IX XORNADA DE USUARIOS DE R EN GALICIA

20 de outubro de 2022

LIBRO DE RESUMOS

> ORGANIZA

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

IX XORNADA DE USUARIOS DE R EN GALICIA

PROGRAMA E RESUMOS

20 de outubro de 2022

Organiza: Asociación de usuarios de software libre da Terra de Melide

IX XORNADA DE USUARIOS DE R EN GALICIA

Editora: María José Ginzo Villamayor

ISBN: 978-84-09-44852-4

Presentación

A Asociación de usuarios de software libre da Terra de Melide (MeLiSA) comprácese en presentar a IX Xornada de Usuarios de R en Galicia.

Pretende ser un punto de encontro para todas aquelas persoas interesadas en intercambiar as súas experiencias e atopar colaboracións co resto da comunidade, ademais trátase de promocionar e difundir o coñecemento libre da linguaxe estatística R e mostrar as súas aplicacións.

O programa contempla vinte e un relatorios ao longo de todo o día. Dos cales sete son convidados e ás outras catorce atenderon á chamada de recepción de propostas.

Entre os participantes figuran especialistas do Instituto Español de Oceanografía (IEO), da Xunta de Galicia: diferentes entidades como a Consellaría de Sanidade ou o Instituto Galego de Estatística, das tres universidades galegas, de universidades estranxeiras: Escuela Politécnica Nacional (Ecuador), Universidade Federal Fluminense (Brasil) e da Academia da Força Aérea (Brasil), do Instituto de investigación sanitaria de Santiago de Compostela e un profesor de Ensino Medio do IES Pedra da Auga (Ponteareas).

Todo isto non sería posible sen o patrocinio de AMTEGA á que agradecemos a súa contribución.

Santiago de Compostela, outubro de

2022 O Comité Organizador

Comité organizador

María José Ginzo Villamayor
Universidade de Santiago de Compostela

Rafael Rodríguez Gayoso
Asociación de usuarios de software libre da Terra de Melide

Miguel Ángel Rodríguez Muíños
Dirección Xeral de Saúde Pública (Consellería de Sanidade)

Comité científico

María José Ginzo Villamayor
Universidade de Santiago de Compostela

Miguel Ángel Rodríguez Muíños
Dirección Xeral de Saúde Pública (Consellería de Sanidade)

Información xeral

IX XORNADA DE USUARIOS DE R EN GALICIA

Data

20 de outubro de 2022

Lugar de celebración

Aula Magna. Facultade de Matemáticas (USC)

Web das xornadas

https://www.r-users.gal/

Certificados

Todos os certificados remitiranse ás persoas solicitantes en formato dixital por correo electrónico unha vez rematada a IX Xornada.

Programa

IX XORNADA DE USUARIOS DE R EN GALICIA

20 de outubro de 2022

09:00 – 09:15	Sesión de apertura Mª Elena Vázquez Cendón (*Decana da Facultade de Matemáticas*), Salvador Naya Ferández (*Vicerreitor de Política Científica, Investigación e Transferencia – Universidade da Coruña*), Rafael Rodríguez Gayoso (Melisa), Miguel Ángel Rodríguez Muíños (*Saúde Pública – Consellería de Sanidade*)
09:15 – 09:35	El paquete biosensors.usc Marcos Matabuena Rodríguez. *Universidade de Santiago de Compostela*
09:35 – 09:55	TUGlabR, un paquete de R para juegos coalicionales Iago Núñez Lugilde. *Universidade de Vigo*
09:55 – 10:15	Aplicacións dos gráficos de control para estimar a carga viral da COVID-19 nas augas residuais Salvador Naya Fernández. *Universidade da Coruña*
10:15 – 10:35	Comparativa entre medidas culturales en el contexto de la toma de decisiones turísticas Yago Atrio Lema. *Universidad de Santiago de Compostela*
10:35 – 10:55	knobi: an R package implementing Known-Biomass Production Models Anxo Paz Cuña. *Instituto Español de Oceanografía – CSIC*
10:55 – 11:15	LearningStats: Un paquete en R para a docencia Sabela Varela Rey. *Universidade de Santiago de Compostela*
11:15 – 12:00	**PAUSA**
12:00 – 12:20	Geoportal: Indice Multidimensional de Pobreza Miguel Flores. *Escuela Politécnica Nacional (Ecuador)*
12:20 – 12:40	Do one thing every day that scares you Ariel Levy. *Universidade Federal Fluminense (Brasil)*
12:40 – 13:00	Colaborar y pedir ayuda en StackOverflow Marcos Fernández Arias. *Xunta de Galicia*
13:00 – 13:20	Aplicación de consulta de datos no marco Input-Output de Galicia en R-Shiny Esther Calvo Ocampo. *Instituto Galego de Estatística*
13:20 – 13:40	Prototipado rápido con R: RAD de aplicacións Shiny e AEDA de datos Miguel Ángel Rodríguez Muíños. *Saúde Pública – Consellería de Sanidade*
13:40 – 14:00	A new measure of dependence: distance correlation María Vidal García. *Universidade de Santiago de Compostela*
14:00 – 16:15	**PAUSA**
16:15 – 16:35	Construindo Mandalas com R Luciane Ferreira. *Academia da Força Aérea (Brasil)*
16:35 – 16:55	Utilização do pacote AHP na tomada de decisão Orlando Celso Longo. *Universidade Federal Fluminense (Brasil)*
16:55 – 17:15	New covariates selection method in dynamic regression models with a public implementation in R language Ana Ezquerro. *Universidade da Coruña*
17:15 – 17:35	Matching para o estudo do impacto de políticas públicas José Manuel Amoedo Meijide. *Universidade de Santiago de Compostela*
17:35 – 17:50	**PAUSA**
17:50 – 18:10	Con R de fRikismo. Learning VS Knowing, the match of the century Álvaro Fernández Theotonio. *Universidade de Santiago de Compostela*
18:10 – 18:30	R para o processamento de textos. Uma aplicação para a ordenação do texto em função da complexidade oracional Afonso Xavier Canosa Rodríguez. *IES Pedra da Auga (Ponteareas)*
18:30 – 18:50	REFREG: Un paquete de R para estimar rexións de referencia Óscar Lado Baleato. *Instituto de investigación sanitaria de Santiago de Compostela*
18:50 – 19:10	Desarrollo de una metodología para la evaluación del riesgo de invasión basada en modelos de distribución de especies: el caso de Vespa velutina en Europa Victoria Formoso Freire. *Universidad de Santiago de Compostela*
19:10 – 19:30	clustcurv: Clustering of nonparametric curves Nora M. Villanueva. *Universidad de Vigo*
19:30 – 19:35	Clausura María José Ginzo Villamayor. *Universidade de Santiago de Compostela – Comité Científico*

Índice

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Matching para o estudo do impacto de políticas públicas

José Manuel Amoedo[1]

[1] Departamento de Economía Aplicada, Grupo de Investigación ICEDE, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela, España.

RESUMO

Neste artigo abordamos o uso do matching para o estudo do impacto das políticas públicas empregando R.

Palabras e frases chave: Matching/ políticas públicas/ innovación/ emparellamento

1. INTRODUCIÓN

O "matching" (tamén coñecido como pareo ou emparellamento) pode ser definido, a grandes riscos, como calquera método que busque igualar (ou equilibrar) a distribución das covariables entre os grupos de control e tratamento. O uso desta técnica permite obter grupos equilibrados (ou balanceados) en estudos non experimentais nos que ambos grupos poden non ser homoxéneos entre si. Desta forma, os individuos de ambos grupos pasan de amosar características non similares antes do matching a amosar características similares reducindo as diferenzas. Polo tanto, a mostra non aleatoria convertese nunha mostra con características propias dunha mostra aleatoria.

O matching é xeralmente empregado para estimar efectos casuais, pero tamén é empregado, en ocasións, para aspectos non causais como, por exemplo, o estudo das desigualdades raciais[1].

Desta forma, este traballo aborda os seguintes contidos. En primeiro lugar, recóllemos as principais distancias e as metodoloxías de matching existentes, de forma moi breve. A continuación, recollemos os principais recursos dispoñibles para o seu uso en R centrándonos na librería MatchIt. En terceiro lugar, recollemos o caso práctico empregando os datos de Colombia. Finalmente, recollemos as principais conclusións obtidas ó longo do proceso.

2. MATCHING PARA EL ANÁLISIS DE POLÍTICAS PÚBLICAS: EL CASO DE LA COMPRA PÚBLICA DE INNOVACIÓN

O matching foi desenvolvido ó longo dos anos e, na actualidade, existen diferentes metodoloxías que empregan diferentes medicións da distancia existente entre cada individuo, así como multitude de ferramentas para o seu uso que comprenden dende extensións de Stata, SPSS ata librerías de R. Neste apartado, incidimos inicialmente sobre as medicións da distancia e as metodoloxías empregadas para, posteriormente, facer o mesmo coas librerías dispoñibles en R, centrándonos na librería MatchIt. Finalmente, nun terceiro punto introducimos o proceso seguido coa librería MatchIt[2] para o caso da Compra Pública de Innovación en Colombia.

2.1 Medición e tipoloxías de Matching.

Os pasos a seguir á hora de aplicar o matching son os seguintes. En primeiro lugar é preciso elixir a distancia e as covariables a empregar para obter o equilibrio de todas as covariables e un bo balance das covariables. En segundo lugar, seleccionar a metodoloxía a empregar para levar a cabo o pareo e analizar se existe balance nas covariables seleccionadas.

No caso da distancia (D_{ij}), esta pode ser definida, para dous individuos calquera, como unha medición da similitude entre ambos individuos i e j. As distancias empregadas na

actualidade son a exacta, a de Mahalanobis, a propensity score e a linear propensity score, definíndose cada unha delas como segue:

I. **Exacta:**

$$D_{ij} = \begin{cases} 0 \; if \; X_i = X_j \\ \infty \; if \; X_i \neq X_{j,} \end{cases}$$

II. **Mahalanobis:**

$$D_{ij} = (X_i - X_j)' \sum^{-1} (X_i - X_j)$$

III. **Propensity score[1]:**

$$D_{ij} = |e_i - e_j|,$$

IV. **Linear propensity score:**

$$D_{ij} = |logit(e_i) - logit(e_j)|,$$

A propensity score raramente é coñecida fora de experimentos aleatorios e por iso debe ser estimada. Calquera modelo que relacione unha variable binaria con unha serie de regresores pode ser empregada. O máis común é empregar a estimación loxística, aínda que tamén se acostuman a empregar metodoloxías non paramétricas. No caso desta estimación os aspectos ós que prestar atención non son os comúns de calquera estimación econométrica, xa que cuestións como a capacidade explicativa do modelo ou a multicolinealidade non son relevantes. O aspecto máis relevante a ter en conta neste contexto é a existencia de equilibrio entre as covariables do grupo de control e o grupo de tratamento.

No que respecta ás tipoloxías de matching cada unha parte dunha forma de pareo distinta que leva a un número de individuos que se manteñen tras o pareo e nos pesos relativos que cada individuo recibe. Concretamente, podemos diferenciar dous grupos de metodoloxías claramente diferentes.

O primeiro grupo son as metodoloxías baseadas no Nearest neighbor matching (veciño máis próximo) e o segundo os métodos de Subclassification, Full matching e Weighting.

I. Nearest neighbor matching.

O veciño máis próximo, na súa forma máis sinxela, 1:1 selecciona para cada individuo de tratamento i un individuo de control a unha menor distancia do individuo i. Para mellorar os resultados obtidos a partir desta metodoloxía base existen diferentes variacións que presentan diferentes alternativas de cara a obter un mellor resultado en termos de equilibrio. Unha das variacións do veciño máis próximo é o Optimal matching, que ten en conta o conxunto de emparellamentos cando se seleccionan os emparellamentos individuais, minimizando a distancia global. Desta forma, o Optimal matching mellora o veciño máis próximo básico á hora de seleccionar pares ben emparellados pero non á hora de buscar o equilibrio entre os grupos.Outra alternativa é o coñecido como Ratio matching que permite asignar a cada individuo de tratamento varias parellas k:1. Isto leva a uns sesgos maiores pero reducen a varianza. Para evitar emparellamentos de mala calidade é frecuente empregar tamén un calibrador (caliper) que seleccione so aqueles que se atopan no rango marcado. Unha última variación é a do uso ou non uso de reemprazo. O uso de reemprazo leva a que un individuo de control poida ser asignado como parella a varios individuos de tratamento. Isto leva a unha redución dos sesgos, pero xera problemas xa que no grupo de control os individuos non son independentes e o grupo pode chegar a ser moi reducido.

