. No tocante á serie financeira, realízase unha análise sobre o IBEX 35, para o que se escollen os prezos de peche diarios que a entidade BBVA tivo na bolsa de Madrid nos últimos 10 anos, axustando un modelo *ARIMA* e un *GARCH* para a modelización da media e da varianza, respectivamente. Conclúese coa predición dos 40 futuros valores e da correspondente volatilidade.

Regresión con datos dependentes.

Ana Belén Buide Carballosa.
Directora: Rosa M. Crujeiras Casais.

Este traballo consiste nunha introdución ao modelo de regresión lineal con erros dependentes. Con este tipo de modelos atopámonos na práctica, por exemplo, cando se observan procesos ao longo do tempo ou en localizacións do espazo. Comezamos facendo unha revisión dos conceptos básicos da regresión lineal. Logo, levamos a cabo unha formulación do modelo incluíndo unha estrutura de dependencia nos erros, que se recolle na matriz de covarianzas dos erros a cal pasa a ter elementos non nulos fóra da diagonal. Presentamos os estimadores (por mínimos cadrados xeneralizados) dos parámetros do modelo, nos cales se ten en conta a matriz de covarianzas dos erros, dando solución á falta de eficiencia estatística que presenta o método de mínimos cadrados ordinarios cando os erros presentan dependencia. Tratamos o modelo de regresión con heterocedasticidade nos erros e o modelo de regresión con erros autorregresivos de orde 1. Ademais, facemos unha simulación en R coa que se ilustra o bo funcionamento do estimador do modelo, tendo en conta a matriz de covarianzas dos erros, a través de técnicas de Monte Carlo. Para rematar, introducimos o modelo de regresión con dependencia espacial e ilustrámolo con datos reais.

Modelos de regresión con covariables en alta dimensión

Patricia Calvo Pérez.
Director: Wenceslao González Manteiga.

Neste traballo descríbense os modelos de regresión en alta dimensión, que se caracterizan porque o número de variables explicativas é moi grande en comparación co tamaño de mostra. Estes modelos xorden de forma natural en moitas áreas tales como a bioinformática, a economía, a xenética, etc, e supoñen un reto para as técnicas estatísticas tradicionais. O traballo comeza introducindo os modelos de regresión simple e múltiple, ambos os casos particulares do modelo linear xeral. Descríbese con detalle o tradicional método de estimación por mínimos cadrados, xunto con algunhas nocións de inferencia sobre o modelo. A continuación, defínense os modelos de regresión en alta dimensión e arguméntase por que a estimación por mínimos cadrados non é axeitada. Motivadas polo fenómeno da ?sparsity?, que asume que só un subconxunto das variables explicativas está realmente asociado coa resposta, formúlanse técnicas alternativas como os métodos de selección, regularización e redución da dimensión. Para rematar, exemplifícanse os métodos discutidos utilizando unha base de datos real a través do software estatístico R.

Introdución aos modelos de colas

Sandra Carballido Regueiro.
Directora: María de los Ángeles Casares de Cal.

A teoría de colas é a disciplina que se encarga do estudo da formación de liñas de espera dentro dun sistema. Este sistema constitúeno uns clientes que requiren un servizo ofrecido por uns servidores que non son capaces de satisfacer simultaneamente toda a demanda. O obxectivo desta teoría é obter as ecuacións que modelan esas colas para así acadar información sobre factores tales como o tempo de espera medio ou o número de clientes que agardan a ser atendidos. No traballo introdúcense os conceptos necesarios para o desenvolvemento da teoría de colas e estúdase un modelo sinxelo: o modelo M/M/s. Así mesmo preséntanse exemplos prácticos onde se pon de manifesto a utilidade das expresións obtidas previamente. Ademais, a existencia de cuestións complexas que non poden abordarse de xeito analítico fai necesario o uso dunha nova ferramenta que aproxime a solución do problema. Por iso faise tamén unha introdución á simulación no campo da teoría de colas.

Implicacións do comportamento racional en teoría de xogos

Joaquín Carballo Mato.
Director: Julio González Díaz.

Esta memoria iníciase cunha breve introdución á teoría de xogos co fin de demostrar o Teorema de Nash, teorema que deu lugar a unha expansión sen precedentes desta disciplina. Na parte principal do traballo introduciremos un modelo de información que nos aporta unha descrición completa da información que, previa ao inicio do xogo, é relevante. Este modelo permitiranos traballar con situacións onde os xogadores teñen distinta información antes de que o xogo comece. Con esta nova ferramenta definiremos un novo concepto de equilibrio, o equilibrio correlacionado, que permitirá, ao contrario que no equilibrio de Nash, que os xogadores coordinen as súas estratexias de certa maneira. Trala introdución do concepto de solución denominado "eliminación iterada de estratexias estritamente dominada" remataremos demostrando as implicacións que ten o concepto de racionalidade e de que este sexa "common knowledge" no concepto de solución estudados ao longo da memoria.