II. Subclassification, Full matching e Weighting.

O forte desta tipoloxía fronte á anterior é o feito de que nela empréganse todos os individuos de control dispoñibles, sen descartar aqueles que non son emparellados. Isto lévase a cabo empregando ponderadores, que toman un valor entre cero e un para os individuos de control. A metodoloxía de subclassification forma grupos de individuos similares tomando algún criterio formando un número determinado de subclases que debemos seleccionar previamente. O full matching é unha forma máis sofisticada de subclassification que selecciona de forma automática o número de subclases. Por

[1] Onde e_k es la propensity score do individuo k definida en detalle máis adíante.

último, o Weighting emprega a propensity score para ponderar os individuos de tratamento e control en ambos grupos.

2.2 Recursos para o seu uso en R.

Para empregar o matching existe un número considerable de librerías dispoñibles para o seu uso en R[2]. Desta forma, as librerías máis relevantes son MatchIt, Matching, twang, cem, optmatch, PSAgraphics, Synth, Cobalt, CBPS e ebal.

Neste caso centrámonos na librería MatchIt[2], que empregamos posteriormente para analizar o caso de CPI en Colombia. A librería MatchIt permite empregar diferentes metodoloxías e medicións da distancia con opcións adicionais, así como obter algunhas medidas e recursos gráficos para valorar o balance das covariables e obter o conxunto de datos emparellado (individuos de tratamento, control ou ambos).

No que respecta ás metodoloxías, a librería permite empregar Exact matching, Subclassification, Nearest neighbor matching, Optimal matching, Full matching e Genetic matching. Ademais, esta librería permite tamén empregar exact matching xunto ó Nearest neighbor matching, algo de utilidade se se emprega un número elevado de variables.

Adicionalmente, a librería recolle diferentes argumentos que permiten mellorar os resultados do matching. Concretamente, o argumento "discard" permite descartar individuos dun ou de ambos grupos, "reestimate" decide se reestimar o modelo para mellorar os resultados logo de descartar individuos. O argumento "m.order" que sinala a orde en que facer o emparellamento. O argumento "replace", que sinala se unha unidade de control pode ser asignada a máis dun individuo. O argumento "ratio" o número de individuos de control a asignar a cada individuo de tratamento. O argumento "caliper" o número de desviacións estándar que marca o rango no que un emparellamento é válido. Finalmente, o argumento "exact" permite introducir a lista de variables coa que empregar a distancia exacta.

II.3 O caso da CPI en Colombia.

Para exemplificar o uso desta metodoloxía e da librería en cuestión empregamos os datos de dúas enquisas, a "Encuesta de Desarrollo e Innovación Tecnológica" que ten dúas modalidades. Unha para a industria e outra para os servizos nos bienios 2017-18 e 2018-19. Desta forma, en ambos casos empregamos o primeiro ano como covariables previas á pandemia e o segundo como variables de resultado. Neste apartado non incidiremos en exceso sobre o filtrado dos datos, so introducimos a especificación do matching, as medidas do balance e os resultados.

Comezando co matching, cabe sinalar que empregamos a seguinte especificación:

```
> equation <- CPI ~ micro +peque +mediana +grande + GASTOID +SAT + SAMT + SBMT +SBT +SIC+ SNIC +SUBVENCIONID + IDTOTAL1 + PACTI +TRPI + IDINTERNO1 + PFOR1 + EXP + c_prov + c_gobierno + c_clientes
> m.out <- matchit(formula = equation, data, method = "nearest", ratio = 50,  distance = "probit", reestimate = TRUE, verbose = TRUE, replace = TRUE, discard = "both")
```

En canto ás medidas de equilibrio, esta especificación do matching amosa un bo balance nos principais indicadores empregados para medilo e na análise gráfica[3].

	distan-cia	micro	peque	mediana	grande	GASTOID	SAT	SAMT	SBMT
Antes	102,57%	-71,85%	-65,90%	-4,05%	66,14%	184,12%	0,79%	1,36%	-36,81%
Despois	2,6%	-8,5%	3,6%	2,7%	-3,6%	-6,2%	-2,5%	-6,1%	-2,9%
Antes	SBT	SIC	SNIC	SID	IDTOT	PACTI	TRPI	IDINTER	PFOR
Despois	-55,36%	94,32%	-63,52%	87,34%	40,33%	57,48%	13,27%	27,82%	36,81%
Antes	-4,0%	7,8%	-0,3%	-4,7%	3,0%	0,7%	1,4%	-2,7%	-0,2%
Despois	EXP	c_prov	c_gobierno	c_clientes	PERS*	VENT*	Sesgos modelo	Sesgos	
Antes	6,23%	55,81%	52,01%	48,42%	42,84%	15,61%	47,79%	45,32%	
Despois	-4,5%	-2,9%	4,1%	-0,2%	-1,8%	0,6%	3,68%	3,35%	
*indica que a variable non foi introducida no matching.									

Táboa 1: Medidas do balance, sesgos estandarizados por variable e total.

	Tratados	Control	Pseudo-R^2	Test de verosimilitude	Prob. Test verosimilitude
Antes	450	16089	0,3275	1.352,85	0,0000
Despois	450	4382	0,0035	19,43	0,8296

Táboa 2: Medidas do balance, individuos, pseudo-R^2 e test de verosimilitude.

En canto á análise gráfica, obtemos os seguintes resultados.

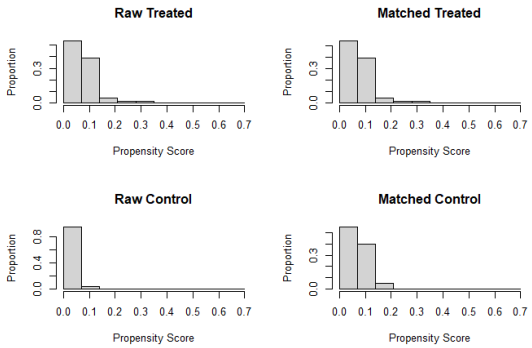


Figura 1: Análise gráfica do balance.

En canto ós resultados, a seguinte táboa recolle o impacto porcentual da CPI.

	Medias ponderadas		Impacto (diferencia %)	Desviación estándar		Test-t/Test-u
Variable	Tratados	Control		Tratados	Control	
Nuevos bienes y servicios	3,8533	1,5596	147,1%	24,41	5,03	23,3732***
Bienes y servicios mejorados	1,7422	0,7523	131,6%	3,37	2,50	1,8063***
Nuevos métodos de prestación	0,8400	0,6919	21,4%	2,01	1,42	1,9841***
Nuevos métodos organizativos	0,7556	0,5749	31,4%	1,94	1,18	2,6797***
Nuevas técnicas de comercialización	0,3978	0,3463	14,9%	0,96	0,81	1,4100***
Personal	621,28	638,93	-2,8%	1080,39	1163,37	0,8579***
Ventas (millones pesos colombianos)	229106,224	229244,563	-0,1%	1129827,36	975676,90	1,3330***
Peso exportaciones	0,0361	0,0428	-15,5%	0,11	0,12	0,7420***
Peso gasto I+D total sobre ventas	0,0105	0,0134	-21,8%	0,04	0,08	0,2632***
Peso gasto I+D interno sobre ventas	0,0360	0,0302	19,3%	0,08	0,10	0,6388***
Peso personal formado sobre personal (1)	0,0602	0,0618	-2,6%	0,13	0,13	1,1294***
Peso personal I+D sobre total	0,0919	0,0804	14,3%	0,12	0,12	0,9694***
Patentes obtenidas	2,7578	3,6629	-24,7%	12,08	16,72	0,5188***
(1) Significa a dicha variable se le aplica el test-u						
* p<0,1, **p<0,05,*** p<0,01						

Táboa 2: Medidas do balance, individuos*, pseudo-R^2 e test de verosimilitude*.

3. CONCLUSIÓNS

A CPI parece ter efectos significativos nas variables de resultados analizadas. O matching amósase como unha boa metodoloxía para estudar estes aspectos. Ademais R, e máis concretamente a librería MatchIt, amósanse como ferramentas útiles para aplicar a metodoloxía en cuestión múltiples como as políticas públicas e outros aspectos nos que os estudos experimentais non se poden levar a cabo.

Referencias

[1] Stuart, E. A. (2010). Matching methods for casual inference: A review and a look forward. *Statistical Science,* 25(1), 1-21. https://doi:10.2307/41058994

[2] Ho, D. E., Imai, K., King, G. e Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, 42(8) 1-28, https://doi.org/10.18637/jss.v042.i08

[3] Guerzoni, M. e Raiteri, E. (2015). Demand-side vs. supply-side technology policies: Hidden treatment and new empirical evidence on the policy mix. Research Policy 44, 726-747. https://doi.org/10.1016/j.respol.2014.10.009

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Comparativa entre medidas culturales en el contexto de la toma de decisiones turísticas

Yago Atrio Lema[1], Isabel Neira Gómez[2], Eduardo Sánchez Vila[3] y Paula Saavedra Nieves[4]

[1] Universidad de Santiago de Compostela, VALFINAP, ECOBAS
[2] Universidad de Santiago de Compostela, VALFINAP, ECOBAS
[3] Universidad de Santiago de Compostela, GSI
[4] Universidad de Santiago de Compostela, CITMAGA

RESUMO

La importancia de la cultura como elemento clave en la toma de decisiones turísticas ha sido objeto de numerosos trabajos en la literatura empírica. Nuestra hipótesis central de trabajo parte de la idea de que la cultura del país de origen del turista afecta a su proceso de toma de decisiones. Para probar nuestra propuesta, analizamos la elección de los peregrinos sobre la ruta para realizar el Camino de Santiago, comparando el poder explicativo de dos posibles variables: la distancia cultural (distancia entre valores individuales, distancia de Kogut y Singh, distancia de Jackson y la distancia de Kandogan), y el clúster cultural al que pertenece el país de origen del peregrino utilizando el modelo (Inglehart & Baker, 2000). Se ha empleado una metodología basada en modelos de elección discreta utilizando la formalización logística multinomial. Los resultados obtenidos validan la hipótesis principal analizada y abren un nuevo campo de estudio del impacto de la cultura de origen del viajero en su proceso de toma de decisiones turísticas.

Palabras e frases chave: Camino de Santiago, toma de decisiones, cultura, distancia cultural, clúster cultural.

1. INTRODUCIÓN

La cultura es un fenómeno complejo que se caracteriza como el conjunto de creencias, valores, conocimientos o normas aceptadas por un grupo social. Estas características pueden explicarse debido a factores geográficos, climáticos y socioeconómicos (Minkov & Hofstede, 2013; Van der Westhuizen et al., 2012).

El impacto de la cultura sobre el comportamiento de los turistas se ha analizado a fondo por medio de los métodos cuantitativos, a través de cuatro variables explicativas: la cultura interiorizada del turista, la cultura de la región de origen, la cultura de la región de destino, y la distancia entre la cultura de la región de destino y la cultura de la región de origen. Sin embargo, en la literatura de turismo relativa al proceso de toma de decisiones en la elección de alternativas,

como por ejemplo el destino, paquetes turísticos, etc…, sólo se ha estudiado en detalle el impacto de la distancia cultural, no prestando atención a otras alternativas de medición (Liu et al., 2018; Ng et al., 2007).

Para contrastar nuestras hipótesis utilizamos modelos logit multinomiales, en tres especificaciones (distancia cultural de Kogut y Singh, distancia cultural euclidiana y clúster cultural), siendo la variable a explicar el camino de Santiago elegido por los peregrinos. La computación de los modelos se efectuó a través del software R en su versión 4.2.1., utilizando el paquete mlogit en su versión 1.1.1. para el calculo de los modelos y la colección de paquetes de tidyverse para facilitar la tarea de la gestión de los datos con dplyr.

2. RESULTADOS RESUMIDOS

		Francés	Portugués	Inglés	Norte	Vía de la Plata	Portugués	Primitivo
Distancias	Kogut y Singh (Modelo 1)	-0.125 (0.002) ***	0.362 (0.006) ***	-0.413 (0.008) ***	-0.242 (0.006) ***	-0.390 (0.007) ***	0.348 (0.002) ***	-0.184 (0.006) ***
	Euclidiana (Modelo 2)	-0.197 (0.002) ***	0.545 (0.006) ***	-0.371 (0.006) ***	-0.150 (0.004) ***	-0.359 (0.006) ***	0.497 (0.002) ***	-0.277 (0.005) ***
Clúster (Modelo 3)	Sudeste asiático	0.121 (0.016) ***	1.151 (0.040) ***	-0.684 (0.051) ***	-0.488 (0.041) ***	-0.842 (0.055) ***	0.266 (0.020) ***	-0.517 (0.043) ***
	Europa católica	-0.468 (0.003) ***	0.906 (0.012) ***	-0.122 (0.009) ***	-0.139 (0.007) ***	-0.294 (0.008) ***	0.894 (0.004) ***	-0.290 (0.008) ***
	América latina	0.200 (0.007) ***	0.393 (0.025) ***	-0.925 (0.024) ***	-0.533 (0.017) ***	-0.876 (0.022) ***	0.245 (0.009) ***	-0.148 (0.015) ***
	Europa ortodoxa	-0.865 (0.014) ***	1.882 (0.029) ***	-0.393 (0.044) ***	0.530 (0.026) ***	-0.863 (0.055) ***	0.923 (0.016) ***	0.425 (0.028) ***
	Europa protestante	-0.252 (0.004) ***	0.806 (0.014) ***	-0.898 (0.014) ***	0.363 (0.008) ***	-0.256 (0.011) ***	0.583 (0.005) ***	-0.593 (0.012) ***
	Confucionistas	1.060 (0.012) ***	0.058 (0.039) **	-1.880 (0.051) ***	-0.526 (0.023)	-1.479 (0.042) ***	-0.742 (0.017)	-1.225 (0.035) ***
	Angloparlantes	0.319 (0.004) ***	0.939 (0.014) ***	-0.467 (0.012) ***	-0.419 (0.010) ***	-0.865 (0.014) ***	-0.001 (0.006)	-0.848 (0.013) ***

Tabla 1. Coeficientes de los modelos de elección analizados: Modelo I, con la "distancia cultural de Kogut y Singh, Modelo II, con la "distancia euclidea", y Modelo III con el clúster cultural al que pertenece el turista

3. CONCLUSIÓNS

Los resultados apoyan la veracidad de la primera hipótesis de que la distancia cultural afecta sobre la toma de decisiones de los peregrinos, debido a que esta variable es significativa tanto en el modelo 1 como en el 2. Sin embargo, el signo del coeficiente es diferente al esperado, dado que numerosos autores (Kandogan, 2012; Manosuthi et al., 2020) apuntaban a que una mayor distancia cultural debería afectar de manera negativa sobre la llegada de turistas, así como debería provocar una disminución de la probabilidad de elección de una determinada alternativa turística. Esto no es así en nuestra modelización, donde existen alternativas (Caminos de Santiago) donde la distancia cultural afecta positivamente a la elección de los mismos. Esta es una relación que merecerá la pena estudiar en subsiguientes análisis.

Los análisis realizados también permiten confirmar la segunda hipótesis, es decir, que pertenecer a un determinado clúster cultural afecta a la toma de decisiones de los peregrinos en base a los resultados. Esto se puede ver con la significación de las variables de clúster en cada elección del camino. Consideramos que la aceptación de esta hipótesis está relacionada con diversos factores socioculturales de cada grupo, como la cultura del esfuerzo, la popularidad del turismo de ruta, la espiritualidad, el gusto por participar en actividades de grupo.

A pesar de que como hemos visto en el comentario de los resultados y la primera parte de la discusión, los modelos tienen una capacidad predictiva similar, la principal diferencia entre la medición de la influencia de la cultura en la toma de decisiones de los individuos con los clústeres culturales y con la distancia cultural es que esta última permite considerar idénticos a países con valores culturales muy diferentes, mientras que esto no puede ocurrir en la primera.

Los resultados son contundentes en cuanto a la necesidad de incluir variables que caractericen la cultura de los turistas en los análisis de toma de decisiones, dado que se pudo observar que tanto la distancia cultural como la cultura de origen del peregrino, tuvieron un efecto significativo sobre sus elecciones y mejoran de modo significativo las predicciones sobre las alternativas de decisión. Además de estos resultados constatamos la necesidad de utilización de diferentes aproximaciones a las variables distancia cultural en los modelos de impacto de la cultura en la toma de decisiones, y más concretamente la posibilidad de incluir variables relativas al entorno de la cultura de la que procede el viajero y no sólo a la distancia con respecto al país de destino, variable que constituye la principal aportación de este estudio, así como la metodología empleada que considera variables relativas no sólo a las características del individuo, sino a las de la alternativa elegida.

Referencias

Inglehart, R., & Baker, W. E. (2000). Modernization , Cultural Change , and the Persistence of Traditional Values. *American Sociological Review*, *65*(1), 19–51.

Kandogan, Y. (2012). An improvement to Kogut and Singh measure of cultural distance considering the relationship among different dimensions of culture. *Research in International Business and Finance*, *26*(2), 196–203. https://doi.org/10.1016/j.ribaf.2011.11.001

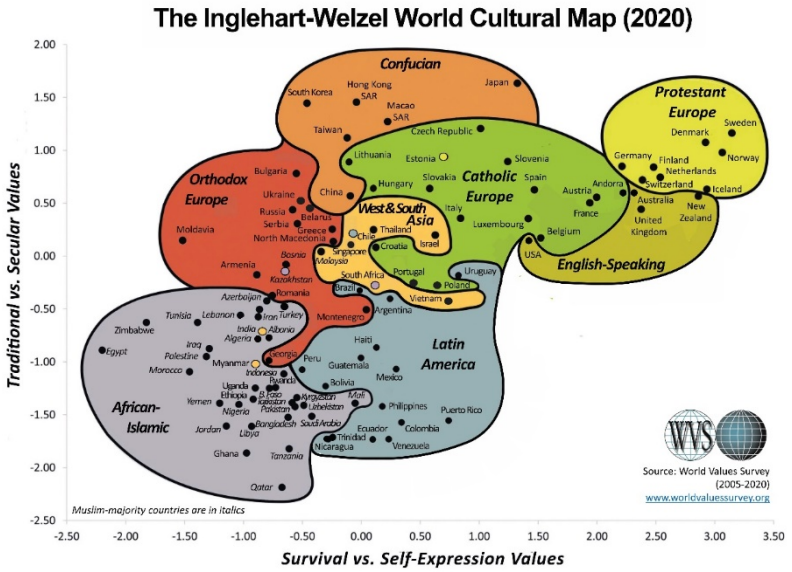
Liu, H., Robert, X., Cárdenas, D. A., & Yang, Y. (2018). Perceived cultural distance and international destination choice : The role of destination familiarity , geographic distance , and cultural motivation. *Journal of Destination Marketing & Management*, *9*, 300–309. https://doi.org/10.1016/j.jdmm.2018.03.002

Manosuthi, N., Lee, J., & Han, H. (2020). Impact of distance on the arrivals , behaviours and attitudes of international tourists in Hong Kong : A longitudinal approach. *Tourism Management*, *78*, 103963. https://doi.org/10.1016/j.tourman.2019.103963

Minkov, M., & Hofstede, G. (2013). *Cross-cultural analysis*. SAGE Publications Inc.

Ng, S. I., Lee, J. A., & Soutar, G. N. (2007). Tourists ' intention to visit a country : The impact of cultural distance. *Tourism Management*, *28*, 1497–1506. https://doi.org/10.1016/j.tourman.2006.11.005

Van der Westhuizen, D. W., Pacheco, G., & Webber, D. J. (2012). Culture, participative decision making and job satisfaction. *International Journal of Human Resource Management*, *23*(13), 2661–2679. https://doi.org/10.1080/09585192.2011.625967



IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

APLICACIÓN DE CONSULTA DE DATOS DO MARCO INPUT-OUTPUT DE GALICIA EN R-SHINY

Calvo Ocampo, Esther[1]

[1] Instituto Galego de Estatística

RESUMO

A presentación dun modo claro e comprensible dos datos difundidos é un dos desafíos dos produtores de estatísticas oficiais e, ao mesmo tempo, unha das boas prácticas recollidas no Código de conduta das estatísticas europeas.
O Marco Input-Output de Galicia (MIOGAL) constitúe unha ferramenta complexa pero fundamental para describir os fluxos económicos da nosa rexión, aínda que habitualmente dominou a súa vertente analítica a través das múltiples aplicacións da análise input output.
Non obstante, a información dispoñible nesta publicación ten un gran poder descritivo que se quixo potenciar no Instituto Galego de Estatística (IGE) co desenvolvemento da aplicación de consulta deseñada en R-Shiny que presentamos nesta ponencia. O obxectivo é ofrecer ao usuario unha visión amigable da información contida nunhas matrices de datos que, polo seu tamaño e pola metodoloxía coa que se constrúen, requiren certa especialización e coñecementos previos para o seu uso e a súa interpretación.
Nesta aplicación preséntase a información coa dobre visión que achegan as táboas de orixe e destino contidas no MIOGAL: por un lado a visión da rama de actividade e polo outro a visión dos produtos. É unha aplicación dinámica, que combina información en táboas e gráficos cos que o usuario pode coñecer a información das 72 ramas de actividade e os 110 produtos contidos na publicación.

Palabras e frases chave: R-shiny, marco input-output

1. RESULTADOS

Co fin de sintetizar e organizar a cantidade de información que ofrece o Marco Input-Output, desenvolveuse en R-Shiny unha aplicación que permite consultar a información por dúas vías: por rama de actividade e por produto.

Información por produto

No caso dos produtos, tratouse de condensar a información en forma de táboas e gráficos de forma que se amose o equilibrio oferta-demanda. Con este fin, incorporáronse nesta parte un tipo de gráficos que permiten estudar fluxos e que permitiron representar os fluxos da oferta e demanda de cada produto, os diagramas de Sankey. Estes diagramas representan fluxos, é dicir, conexións ponderadas que van dun nodo a outro.
No contexto do Marco Input-Output, o diagrama de Sankey empregouse para visualizar os fluxos da oferta e a demanda do produto. Os nodos fonte son neste caso as ramas de actividade produtoras do ben ou servizo (produto) considerado, incluíndo tamén os nodos correspondentes ás importacións do resto de España, importacións do resto da Unión Europea e importacións do Resto do Mundo.
Á súa vez, toda esta oferta do produto considerado agrúpase en dous nodos centrais,

que diferencian a procedencia da oferta do produto (importación ou interior).
Por último, os nodos de destino están formados polas ramas de actividade como demandantes do produto seleccionado así como polas restantes compoñentes da demanda (gasto en consumo final, formación bruta de capital, exportacións con destino ao resto de España, exportacións con destino ao resto da Unión Europea, exportacións con destino ao Resto do Mundo).
A librería que se empregou para crear este tipo de gráficos foi networkD3, htmlwidget que crea automaticamente gráficos interactivos.
Os datos de entrada pódense almacenar en 2 formatos diferentes:

- data frame coas conexións (3 columnas)
- matriz de incidencia (matriz cadrada)

Neste caso empregouse como formato dos datos de entrada un data frame, que enumera unha por unha as conexións e está formado por 3 columnas:

- source: nodo de orixe
- target: nodo de destino
- valor do fluxo

Construíuse este data frame a partir da matriz de orixe, a matriz de destino da produción interior a prezos básicos e a matriz de destino das importacións a prezos básicos. A primeira facilitará a información relativa á oferta e as dous segundas a relativa á demanda.
O data frame que se construíu coas conexións (links), está formado por 3 columnas: source, target y value, que se corresponden co nodo fonte, o nodo destino e a intensidade de cada fluxo, que ven dada neste caso polo valor.
A partir deste data frame cómpre crear tamén outro data frame de nodos no que se listan todos os nodos do diagrama.
Outro aspecto importante a ter en conta ao empregar a librería networkD3 é que as conexións deben proporcionarse mediante id, non mediante o nome que aparece no data frame coas conexións, polo que previamente á realización do gráfico será necesario crear un campo id para source (IDsource) e outro id para target (IDtarget).
Engadiuse ademais unha columna group tanto ao data frame links como ao nodes que permite utilizar unha cor distinta para cada grupo de fluxos ou para cada grupo de nodos.
Neste caso optouse por empregar unha único cor para todos os nodos polo que consideramos un único grupo e dúas cores no caso dos fluxos para distinguir o fluxo con procedencia na produción interior do fluxo con procedencia na importación.
A sintaxe que se empregou para a a construción do diagrama de fluxos foi a seguinte:

```
sankeyNetwork (Links = links, Nodes = nodes,
              Source = "IDsource", Target = "IDtarget",
              Value = "value", NodeID = "name",
              colourScale=my_color, LinkGroup="group", NodeGroup="group",
              fontSize = 10,
              nodePadding = 10, sinksRight=TRUE,iteration=0)
```

O paquete networkD3 crea automaticamente gráficos interactivos, o que permite na visualización web arrastrar nodos e colocar o cursor sobre as ligazóns para obter máis información.