Distancias estatísticas

Brais Durán Cerviño.
Directora: Beatriz Pateiro López.

O concepto de distancia ou medida de similitude entre individuos é esencial en moitas técnicas da análise multivariante, por exemplo en problemas de clasificación ou *clustering*. Neste traballo preténdese facer unha revisión exhaustiva das principais distancias estatísticas, das súas propiedades e aplicacións. Para iso, teremos que ter en conta o estudo das principais medidas de proximidade ou de asociación e a formación de conglomerados. En segunda instancia, a partir dunha base de datos, ilustrar, mediante o software estatístico R, parte da metodoloxía anteriormente descrita, coa conseguinte explicación dos resultados acadados.

O Bootstrap

Víctor González Carro.
Director: Alberto Rodríguez Casal.

O *Bootstrap* é a técnica de remostraxe máis empregada na estatística dende a súa introdución por B. Efron en "Bootstrap methods and their applications" (1979), superando a outras tan coñecidas como o *Jacknife*. A pesar de que inicialmente foi formulado para estimar a varianza dun estatístico, unha das máis notables aplicacións do *Bootstrap* é a creación de intervalos de confianza para parámetros poboacionais. Tanto a estimación da varianza coma os intervalos de confianza son o eixo en torno ao que xira o traballo, mostrando os procedementos clásicos para ditos cometidos e aportando algunhas solucións en termos do *Bootstrap* ao tempo que se sinalan os seus inconvenientes. Para levar a cabo esta tarefa, ademais de presentar ambos os casos de xeito teórico, engadíronse un estudo de simulación e un problema con datos reais que axudan a comprender esta comparativa no caso dos intervalos de confianza. En particular, a partir de dúas equisas sobre novas tecnoloxías realizadas polo IGE aos fogares galegos nos anos 2008 e 2013, trátase de comparar as proporcións de domicilios con internet en Galicia nestes anos.

Técnicas de redución da dimensión en datos de alta dimensión

Diego Juárez Rodríguez.
Director: Wenceslao González Manteiga.

As técnicas de redución da dimensión abranguen diversos procedementos de procesamento de datos que pretende reducir o número de variables consideradas nun modelo. O que se consegue é atopar unha representación máis compacta dos datos. A análise de compoñentes principais é o método máis coñecido para a redución da dimensión e ten tido éxito en moitos ámbitos, como a xenética e a intelixencia artificial, utilizada para dixitalización de documentos manuscritos ou no recoñecemento facial. Por outra banda, a análise en correlación tense utilizado con notable éxito en xenética para detectar a relación entre a expresión de determinados xenes e a aparición de distintos tipos de cancro. Aínda que estes métodos tivesen éxito en moitos ámbitos, tamén teñen desvantaxes. Cando se teñen datos de grandes dimensións perden interpretabilidade. Se, ademais, temos unha mostra de tamaño pequeno, estes procedementos fallan. Para solucionalo, desenvolvéronse posteriormente unha serie de técnicas denominadas *sparse*. Estas últimas teñen acadado un notable éxito tanto nos campos antes mencionados como noutros novos, como o ámbito da compresión de imaxes. Tanto é así que calquera transmisión de audio e vídeo en alta definición aprovéitase delas. Neste traballo desenvolverase a análise en compoñentes principais e a análise en correlación canónica. Veranse as súas vantaxes e inconvenientes e desenvolveranse as súas correspondentes técnicas *sparse*. Por último, aplicaranse estas ferramentas a unha base de datos reais.

O problema do viaxante

Daniel Mato Regueira.
Director: Alberto Rodríguez Casal.

O problema do viaxante (ou the traveling salesman problem, TSP) consiste en atopar a ruta óptima para un vendedor que debe visitar unha soa vez, comezando nunha orixe, un número prefixado de cidades para logo regresar ao punto de partida. Esta ruta óptima será a que teña a menor lonxitude. O TSP é un dos problemas mellor estudados dentro da optimización, cunha ampla gama de problemas aos que se pode adaptar. Aínda que existen algoritmos que permiten calcular a solución exacta, a súa complexidade computacional fan aconsellable utilizar aproximacións heurísticas. Neste traballo, ademais de facer un breve percorrido pola historia do TSP, revisáronse os elementos básicos para a introdución e aplicación dos máis coñecidos algoritmos para a resolución do problema, centrándonos no estudo e programación en R do algoritmo heurístico das formigas. Finalmente obtívose unha ruta para un carteiro con distintos algoritmos, considerando o problema como un TSP, e comparáronse os resultados, tendo en conta tamén a posible ruta que o profesional escollería.