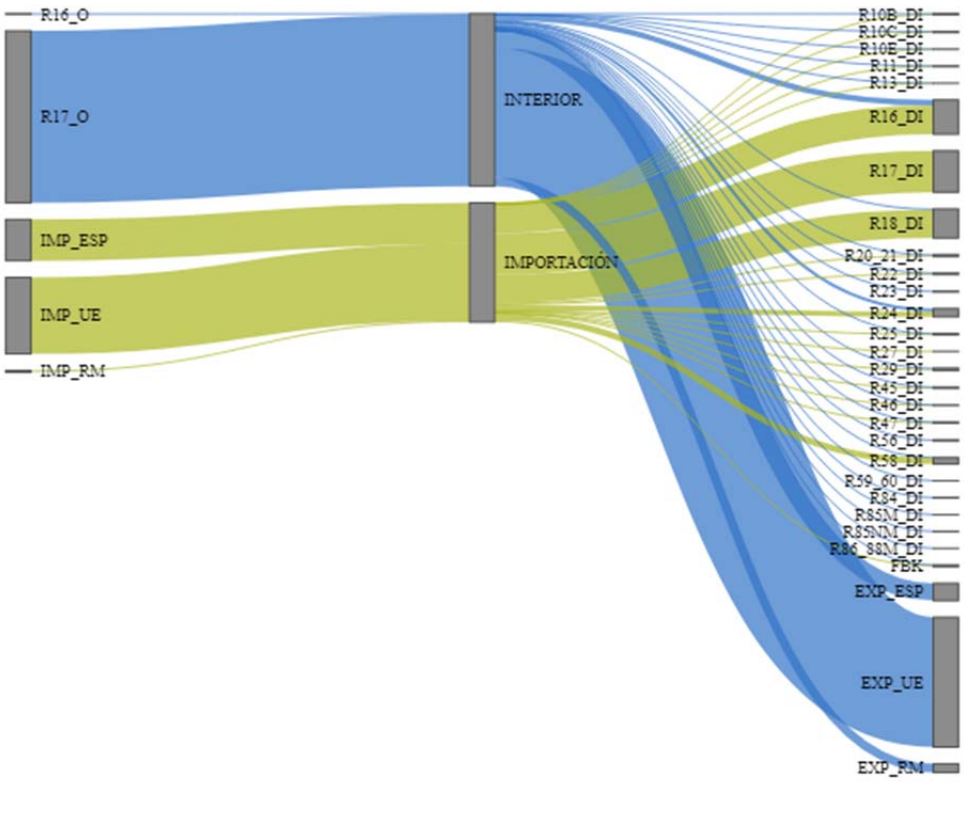


Figura 1: Diagrama de Sankey do produto 17A

Información por rama de actividade

Polo que respecta á información por rama de actividade, a novidade neste apartado é a utilización dos mapas de árbores xerárquicas ou treemaps.
Este tipo de gráficos permiten a comparación de cantidades e a visualización de patróns dalgunha estrutura xerárquica mediante un uso eficiente do espazo. Están compostos de rectángulos aniñados con áreas proporcionais aos rectángulos que representan e empréganse para analizar como se divide o todo e identificar rapidamente as compoñentes máis grandes e máis pequenas.
No contexto do Marco Input-Output, os gráficos treemap empregáronse para visualizar os consumos intermedios de cada rama de actividade. Neste caso, o tamaño dos rectángulos está asociado co peso do consumo intermedio de cada produto na rama de actividade e, por outra parte, a cor empregouse para diferenciar os produtos segundo a porcentaxe de consumo intermedio interior; é dicir, represéntanse para cada rama de actividade dúas variables: o peso do consumo intermedio por produto e a porcentaxe de consumo intermedio interior de cada produto.

Para implementar este tipo de gráficos empregouse o paquete treemapify, que permite crear treemaps en ggplot2. Cómpre utilizar geom_treemap () e especificar dentro de aes as variables. A sintaxe que se empregou foi a seguinte:

```
ggplot (df, aes (area = Peso, fill = `% CI Interior`,label = Codigo_completo)) +
  geom_treemap() +
  geom_treemap_text (colour = "white", place = "centre", grow = FALSE, reflow =TRUE,
min.size = 5) +
  scale_fill_gradient (low = '#99b9e3', high = '#0051ba')
```

onde df es un data frame formado por 3 columnas: produto, peso dese produto no consumo intermedio da rama e porcentaxe do consumo intermedio dese produto que é interior. A información contida no data frame é relativa neste caso a unha rama de actividade.
O algoritmo que geom_treemap () utiliza por defecto para colocar os rectángulos por defecto é o "squarified", que se basea nunha estratexia que busca que cada bloque sexa o máis cadrado posible para facilitar a comparación entre eles. Neste algoritmo, a colocación dos rectángulos procede dunha esquina, colocando os rectángulos en filas ou columnas ata que se colocan todos.

3. CONCLUSIÓNS

O Marco Input-Output de Galicia (MIOGAL) constitúe unha ferramenta complexa pero fundamental para describir os fluxos económicos da nosa rexión. Ten unha dobre vertente: descritiva e analítica. Sempre primou a analítica, como base da análise input-output, pero non é desprezable e debe promoverse a vertente descritiva.
Cada vez son máis frecuentes as situacións nas que nos atopamos con grandes cantidades de datos e cada vez faise tamén máis patente a necesidade de apoiarse en distintas ferramentas que permitan facilitar a comprensión da información. Cómpre favorecer a visualización de grandes cantidades de datos de xeito sinxelo e dunha sola ollada.
A incorporación na difusión do Marco Input-Output dos gráficos Sankey e dos treemap non deixa de ser un novo paso na procura de ferramentas que axuden a unha mellor comunicación e comprensión dos resultados estatísticos.
Non obstante, este tipo de visualizacións tampouco son sempre doadas de ler polo que a veces cómpre avaliar si un gráfico de barras ou un gráfico circular conta o mesmo aínda que visualmente sexa menos intenso.

Referencias

[1] Veiguela, N. (2016). R-Shiny: Una herramienta para mejorar la difusión de las operaciones del sistema de cuentas económicad de Galicia. XIX Jornadas de Estadística de las Comunidades Autónomas. Madrid

[2] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny
[3] David Wilkins (2021). treemapify: Draw Treemaps in 'ggplot2'. R package version 2.5.5. https://CRAN.R-project.org/package=treemapify
[4] J.J. Allaire, Christopher Gandrud, Kenton Russell and CJ Yetman (2017). networkD3: D3 JavaScript Network Graphs from R. R package version 0.4. https://CRAN.R-project.org/package=networkD3
[5] IGE : Marco input-output de Galicia.
https://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0307007003
[6] IGE: Aplicación gráfica de consulta do Marco Input-Output de Galicia 2016.
http://www.ige.eu/Shiny/MIOGAL/

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

R para o processamento de textos. Uma aplicação para a ordenação do texto em função da complexidade oracional

Afonso Xavier Canosa Rodrigues[1]

[1] IES Pedra da Auga, Ponteareas

RESUMO

Apresentamos um script em R que tokeniza um texto e extrai as orações para, de seguida, analisar o vocabulário e reorganizar as orações em função da sua complexidade medida no eixo sintático como longitude em número de tokens e no semântico conforme à ordenação do léxico segundo a primeira lei de Zipf

Palavras e frases chave: processamento de texto, sentence complexity, primeira lei de Zipf

1. INTRODUÇÃO

No trabalho com textos em que a língua apresenta dificuldades de compreensão, pode ser útil tentar reordenar o texto para ordenar as orações em função da sua complexidade. No exemplo usado para esta comunicação parto da própria experiência com o trabalho dum texto medieval em galês, língua céltica da Grã Bretanha. O corpus contém um número de termos de vocabulário desconhecidos, não só por possuirmos um vocabulário reduzido ao nível de aprendizagem, mas também por apresentarem variantes não facilmente recuperáveis num dicionário contemporâneo. A necessidade de organizar as orações em função da sua dificuldade propiciou uma série de experimentos para melhor afrontar a tradução do texto.

2. MATERIAIS

O procedimento descrito em baixo utiliza um script em R que permite replicar os experimentos e aplicá-los a qualquer outro documento com simplesmente modificar o nome do ficheiro no script[1]. Para as operações sobre o texto utilizaram-se os pacotes *stringr* [1] e t*okenizers* [2] de R. Como material complementário no repositório incluem-se os resultados da análise em formato csv, o texto utilizado como exemplo e um script adicional para a elaboração de gráficos e o estudo dalgumas propriedades do corpus.

3. PROCEDIMENTO

3.1. Vocabulário e frequências

O texto foi depurado de marcas de anotação (tags tipo XML), ficando um documento com apenas as palavras e sinais de pontuação. Os tokens são então agrupados em tipos únicos, considerados aqui como cada uma das palavras que aparece no texto, independentemente da sua frequência, sem lematização nenhuma. Isto é, cada variante gramatical de um mesmo lexema, caso, por

1 Os scripts e dados estão disponíveis no repositório: https://github.com/afonsoxavier/order_sentences.

exemplo, de mutações em início de palavra e morfemas gramaticais em posição final, é considerada um tipo único diferente.

Os tipos únicos são organizados de maior a menor frequência (o número de ocorrências no texto) criando uma tabela que responde ao critério de ordenação duma distribuição zipfiana cuja formulação tem expressão[3]:

$$P(r) = C / r^{\alpha}$$

em que P_r representa a frequência, *r* é dado pela ordenação decrescente das frequências, $C \approx 0.1$ e $\alpha \approx 1$. A sua representação gráfica tem a forma de uma curva monotonamente decrescente e uma linha reta com declive -1 ao realizarmos a sua transformação logarítmica.

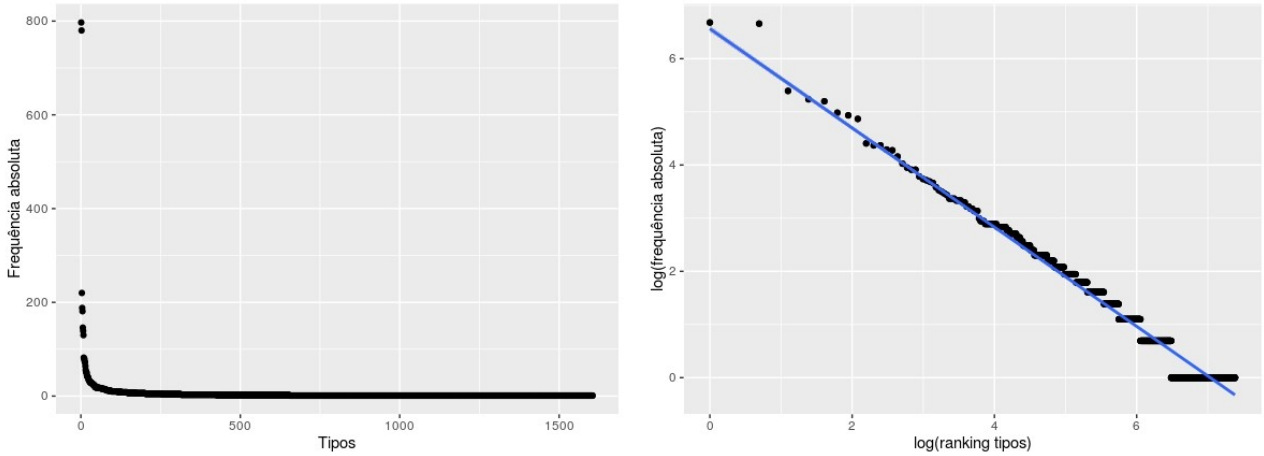


Figura 1:**:** Frequências dos tipos do corpus em ordem decrescente (esquerda) e transformação logarítmica com uma linha de regressão (direita), mostrando uma relação aproximada à expressada na primeira lei de Zipf.

O gráfico da figura 1 esquerda mostra a distribuição do corpus com os termos de maior frequência decrescendo em intervalos progressivamente menores até formar a linha dos *dislegomena* e a maior dos *hápax legomena*, dominante no eixo das ordenadas. As frequências aparecem com valores absolutos. A transformação das frequências absolutas representadas em função do logaritmo dos tipos ordenados mostra uma relação linear (fig. 4.2 direita), o desvio dos dados do corpus da linha de regressão aumenta nas frequências mais altas, confirmando que a lei de Zipf apenas supõe uma aproximação[4], não obstante, o conjunto da distribuição representa um comportamento observado repetidamente na análise de corpora [5].

3.2 A reordenação do texto

Uma vez segmentado o texto em orações e ordenados os tipos únicos numa distribuição zipfiana aplicamos uma série de métricas que combinam as frequências com a longitude da oração em número de tokens. A frequência relativa representa o peso do termo no vocabulário, isto é, quanto maior for o valor da frequência, mais relevante é o termo para a compreensão do texto. De outra parte, a longitude da oração assume-se como representativa da complexidade, uma oração com maior número de tokens implica um maior número de relações sintáticas envolvidas. O valor de uma oração obtém-se relacionando as frequências relativas dos tokens com a longitude da oração por meio da fórmula:

$$O = \frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^{w}}{t^{s}}$$

em que o valor duma oração *O* é obtido a partir de *t*, número de tokens da oração, a frequência absoluta do termo no corpus *f*, o total de tokens do corpus *T*, e os valores *w* e *s* outorgados numa escala de maior ou menor relevância da longitude da oração sobre o vocabulário. Enquanto *T*, *t* e *f* são variáveis empíricas, obtidas da análise do corpus, *w* e *s* relativizam o peso do léxico ou da sintaxe segundo as preferências do usuário com o objetivo de discriminar distintas ordenações do texto.

4. RESULTADOS E DISCUSSÃO

Da aplicação de distintos valores de escala sobre os dados do corpus obtiveram-se ordenações que foram classificadas em 5 grupos, segundo maior relevância do vocabulário ou da complexidade da oração e a combinatória de ambas.

Critério de ordenação	Orações no topo da reordenação
A) Frequência dos termos no contexto	**1.** A chyuodi y uynyd, a dodi y deudroet yn y got, a troi o Pwyll y got yny uyd Guawl dros y penn yn y got ac yn gyflym caeu y got, a llad clwm ar y carryeu, a dodi llef ar y gorn. **2.** Ef a aeth ryngtaw a llys Eueyd Hen, ac ef a doeth y r llys, a llawen uuwyt wrthaw, a dygyuor a llewenyd ac arlwy mawr a oed yn y erbyn, a holl uaranned y llys wrth y gynghor ef y treulwyt. **3.** Ac yskynuaen a oed odieithyr y porth, eisted gyr llaw hwnnw beunyd, a dywedut y pawb a delei o r a debygei nas gwyppei, y gyffranc oll, ac o r a attei idi y dwyn, kynnic y westei a phellynic y dwyn ar y cheuyn y r llys.
B) Frequência dos termos mais relevante do que a complexidade da oração	**1.**A hynny a wnaeth y makwyf. **2.**Y llys a gyrchyssant. **3.**A chyrch y llys.
C) Complexidade da oração mais relevante do que a frequência dos termos	**1.** heb y Pwyll. **2.** Y llys a gyrchyssant. **3.**A chyrch y llys.
D) Complexidade da oração muito relevante, frequência dos termos relevante	**1.** heb ef. **2.** heb y Pwyll. **3.** heb hi.
E) Complexidade da oração com a maior relevância	**1.** heb ef. **2.** heb hi. **3.** heb wy.

Tabela 1. Classificação das reordenações do texto e primeiras orações no topo de cada reordenação

A tabela 1 mostra os resultados obtidos em 5 reordenações do texto obtidas sempre como os mesmos dados empíricos do corpus (frequências e longitude em tokens) pela alteração do valor da escala na relação entre as variáveis *w* e *s*. Assim, na primeira ordenação (grupo A na tabela 1), as orações no topo são as que oferecem maior contexto para a compreensão dos termos e, dentro destas, procura-se que os termos tenham os maiores índices de frequência no corpus, de outra parte a complexidade da oração tem uma relevância mínima. No extremo oposto, a complexidade da oração com a maior relevância (grupo E na tabela 1), prioriza as orações mais simples. Porém, quando as orações têm igual complexidade, aquelas cujos termos apresentam maior frequência (aparecem, portanto, em mais orações) obtêm uma posição mais alta na reordenação. De facto, uma oração com apenas um termo (n. 4 no grupo E), fica abaixo de orações com dois termos, enquanto as primeiras estão compostas pelas unidades com maior frequência no corpus, no topo da curva zipfiana, frente a uma oração com um só termo que haveria que procurar na linha alargada de hápax da distribuição (figura 1). As ordenações intermédias (B, C, D) procuram um maior balanço entre a relação frequência do vocabulário e complexidade da oração, na pretensão de otimizar o processamento tanto da gramática quanto do léxico num processo de aprendizagem em que procuramos orações simples (mas não necessariamente as mais simples) com os termos mais comuns.

Os resultados pretendem ser simplesmente demonstrativos das possibilidades de ordenação. A classificação das distintas reordenações tem necessariamente um carácter subjetivo, enquanto as variáveis de peso (*w* e *s*) têm um valor não empírico e não se estabeleceu nenhuma medida avaliativa externa que considere a adequação dos resultados a uns objetivos concretos. Como trabalho futuro nesta linha, fica obter valores empíricos para *w* e *s*, por exemplo mediante a análise de ordenações feitas a priori. A mesma fórmula empregada neste experimento pode ser simplificada mediante o uso de novas variáveis a partir de reordenações usadas como treino, aplicando modelos de aprendizado de máquina.

A complexidade pode ser medida nos diversos componentes gramaticais. Particularmente relevante para a diversidade lexical é a morfologia. No âmbito da sintaxe, termos comuns e frases simples podem estabelecer relações mais complexas quando, por exemplo, requerem uma posição fixa na oração. A ampliação de observações gramaticais, particularmente aquelas obtidas automaticamente com ferramentas de PLN, contribuiria a uma apreensão mais abrangente da complexidade do texto.

5. CONCLUSÕES

Apresentamos aqui o caso de um texto que, pela dificuldade de interpretação com vista a um trabalho de tradução, se considerou conveniente reordenar para assim obter aquelas orações mais acessíveis segundo critérios de dificuldade semântica e sintática. Com apenas duas variáveis quantitativas, as frequências dos termos ordenadas numa distribuição zipfiana e a longitude da oração, aprecia-se a maior ou menor relevância da complexidade sintática em relação com o léxico segundo preferências definíveis pelo usuário. Ainda sendo representativas apenas do essencial do significado pelo contexto e a complexidade sintática, obtêm-se resultados mediante um procedimento unicamente quantitativo, que não requer a consideração de variáveis qualitativas empregadas tradicionalmente na análise semântica e morfossintática e portanto simplificam enormemente o processo de tratamento do texto.

Referências

[1] Wickham, H. (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. http://stringr.tidyverse.org, https://github.com/tidyverse/stringr.

[2] Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, *3*(23), 655.

[3] Li, W. (1991). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, *38*(6), 1842-1845

[4] Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112-1130.

[5] Kornai, A. (2008). *Mathematical Linguistics*. London: Springer.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

New covariates selection method in dynamic regression models with a public implementation in R language

Ana Ezquerro[1], Germán Aneiros[2], Manuel Oviedo[3]

[1]University of A Coruña, ana.ezquerro@udc.es
[2]CITIC, Grupo MODES, Departamento de Matemáticas, University of A Coruña, german.aneiros@udc.es
[3]CITIC, Grupo MODES, Departamento de Matemáticas, University of A Coruña, manuel.oviedo@udc.es

Abstract

This work introduces a new approach in time-series analysis field for automatic covariates selection in dynamic regression models. Based on [1] and [2] previous study, a forward-selection method is proposed for adding new significant covariates from a given set. This algorithm has been implemented and optimized in R as a package, and we openly publish its sources in order to make it available for all the R community. Our method has been applied to multiple simulations to validate its performance. Finally, the obtained results from the IRAS database of Catalonia are presented to analyze the COVID-19 evolution.

Keywords: Time series, dynamic regression models, selection methods, forecasting.

Introduction

In time-series analysis, the well-known dynamic regression models allow formally modelling the dependence between a set of covariates and a dependent variable considering the intrinsic temporal component of all participant variables. Thus, this type of regression models are of widespread application in diverse scenarios where it is desired to analyze the effect of recollected data in a time series of interest.

Formally, dynamic linear regression models define the linear dependence between a stochastic process Y_t (the dependent variable) and a set of processes $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ (candidates for regressor variables) in times non-greater than t:

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \cdots + \beta_m X_{t-r_m}^{(m)} + \eta_t \tag{1}$$

where $r_i \geq 0$, for $i = 1, ..., m$, and $\eta_t \sim$ ARMA(p,q).

In this work we formally introduce a new algorithm to select covariates which significantly influence the behavior of a dependent variable. The implementation of this selection method is publicly available[1].

Methodology

Following the definition in 1, [1] proposed a method named *prewhitening* for removing spurious correlation (false linear correlation) between two processes X_t and Y_t (where one of them is not white noise and/or the other is not stationary) by analyzing the cross correlation function $\rho_k(\ddot{X}_t, \ddot{Y}_t) = \dfrac{\text{Cov}(\ddot{X}_t, \ddot{Y}_{t-k})}{\sigma_{\ddot{X}_t}\sigma_{\ddot{Y}_t}}$ where σ_{Z_t} denotes the standard deviation of a stochastic process Z_t

[1] https://github.com/anaezquerro/dynamic-arimax

and $\ddot{X}_t$ and $\ddot{Y}_t$ are obtained via some linear filter application to X_t and Y_t ensuring one of them is white noise and the other is a stationary process. Specifically, [1] proposes a real linear correlation between X_t and Y_t if exists some k where $\rho_k(\ddot{X}_t, \ddot{Y}_t)$ is statistically significant. This method is applied to obtain the optimal lags of each regressor in 1, considering the condition of k being less or equal than 0.

Our approach iteratively adds dependent processes to a model by checking if a significant correlation (as in [1]) exists between a new process (candidate for regressor variable) and the residuals η_t of a simpler model.

Let Y_t be the stochastic dependent process and $\mathcal{X}$ be the set of processes that might act as regressor variables in the model (candidates), and an information criterion (IC) for model evaluation. Our method proceeds as follows:

1. Initialization. Consider the process $\tilde{Y}_t = Y_t$ that will be used to check the existence of linear correlation between Y_t and each $X_t \in \mathcal{X}$ with [1] method, $\nu = \infty$ the value of the IC corresponding to the best model with 1 form, $\mathcal{X}^{(s,r)}$ the set of selected covariates paired with their respective optimal lags and $\mathcal{X}^{(s)}$ the set of selected covariates (with no lag information). Let $\mathcal{M}(\mathcal{Z})$ be the fitted dynamic regression model regarding Y_t where $\mathcal{Z}$ is the set of covariates paired with their optimal lags:

$$\mathcal{M}(\mathcal{Z}) := Y_t = \beta_0 + \sum_{(Z_t,r)\in\mathcal{Z}} \beta^{(Z_t,r)} Z_{t-r} + \eta_t$$

   where $\beta^{(Z_t,r)}$ is obtained via some estimation.

2. Iterative selection. For each $X_t \in \mathcal{X} - \mathcal{X}^{(s)}$, obtain the optimal lag where the maximum linear cross correlation between X_t and $\tilde{Y}_t$ occurs (via [1] method). Consider the process $X_t^{\text{best}} \in \mathcal{X} - \mathcal{X}^{(s)}$ that minimizes and improves ν value, based on the selected IC, by including it in the model with its optimal lag ($r_{X_t^{\text{best}}}$):

$$X_t^{\text{best}} = \underset{X_t\in\mathcal{X}-\mathcal{X}^{(s)}}{\arg\min} \left\{ \text{criteria}\left(\mathcal{M}\left(\mathcal{X}^{(s,r)} \cup \left\{(X_t, r_{X_t})\right\}\right) \right) \right\} \tag{2}$$

   conditioned to criteria$(\cdot) < \nu$[2]. If X_t^{best} exists, consider $\mathcal{X}^{(s,r)} = \mathcal{X}^{(s,r)} \cup \{(X_t^{\text{best}}, r_{X_t^{\text{best}}})\}$, $\tilde{Y}_t = \eta_t$ and $\nu = \text{criteria}(\mathcal{X}^{(s,r)})$[3]. Repeat this step until no process $X_t \in \mathcal{X} - \mathcal{X}^{(s)}$ can be added to the model, i.e. X_t^{best} does not exist.

3. Finalization. If the errors η_t of $\mathcal{M}(\mathcal{X}^{(s,r)})$ are not stationary and no model with $\eta_t \sim$ ARMA(p,q) and $\mathcal{X}^{(s,r)}$ covariates can be adjusted, consider the regular differentiation of all data (dependent variable and regressor candidates) and return to (1). Otherwise, it is proven that $\mathcal{M}(\mathcal{X}^{(s,r)})$ with stationary errors defines the significant correlation between the set of $\mathcal{X}^{(s)}$ regressor variables and the dependent process Y_t.