Estimación de mesturas de distribucións

Silvia Novo Díaz.
Directora: Rosa María Crujeiras Casais.

As mesturas finitas de distribucións proporcionan unha ferramenta moi útil para describir situacións reais nas que aparecen fenómenos de multimodalidade ou asimetrías, características que adoitan ter a súa orixe na presenza de categorías ou grupos ocultos dentro da poboación de estudo. O contrapunto á grande versatilidade destes modelos reside, a miúdo, na presenza dun elevado número de parámetros que, polo clásico método de máxima verosimilitude, resultan difíciles de estimar. Neste Traballo de Fin de Grao, analizaremos a estrutura dos modelos de mesturas e veremos distintas aplicacións dos mesmos, tales como a aproximación da densidade dos datos anuais de renda do ano 2012. Así mesmo, mostraremos a solución que o algoritmo xeral *Expectation-Maximization* (EM) proporciona para a estimación dos seus parámetros, facendo especial finncapé nas mesturas finitas de normais. Finalmente, por medio dun estudo de simulación, analizaremos as propiedades dos estimadores proporcionados polo algoritmo para o caso da mestura finita de normais unidimensionais.

Estimación de conxuntos: algoritmos para o cálculo do "medial axis".

María Jesús Pérez Pena.
Directora: Beatriz Pateiro López.

A estimación de conxuntos ten como obxectivo recuperar un conxunto descoñecido a partir dunha mostra de puntos provinte del. Podemos considerar parte da estimación de conxuntos o feito de estimar unha característica do conxunto. Se nos paramos a pensalo, posiblemente coñecendo a área ou o volume dun conxunto, debemos ter en conta con que tipo de conxunto estamos traballando, poidamos resolver unha gran variedade de problemas sen necesidade de chegar a obter o conxunto en si. Isto pode facilitar o noso traballo xa que estimar unha destas características resulta, en xeral, menos custoso. Forma parte das características máis importantes o "medial axis". Esta característica ten gran relevancia en distintos campos, como por exemplo na análise de imaxes. A súa importancia implicou que fose estudado por diversos autores. Podemos pensar no "medial axis" como a parte central ou o corazón do conxunto. Dun xeito máis formal debemos definir o "medial axis" como o conxunto de puntos que teñen máis dunha proxección sobre a fronteira do conxunto. Neste traballo recóllese a noción de "medial axis" atendendo ás súas propiedades e aos distintos algoritmos existentes para o seu cálculo. Na parte final do traballo realizouse en R a programación dun algoritmo para a estimación do "medial axis" de conxuntos convexos.

Coñecéndonos

Coñencendo a un socio da SGAPEIO: Juan Carlos Pardo Fernández.

Juan Carlos Pardo Fernández é profesor titular na Univesidade de Vigo, onde imparte docencia na Escola de Enxeñaría Industrial e na Facultade de Económicas. A súa investigación céntrase no estudo de contrastes para modelos de regresión nun contexto non paramétrico ou semiparamétrico.

Cantos anos levas traballando en estatística ou IO?

Desde o ano 2000. Ao rematar a carreira estiven un ano traballando como bolseiro no Instituto Galego de Estatística e despois xa me incorporei á Universidade de Vigo.

Descríbenos o teu traballo

Son profesor universitario, así que iso implica que o meu traballo se reparte entre a docencia e a investigación. A maior parte da miña docencia está na Escola de Enxeñaría Industrial de Vigo e no Máster en Técnicas Estatísticas, que coorganizan as tres universidades galegas. A miña investigación está enfocada fundamentalmente ao desenvolvemento de técnicas para contrastar hipóteses en modelos de regresión non paramétrica e semiparamétrica.

Dende cando es socio/a da SGAPEIO?

Pois tería que consultalo con detalle, pero creo que desde o ano 1999.

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

Curiosamente, ao primeiro congreso da SGAPEIO ao que asistín non o fixen en calidade de asistente, senón como compoñente do Coro Universitario da USC. Recordo que fixemos unha pequena actuación no acto de clausura do congreso do ano 1997 en Lugo. Xa como congresista, o meu primeiro congreso foi o de Compostela do ano 1999.