This algorithm was implemented in R programming language. The step 2 was optimized by parallelizing the fit of independent models of each candidate in $\mathcal{X}$. Dickey-Fuller test is used for checking processes stationary, Ljung-Box to check the independence, Shapiro-Wilks and Jarque-Bera tests for normality and t-test for zero mean of ARIMA residuals.

Simulation results

In order to validate the performance of our selection method, we simulate multiple scenarios where a time series Y_t was artificially constructed with other variables (introduced with their respective coefficients and lags as in 1), which were added to a set of candidates along with more variables which do not influence in the construction of Y_t. The algorithm was tested when the residuals of the model η_t were stationary and non-stationary.

Specifically, we simulate $M = 100$ times the following scenario:

[2] for simplicity, we denote the expression in criteria() in 2 as $\cdot$

[3] once X_t^{best} has been added to the model

Figure 1: Example of code output and results of `drm.select()` when running the selection method

```
beta0 <- -0.6; beta1 <- 1.7; beta2 <- -2.2; beta3 <- 1.3; r1 <- 2; r3 <- 3
Y <- beta0 + beta1*lag(X1,-r1) + beta2*X2 + beta3*lag(X3,-r3) + residuals
xregs <- cbind(X1, X2, X3, X4, X5, X6)
ajuste <- drm.select(Y, xregs, ic='aicc', st_method='adf.test', show_info=F)

print(ajuste$history, row.names=F)

 var lag                ic
  X2   0 -1156.68486061937
  X1  -2 -2171.66958134745
  X3  -3 -3108.15443209894

print(ajuste, row.names=F)

Series: serie
Regression with ARIMA(0,0,4) errors

Coefficients:
         ma1     ma2  ma3     ma4  intercept       X2      X1      X3
      0.2498  0.3360    0  0.1589    -0.5947  -2.1868  1.6949  1.3083
s.e.  0.0304  0.0302    0  0.0300     0.0033   0.0105  0.0089  0.0320

sigma^2 = 0.002377:  log likelihood = 1562.15
AIC=-3108.3   AICc=-3108.15   BIC=-3069.26
```

1. We generate seven different independent time series (each modelable by an ARIMA), of which six of them act as the covariate candidates set: $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(6)}\}$; and the remaining as the residuals η_t of the model.

2. We construct the dependent variable Y_t by a linear combination of $\{X_t^{(1)}, X_t^{(2)}, X_t^{(3)}\}$, randomly lagged $r = 0, ..., 6$ moments (where the coefficients are randomly generated), with an intercept β_0 and the generated residuals η_t. Formally,

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \beta_3 X_{t-r_3}^{(3)} + \eta_t$$

   where $\beta_0, ...\beta_3$ are randomly generated, and $r_i \in [0, 6]$ for $i = 1, 2, 3$.

3. We launch our selection method with different configurations:

   - Using as the information criterion the AIC, BIC and AICc.
   - Using as the method to check stationary the Dickey-Fuller test or via analyzing the differentiation order when an ARIMA is adjusted.

4. Evaluate the selection method using as metrics the percentage of times a covariate is:

   (a) correctly added to the model (*true positive*),

   (b) incorrectly added to the model (*false positive*),

   (c) correctly not added to the model (*true negative*),

   (d) incorrectly not added to the model (*false negative*).

Figure 1 displays the result of calling the function that implements the selection method. The DataFrame stored in `$history` provide information about the covariates iteratively added to the model, the IC value achieved and the lag they were added with. When printing the resultant object (`ajuste`) we see that the estimated regression coefficients are nearly the same than the real values artificially set. Also, the lags estimated by the method are correct, and the errors of the final model are stationary.

Table illustrates a resume of the results of our approach when running our approach $M = 100$ times with different configurations and using stationary an non-stationary errors.

Table 1: Percentage data results with different configurations when residuals are stationary

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	97.66%	97.66%	97.66%	3.66%	1.33%	3.66%
auto.arima	98.33%	98.33%	98.33%	3.66%	1.33%	3.66%
	correctly added (TP)			incorrectly added (FP)		

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	96.33%	98.66%	96.33%	2.33%	2.33%	2.33%
auto.arima	96.33%	98.66%	96.33%	1.66%	1.66%	1.66%
	correctly **not** added (TN)			incorrectly **not** added (FN)		

Table 2: Information about the dynamic regression model constructed via selection of multiple vaccination variables to model COVID19 evolution

Covariate	**Lag**	**Coefficient est. (s.e)**
`vac4565`	-3	-0.0410 (0.0057)
`vac6580`	-2	-0.0468 (0.0120)
`vac1845`	-6	-0.0901 (0.0047)
`vac1218`	Not included in the model	
`vac80`	Not included in the model	
residuals	ARIMA(4, 0, 0)	$\phi_1 = 2.0816(0.0810)$ $\phi_2 = -1.2837(0.1152)$ $\phi_4 = 0.1919(0.0432)$

Application to COVID19 evolution

Due to the impact of COVID-19 around the world, we use this method to formalize and study the relation of the COVID-19 evolution in Catalonia (Spain) with the flu syndrome, COVID-19 vaccination and other recollected variables from the IRAS database. Individual data was aggregated by age ranges and Health Areas to study the correlation between groups and their influence in the global evolution.

Table 2 resumes the algorithm trace and the order of covariates addition to the model. The covariates named `vac1218`, `vac1845`, `vac4565`, `vac6580` correspond the vaccination data in population from 12, 18, 45 and 65 up to 18, 45, 65 and 80 years old (exclusive), and `vac80` corresponds the vaccination in population from 80 years. We can analyze the vaccination has a negative impact in the expansion of COVID19, specifically, the vaccination of working-age population.

Future work

Our approach has considered DRM covariates modelable by ARIMA models, which successfully covers a wide real-life applications. However, other cases might be considered, such as adding functional variables and discrete variables to the set of candidates.

Acknowledgements

To Banco Santander for the scholarships offered in 2021/2022, which helped the investigation of this proposal, and to *Maths Department* of University of A Coruña.

References

[1] Jonathan D Cryer and Kung-Sik Chan. *Time series analysis: with applications in R*, chapter 11. 2, 2008.

[2] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*, chapter 9. OTexts, 2018.

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Con R de fRikismo. *Learning VS Knowing, the match of the century.*

Álvaro Fernández Theotonio[1]

[1] Universidade de Santiago de Compostela

RESUMO

El presente proyecto nace por el deseo de aplicar los conocimientos del lenguaje estadístico "R", adquiridos durante el Máster en Técnicas Estadísticas, a una importante esfera del desarrollo humano: el ocio.

"Android: Netrunner"(FFG, 2012) es un juego de cartas para dos jugadores ambientado en un futuro distópico en el que grandes corporaciones, dispuestas a cualquier cosa para cumplir sus planes, deben proteger sus valiosos datos de hackers de élite conocidos como "*runners*"(FFG, 2012). Más allá de la temática, en dónde reside el interés por este juego es en su mecánica, tópico en el que se centrará este estudio.

La mecánica se fundamenta en el robo y manejo de cartas de un mazo de longitud variable y en la gestión de recursos. Tratándose de una mecánica de robo de cartas de un mazo boca abajo, el componente probabilístico asociado a obtener cartas concretas cobra una vital importancia de cara a la optimización de la toma de decisiones.

En "Android: Netrunner" los jugadores disponen de 2 tipos de recursos: los "créditos" y las cartas. Los créditos son la moneda del juego, el "dinero", mientras que las cartas son efectos que para ser jugados han de ser pagadas con créditos. Los efectos producidos por estas cartas dan lugar a cambios en el estado del juego, que, a priori, incrementan las probabilidades de victoria del jugador que las usa. A este respecto, es importante resaltar que el potencial de las cartas es variable (Waddell, 2022), ya que las ganancias obtenidas por el uso de una carta determinada son dependientes del contexto en el que son utilizadas, como se puede observar en el siguiente ejemplo.

Una carta que cuesta 2 créditos usar cuyo efecto es proporcionar 1 crédito por turno, tendrá un valor elevado si es jugada en los momentos iniciales de la partida, mientras que de hacerlo al final está no tendrá tanto valor.

La estructura del juego se basa en la sucesión alterna de turnos entre los 2 jugadores hasta que alguno de ellos obtenga la victoria. En cada turno el jugador dispone de 4 acciones, que puede invertir en diferentes efectos. *Nota: debido a que los únicos efectos relevantes para el presente proyecto son los asociadas al robo de cartas, se omitirán el resto de efectos posibles.*

Al comenzar la partida, cada jugador tiene un mazo y una mano inicial de 5 cartas. A partir de este momento, los jugadores puedan robar cartas de su mazo "gastando" una de sus acciones disponibles a través de una de las dos siguientes posibilidades:

- Robar una carta (conocido como "acción básica de robo")
- Usar una carta que permita robar un número de cartas determinado.

Mientras que la obtención de cartas en el caso de la acción básica de robar permite conseguir solo una carta, el usar cartas que permitan robar resultará siempre más eficiente, ya que el número de cartas obtenidas por este método es siempre mayor que el número obtenido a través de la acción básica de robo.

Como se adelantó previamente, el beneficio obtenido del uso de cartas está íntimamente relacionado con el momento en el que éstas son utilizadas. Existen cartas que, en el caso de ser jugadas durante el primer turno de juego desarrollan su máximo potencial y producen el mayor incremento en términos de probabilidades de victoria (*win-rate*)(Waddell, 2022). En base a esta idea, resulta de capital importancia maximizar las probabilidades de poder utilizar esas cartas durante el primer turno, con el fin de obtener los mejores resultados durante una competición. Es precisamente en este instante donde el software estadístico R entra en la ecuación.

Mientras que calcular las probabilidades de obtener una determinada carta del mazo en la mano inicial (5 cartas) resulta muy simple, calcular las probabilidades de obtenerla durante el primer turno presenta una elevada complejidad. Esto es así debido a la variabilidad existente entre las cartas que

se pueden utilizar para robar cartas, y a las diferentes decisiones que se pueden tomar. Para ilustrar esta idea, podemos observar el siguiente ejemplo:

> *Queremos calcular las probabilidades de tener en la mano una carta llamada "Rezeki"al final del primer turno. Se asume que el jugador "A" va a tomar las decisiones que incrementen al máximo las probabilidades de obtención de esta carta. Primero, "A" recibirá su mano inicial de 5 cartas, en las que puede estar ya incluida "Rezeki", o no. De no estar incluida, "A" tratará de robar el máximo número de cartas con el fin de maximizar probabilidad de obtener la carta deseada. De esta manera, sabemos que en el peor de los casos, este jugador robará 4 cartas más, ya que dispone de 4 acciones y, en el caso de no disponer de efectos más eficientes siempre puede invertir una acción en robar una carta. Sin embargo, también es posible que "A" disponga de efectos en su mano que produzcan un mayor robo de cartas que el proporcionado por la acción básica (por ejemplo la carta "Diesel" permite robar 3 cartas por una acción). De esta manera, siempre que le sea posible, "A"utilizará una carta de este tipo en vez de la acción básica. Debido a esto, el número de cartas al que "A" es capaz de acceder resulta variable.*

Esta última cuestión produce un incremento notorio en la complejidad del problema, ya que no solo resulta necesario evaluar si hemos obtenido (o no) la carta deseada, sino también el grado en el que las cartas obtenidas incrementan las probabilidades de conseguir la carta deseada. Para solventar esta cuestión, se pretende a través del lenguaje estadístico "R" obtener, mediante simulación, las probabilidades asociadas a diferentes configuraciones del mazo de cartas, con el fin de evaluar en qué medida cada mazo es capaz de robar sus cartas importantes y proporcionar un indicador de su nivel de robo.

Palabras e frases chave: Simulation, sampling, card-game, Netrunner

Referencias

Waddell, J. (2022, 16 de junio). *Android: Netrunner Full Review*. https://riptidelab.com/android-netrunner-full-review/

Fantasy Flight Games [FFG] (2012). *Android: Netrunner the Card Game*. https://www.fantasyflightgames.com/en/products/android-netrunner-the-card-game/

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Construyendo Mandalas con R/Construindo Mandalas com R

Luciane Ferreira Alcoforado[1]
[1] Academia da Força Aérea/Divisão de Ensino

RESUMO

A relação entre a Mandala e a matemática encontra-se no seu significado que vem do sânscrito e significa círculo. As mandalas são compostas por figuras geométricas que se repetem através de transformações matemáticos como isometrias e homotetias.
A partir das equações parametrizadas de uma seleção de curvas clássicas da matemática, combinados com as transformações matemáticas, é possível construir diversas formas de mandalas. Como recurso computacional a linguagem R.

Palabras e frases chave: curvas paramétricas, transformações, mandala, programação, linguagem R.

1. Introdução

Neste trabalho apresenta-se as equações parametrizadas de curvas clássicas da matemática, como gerar os pontos do eixo cartesiano e como programá-las no pacote ggplot2, Wickham (2016) da linguagem R, empregando transformações matemáticas com o objetivo de construir mandalas.
De acordo com Bezerra (2022), a mandala é, originalmente, um círculo que contém em seu interior desenhos de formas geométricas, figuras humanas e cores variadas e podem ser empregadas na educação como um recurso didático utilizado por professores de matemática, pois este símbolo serve para ensinar vários tópicos como geometría analítica, representação gráfica e lógica de programação.
Apresenta-se inicialmente as equações paramétricas de curvas como o círculo. Uma equação paramétrica, de acordo com a definição de MathWorld (2022), é um conjunto de equações que expressam um conjunto de pontos como funções explícitas de uma série de variáveis independentes, conhecidas como "parâmetros". A partir da equação paramétrica de uma curva, utiliza-se processos de transformações matemáticas gerando-se assim a mandala.

2. Transformações Matemáticas

Definição 1: Uma transformação no plano é uma função bijetora do conjunto dos pontos do plano sobre si mesmo.
Definição 2: Isometrias são transformações no plano que preservam distâncias, isto é, se T é uma isometria, para qualquer par de pontos A e B vale a relação $T(A)T(B) = AB$.

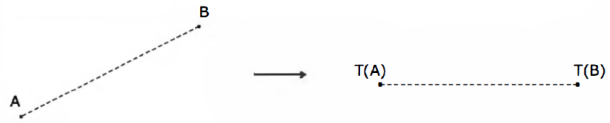


Figura 1: Isometria. Fonte: Rezende & Queiroz (2008).

Definição 3: Sejam A e B pontos distintos do plano. A translação T_{AB} é a isometria que leva um ponto X do plano ao ponto $T_{AB}(X) = X'$, tal que $ABX'X$ é um paralelogramo, se A, B e X não são colineares. Se A, B e X são colineares, então T_{AB} é tal que XX' está na reta

AB e os segmentos *AX'* e *BX* têm o mesmo ponto médio.

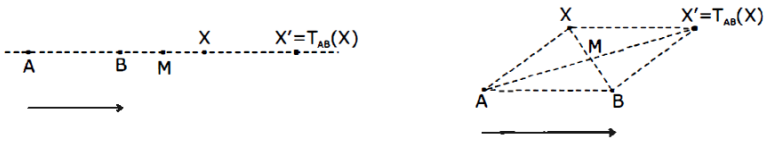


Figura 2: Translação. Fonte: Rezende & Queiroz (2008).

Considerando um ponto X de coordenadas (x,y), a translação para o ponto X´ de coordenadas (x´,y´) é obtido por
x´= x + a, a ∈ R; y´= y + b, b ∈ R
Definição 4: Seja O um ponto do plano e t um número real com $t \in [0,2\pi]$. A rotação de centro O e ângulo t é a isometria I_{Ot} que deixa fixo o ponto O e leva o ponto X, X ≠ O, no ponto X' = I_{Ot}(X), tal que OX = OX' e a medida do ângulo orientado (OX, OX') é igual a t, se t≠0 e t≠π. Além disso, OX' = OX, sendo O o ponto médio de XX', se t≠π; e X'= X se t =0.

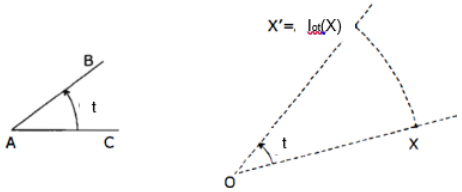


Figura 3: Rotação. Fonte: Rezende & Queiroz (2008).

Considerando um ponto X de coordenadas (x,y), a rotação por um ângulo t no sentido anti horário para o ponto X´ de coordenadas (x´,y´) é obtido por
x´= xcos(t) – ysin(t), t ∈ [0,2π]; y´= xsin(t) + ycos(t), t ∈ [0,2π]
Enquanto duas figuras isométricas têm a mesma forma e as mesmas dimensões, isto é, são congruentes, duas figuras homotéticas conservam apenas a mesma forma. Dizemos que duas figuras homotéticas são semelhantes.

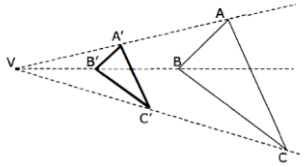


Figura 4: Homotetia de dois triângulos ABC e A'B'C'. Fonte: Rezende & Queiroz (2008).

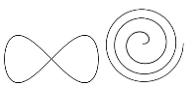
Considerando um ponto X de coordenadas (x,y), a redução por um fator k do ponto X´ de coordenadas (x´,y´) é obtido por
x´= kx, k ∈ (0,1); y´= ky, k ∈ (0,1)

3. Equações Paramétricas de curvas clássicas

Equações paramétricas podem ser obtidas em Stover & Weisstein (2022):
Circulo: $x = rcos(t); y = rsin(t)$, $t \in [0,2\pi]$; Elipse: $x = acos(t); y = bsin(t)$, $t \in [0,2\pi]$

Cardióide: $x = 2r \cdot cos(t) - r \cdot cos(2t); y = 2 \cdot rsin(t) - r \cdot sin(2t)$, $t \in [0,2\pi]$; Limaçon: $x = (b + a \cdot cos(t)) \cdot cos(t); y = (b + a \cdot cos(t)) \cdot sin(t)$, $t \in [0,2\pi]$

Lemniscata: : $x = sin(t); y = \cos(t) . sin(t)$, $t \in [0,2\pi]$; Espiral de Fermat: $x = a\sqrt{t}.cos(t)$; $y = a\sqrt{t}.sin(t)$, $t \geq 0$



4. Programando em R

Considerando as equações paramétricas das curvas clássicas em conjunto com as transformações matemáticas de isometria e homotetia, passaremos ao script na

linguagem R para construção e exibição das mandalas.
O procedimento está decomposto em duas partes, na primeira parte elabora-se a criação de uma base de dados contendo todos os pontos das coordenadas (x,y) que formam o desenho da mandala. Na segunda parte elabora-se a visualização do desenho utilizando-se o pacote ggplot2. O desenvolvimento dos scripts foi realizado com base em Alcoforado (2021).
Mandala 1: Elaborada a partir do círculo com translação e rotação
Inicialmente criamos em x e y os pontos que formam um círculo de raio 1, posteriormente criamos em xt e yt a translação dos pontos iniciais apenas no sentido do eixo x, figura 5:

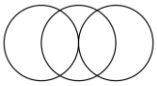


Figura 5: Resultado da translação do círculo na direção do eixo x. Fonte: autora, 2022.

Posteriormente é realizado a rotação da figura 5 nos ângulos de $\frac{\pi}{8}, \frac{\pi}{4}, \ldots, 2\pi$, tais pontos são armazenados em um *dataframe* denominado dt. O número de linhas em dt é formado por 25500, ou seja, 25500 pontos a serem plotados.
Script – Parte 1 gerando pontos e armazenando em dt

```
#Parâmetros
n=500; raio=1; t=seq(0,2*pi, length.out = n)
#pontos para círculo inicial
x=raio*cos(t)
y=raio*sin(t)
#pontos para os 3 círculos
xt=c(x,x-raio,x-2*raio)
yt=c(y,y,y)
```

```
rotacao = (pi/8)*(1:16); n=length(xt); xt1=xt; yt1=yt
for(i in 1:length(rotacao))
  {
xt1=c(xt1,xt[1:n]*cos(rotacao[i])-yt[1:n]*sin(rotacao[i]))
yt1=c(yt1,xt[1:n]*sin(rotacao[i])+yt[1:n]*cos(rotacao[i]))
}
dt= data.frame(xt1,yt1,z="circulo")
```

Script – Parte 2 Visualizando a mandala

```
p= ggplot()+
coord_fixed()+
    theme_void()
p=
  p+
   geom_point(data=dt, aes(x=xt1, y=yt1), color='black')
p
```

Mandala 2: Elaborada a partir do cardióide com rotação
Inicialmente criamos em x e y os pontos que formam um cardióide de raio 1, posteriormente é realizado a rotação dos pontos do cardióide nos ângulos de $\frac{\pi}{4}, \frac{\pi}{2}, \ldots, \frac{7\pi}{4}$, tais pontos são armazenados em um *dataframe* denominado dt. O número de linhas em dt é formado por 4000, ou seja, 4000 pontos a serem plotados.
Script – Parte 1 gerando pontos e armazenando em dt

```
n=500; t=seq(0, 2*pi, length.out = n); rotacao=pi/4*(1:7)
x=c(2*raio*cos(t)-raio*cos(2*t))
y=c(2*raio*sin(t)-raio*sin(2*t))
xt=x; yt=y #rotação dos pontos
for(i in 1:length(rotacao)){
xt=c(xt, x[1:n]*cos(rotacao[i])-y[1:n]*sin(rotacao[i]))
yt=c(yt, x[1:n]*sin(rotacao[i])+y[1:n]*cos(rotacao[i]))
}
dt= data.frame(xt, yt, z="cardióide")
```

Script – Parte 2 Visualizando a mandala (o mesmo já descrito)

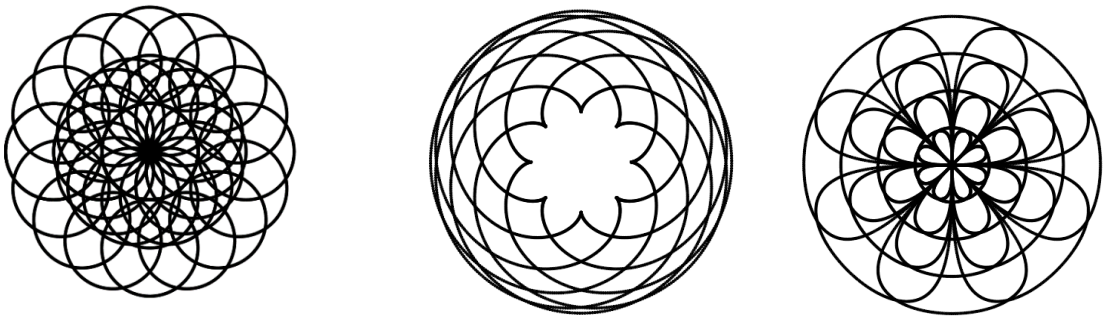


Figura 6: Mandala feita de círculo transladado e rotacionado (esquerda) Mandala feita de cardióide rotacionado (centro) e Mandala feita da Lemniscata com rotação e contração (direita). Fonte: autora, 2022.

Mandala 3: Elaborada a partir da lemniscata com rotação e contração

Inicialmente criamos em x e y os pontos que formam uma lemniscata, posteriormente é realizado a rotação dos pontos da lemniscata nos ângulos de $\frac{\pi}{4}, \frac{\pi}{2}$ e $\frac{3\pi}{4}$, tais pontos são armazenados em xt, yt e finalmente aplica-se uma redução de fator 0.25, 0.5 e 0.75 nos pontos gerados anteriormente, armazenando todos eles em um *dataframe* denominado dt. Em dt há 8000 pontos a serem plotados.

Script – Parte 1 gerando pontos e armazenando em dt

```
n=500; t=seq(0, 2*pi, length.out = n); rotacao=pi/4*(1:3)
x=sin(t);  y=sin(t)*cos(t)
xt=x; yt=y#rotações
for(i in 1:length(rotacao)){
xt=c(xt, x[1:n]*cos(rotacao[i])-y[1:n]*sin(rotacao[i]))
yt=c(yt, x[1:n]*sin(rotacao[i])+y[1:n]*cos(rotacao[i])) }
xtt=NULL; ytt=NULL; red=c(0.25, 0.5, 0.75) #redução
for(i in 1:length(red)){
provx=paste0("x",i); provy=paste0("y",i)
xtt=c(xtt, assign(provx, xt*red[i]))
ytt=c(ytt, assign(provy, yt*red[i])) }
dt=data.frame(x=c(xt, xtt), y=c(yt, ytt), z="lemniscata")
```

Script – Parte 2 Visualizando a mandala (o mesmo já descrito)

Outras combinações a partir de figuras geométricas podem ser realizadas, figura 7:

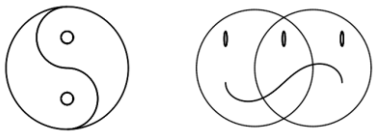


Figura 7: Mandalas feitas com a combinação de figuras geométricas como círculo, elipse e lemiscata. Fonte: autora, 2022.