O tema do teu próximo proxecto/artigo?

Actualmente estou empezando unha colaboración sobre contrastes de especificación con modelos de series temporais, en concreto con modelos GARCH.

Un curso que che gustaría facer?

Hai algúns anos impartín docencia nun máster de lingua e literatura inglesas. Impartía técnicas básicas de estatística para a investigación en lingüística, pero nunca me puiden meter a fondo en temas máis específicos dese ámbito, como por exemplo lingüística computacional. É algo que fun deixando, pero estaría encantado de formarme máis nese campo.

Que libro de estatística ou I.O. recomendarías?

Un libro que me parece moi instrutivo e moi claro é "*Asymptotic Statistics*", de Aad van der Vaart (Cambridge University Press). Creo que explica moi ben todas as técnicas necesarias para a obtención de resultados asintóticos en inferencia estatística.

Se non foses estatístico/a , que che gustaría ser?

Pois non o teño moi claro. Cambio bastante de parecer, pero case seguro que algo relacionado coa música.

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

Claramente, datos. E a continuación o seu tratamento mediante técnicas axeitadas.

Que é o que máis valoras do labor da SGAPEIO?

Creo que a SGAPEIO fai un gran labor de cohesión entre o conxunto de persoas vencelladas á estatística e á investigación de operacións en Galicia. Tamén é importante o labor de divulgación que leva a cabo.

En que cres que deberiamos mellorar?

Pois en pouca cousa. Tendo en conta o carácter completamente altruísta dos participantes na directiva, pouco máis se pode pedir.

Sabías que?

Unha mirada á mosca da froita por Paula López Casado

As enfermidades neurodexenerativas do ser humano foron estudadas a través da retina da Drosophila Melanogaster, ou, comunmente chamada, mosca da froita[i]. Este insecto é un grande aliado en investigacións xenéticas posto que se reproduce facilmente, ten un curto ciclo de vida (15-21 días) e se coñece o mapa completo do seu xenoma. O curioso é que gran parte dos seus xenes son homólogos aos dos humanos[ii]. En particular, o ollo constitúe o órgano elixido para a investigación neurodexenerativa porque ten cantidades similares de dopamina ás do ollo humano. Con todo, detectar exactamente o grao de dexeneración da retina non é tarefa fácil e necesítanse técnicas máis sofisticadas que permitan illar patróns comúns de forma automática.

Os investigadores do Instituto de Bioloxía e Xenética Molecular (UVa-CSIC), entre outros, decidiron facer uso de ferramentas automatizadas de análises de imaxes con algoritmos de clasificación sustentados nun modelo estatístico. Así, tomando unha simple fotografía do ollo da mosca, poderían discernir o seu grao de dexeneración. De cada imaxe tomaron todas as características posibles medibles, eliminando en primeiro lugar aquelas que estaban máis correladas entre si e, a través dun procedemento *stepwise*, as menos significativas. Coa análise de compoñentes principais (PCA) conseguiron discriminar en grupos os diferentes niveis de dexeneración dos ollos e estimar a probabilidade de pertencer a cada clase cun modelo de Regresión Loxística Multinomial. Ademais probaron o propio método utilizando tests de validación cos que comprobaron que o modelo final era adecuado para discernir os diferentes graos da enfermidade.

Con este método conséguese non só reducir os tempos de análises senón tamén aumentar o grao de confianza e robustez no resultado pois proporciona un índice que mide o estado da enfermidade tomando tan só tres características da imaxe. O mellor do método é que podería ser xeneralizado para medir calquera outro patrón biolóxico.

O estudo completo pode atoparse nesta web.

[i] www.seresmodelicos.csic.es

[ii] ciencia.nasa.gov

Dirección:

Paula Raña Míguez

Equipo editorial:

Esther Calvo Ocampo, Mª Jesús Casado Barrio, Balbina Casas Méndez, Tomás Cotos Yáñez, Rosa Crujeiras Casais, Mª Esther López Vizcaíno, Silvia Mª Lorenzo Freire, María Loureiro García, Javier Roca Pardiñas, Rebeca Villaverde López.

Coa colaboración de:

Ana Bouzas, Irene Castro Conde, José Manuel Colmenero, Raúl Jácome Rodríguez, Paula López Casado, David Lozano Núñez, Juan Carlos Pardo Fernández e Luis Alberto Ramil .

COLABORA !!!

Informest é unha publicación cuadrimestral da SGAPEIO que podes atopar na web da sociedade, www.sgapeio.es, no apartado PUBLICACIÓNS.

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

SGAPEIO

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

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