5. Aplicativo Shiny para mandalas

Esta experiência de programar as mandalas levou a produção de um aplicativo shiny disponível em https://lucianealcoforado.shinyapps.io/Mandala/, Alcoforado (2022).

O aplicativo possui uma base de pontos para gerar diversas opções de mandalas e o usuário pode optar por colorir os pontos e o plano de fundo.

Referencias

[1] Alcoforado, L.F. (2022), *Mandala*. Aplicativo Disponível em https://lucianealcoforado.shinyapps.io/Mandala/. Acesso em 11/09/2022.

[2] Alcoforado, L.F. (2021) *Utilizando a Linguagem R: conceitos, manipulação, visualização, Modelagem e Elaboração de Relatório*, Alta Books, Rio de Janeiro.

[3] Bezerra, J. (2022), *Mandalas*, Toda Materia. Disponível em https://www.todamateria.com.br/mandala/. Acesso em 11/09/2022.

[4] Rezende, E.Q.F, Queiroz, M.L.B. (2008). *Geometria Euclidiana Plana de Construções Geométricas*, 2a. ed., Unicamp, Campinas.

[5] Stover, C., Weisstein, E.W. (2022). *Parametric Equations*. From MathWorld - A Wolfram Web Resource. https://mathworld.wolfram.com/ParametricEquations.html, Acesso em 11/09/2022.

[6] Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

Geoportal: Índice Multidimensional de Pobreza

Miguel Flores[1], Andrés Vinueza[2] y Jorge Sosa[3]

[1]Grupo MODES, SIGTI, Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador.
[2]Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador.
[3]Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador.

RESUMEN

Las tecnologías de información y las comunicaciones son el medio para construir ciudades inteligentes (Smart Cities) que permiten administrar recursos de forma eficiente, sostenible y amigable con el ambiente. Los pilares o módulos de un sistema (geoportal) propuesto por la empresa LOGIKA para administrar una Smart City son: información social, gestión territorial, pulso de la ciudad, gestión de tráfico y smart buisness.

El cantón Manta ubicado en el Ecuador, ha centrado sus esfuerzos en priorizar las zonas territoriales donde se necesita una mayor intervención a través de políticas públicas para mejorar el bienestar de los ciudadanos más vulnerables. Para lograrlo ha iniciado con el desarrollo del primer pilar que se refiere a información social.

Este módulo, se ha desarrollado utilizando R para el cálculo de indicadores de accesibilidad y desigualdad social y el paquete shiny [1] para la interfaz gráfica (geoportal). Esta herramienta informática permitirá a las autoridades del cantón Manta: analizar las condiciones de vida de los ciudadanos, e identificar los sectores priorizados que requieren gestionar política pública emergente.

Palabras e frases chave: Análisis de componentes principales no lineales del kernel, Desigualdad social, Condiciones de vida, Heterogeneidad espacial,smart city

1. INTRODUCIÓN

En el post "Big data al servicio de las ciudades" publicado en octubre en la web del Banco Interamericano de Desarrollo (BID), con la colaboración de los autores, se da una visión y reseña del uso de los datos generados por los ciudadanos. Se comenta que "los ciudadanos generamos una gran cantidad de datos e información en nuestro día a día. Esta información, cuando es procesada y analizada, es lo que se conoce como "big data" o datos a gran escala. El buen uso de estos datos a gran escala por parte de las ciudades y municipios puede servir para mejorar la prestación de servicios públicos para medir el impacto en el bienestar de sus habitantes"[2].

Desde la Alcaldía de Manta con el equipo del proyecto se trabajó en el desarrollo e implementación de un geoportal de información social que permita cumplir con el objetivo de conocer la realidad social de los ciudadanos, garantizar su calidad de vida, y guiar adecuadamente las decisiones en política pública en favor de los más necesitados.

Los responsables municipales junto con el Banco Interamericano de Desarrollo (BID) y los autores combinaron esfuerzos para diseñar un sistema web de información social. El que permite la determinación de los niveles de desigualdad, la identificación de las zonas con mayor desigualdad social mediante la definición y construcción de indicadores de accesibilidad y desigualdad social. En el siguiente link, se puede acceder al geoportal y los diferentes reportes: `https://avgeoportal.shinyapps.io/Geoportal/`

La definición y el cálculo de los indicadores, se ha podido realizar considerando dos fuentes de información: la primera, a través de información de OpenStreetMap extraída con la librería en R osmdata [3]; y la segunda, es información facilitada por el municipio y las empresas públicas del cantón. Esta segunda fuente consiste en información geográfica e información relacionada con servicios públicos. Por ejemplo, se contó con información de las empresas públicas de electricidad y agua potable relacionada a los consumos y facturación de los ciudadanos del cantón.

2. METODOLOGÍA: CÁLCULO DE INDICADORES

Para el cálculo de los indicadores de accesibilidad, se ha considerado la metodología propuesta en [4]. El primer indicador mide la distancia desde un sector al servicio público más cercano. Este indicador tiene la siguiente formulación:

$$A_i = \frac{\text{mín}_j\, d_{ij}}{\text{máx}_i\{\text{mín}_j\, d_{ij}\}},\ \forall i \in \{1,\ldots,n\}$$

Donde, d_{ij} son las distancias de un sector $i \in \{1,\ldots,n\}$ a un servicio $j \in \{1,\ldots,m\}$. Estas distancias fueron calculadas con el paquete en R geosphere [5].
El segundo es el indicador de separación espacial, se estima el promedio de recorridos de todas las zonas de origen a todos los puntos de destino o servicios.

$$B_i = \frac{\sum_j d_{ij}}{n},\ \forall i \in \{1,\ldots,n\}$$

Los indicadores tienen como unidad de análisis los sectores censales y estos a su vez se consideran como dimensiones de estudio para identificar zonas con tasas altas de desigualdad social que permiten construir el Índice Multidimensional de Pobreza aplicando el análisis de componentes principales no lineal basado en un kernel[6]. A continuación, se resume el algoritmo aplicando para el cálculo del índice:

1. Dado un conjunto de datos $x_1, x_2, \ldots, x_n$, con $x_i \in \mathbb{R}^{t\times p}$, $i = 1,\ldots,n$
2. Calcular la matriz kernel $K = (K_{ij}), K_{ij} = k(x_i, x_j)$
3. Calcular la matriz del kernel centrado $K = HKN, H = I_n - \frac{1}{n}1_n 1_n^T$
4. Calcular los vectores propios α_i y los valores propios λ_i de K
5. Escoger el i-ésimo vector α_i sobre el cual se va a proyectar
6. Normalizar α_i para que tenga longitud $\frac{1}{\sqrt{n\lambda_i}}$
7. Para la muestra $x_i, i = 1,\ldots,n$ el valor de la proyección de $\Phi_i \in H$ es $\langle \Psi(x_i), u_j\rangle = \sum_{k=1}^n \alpha_k^j \langle \Psi(x_i), \Psi(x_k)\rangle = \sum_{k=1}^n \tilde{K}\alpha_k^j$

El kernel tomado es la función kernel de Gauss que es una función de decaimiento de distancia.

$$k(x, x') = exp(-\gamma||x - x'||_F^2),\ \gamma > 0 \tag{1}$$

Donde $||\cdot||_F$ es la norma de Frobenius, γ es una constante seleccionada adecuadamente a los datos. El conjunto $\{x_1, x_2, \ldots, x_n\}$ representa los datos de n sectores y p variables que se reemplaza por la matriz $w_i X_i$ donde w_i con $i \in \{1,\ldots,n\}$ es el peso geográfico asociado con el $i-$ésimo sector de observación.
El índice desarrollado y aplicado considera: por un lado la ponderación a través del valor del suelo (impuesto predial) y así caracterizar de mejor forma la población; y por otro lado, este índice considera información que considera el consumo en servicios básicos y complementarios, acceso a servicios de salud, educación y seguridad.
Con respecto al kernel utilizado, fue construido con los valores de los predios de cada sector censal y corregido con la tasa entre viviendas y hogares. Al aplicar el ACP se usó todas las variables mencionadas en la parte inicial y se filtró las variables con mayor representatividad.

2. GEOPORTAL: INFORMACIÓN SOCIAL

El producto final del proyecto, es un Sistema de Información Social (SIC) implementado a través de un geoportal (ver Figura 1) que ha potenciado los procesos y conocimientos para priorizar zonas de intervención en la ciudad de Manta. El SIC permite la actualización e incorporación de nueva información cartográfica útil en la definición de los indicadores de desigualdad y en la gestión pública. Así también, permite calcular los indicadores y el índice multidimensional de desigualdad cuantificando y combinando componentes como el valor del suelo y la accesibilidad a salud, educación y seguridad, entre otras variables antes mencionadas.
El SIC, además de presentar y calcular los indicadores e índice multidimensional integra la sección "Escucha ciudadana". Este espacio permitirá acercar a la comunidad con las autoridades municipales, reforzando los procesos de participación ciudadana y rendición de cuentas en la toma de decisiones mediante los incidentes reportados por la aplicación municipal "Manta App". Adicionalmente, el módulo se complementa al explotar la información digital con el monitoreo de noticias y redes sociales

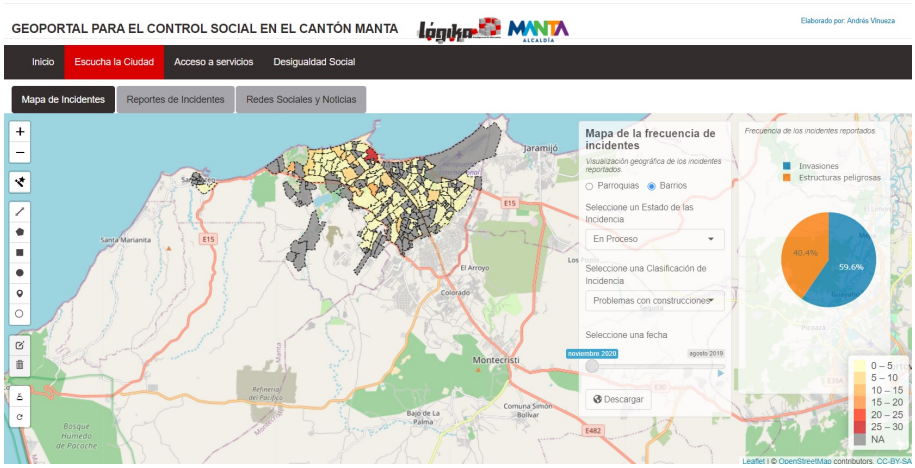


Figura 1: Geoportal: Información Social.

Gracias a los reportes generados y mediante la herramienta desarrollada han permitido complementar la información con la que contaba el municipio para fortalecer la construcción de sus indicadores socio-económicos, territoriales y de gestión integral, entre otros.

3. CONCLUSIONES

La metodología utilizada para la construcción del indicador de desigualdad social toma en cuenta el lugar geográfico donde se encuentra la zona de estudio mediante la construcción de una matriz de pesos espaciales considerando la función de decaimiento de distancia de Gauss y la norma de Frobenius, con la finalidad de modelar el efecto de la distancia en las interacciones espaciales. Además, al considerar la plusvalía y la proporción del número de viviendas y el número de hogares como peso geográfico, se pudo describir la influencia geográfica de las regiones de acuerdo con su ubicación.
Este indicador permite medir la condición de desigualdad de la población de un sector y dónde está ubicada, no es posible obtener la ubicación exacta de los grupos u hogares vulnerables dentro de esa localidad. Esto, no se debe a la metodología, sino a los datos obtenidos. Por tanto, es importante para comprender los factores de desigualdad que influyen en el territorio, considerar niveles más bajos de análisis regional, como la ubicación de la vivienda.

AGRADECIMIENTOS

El proyecto a podido darse a cabo gracias al gobierno autónomo descentralizado de Manta que ha permitido acceder y construir las bases de datos necesarias para el análisis. También, agradecemos el apoyo del BID por el asesoramiento brindado. Finalmente, agradecemos a la empresa LOGIKA inteligencia de mercados y sus colaborados Nicolas Mafla, Abigail Valencia, Jasmina Vizuete y Ana Cabezas por el aporte brindado.

Referencias

[1] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2021. R package.

[2] Flores M., Vinueza A., Cabezas B., Vasquez H., and Rosero D. Big data al servicio de las ciudades, October 2021. `https://blogs.iadb.org/ciudades-sostenibles/es/big-data-al-servicio-de-las-ciudades/`, Last accessed on 2022-09-24.

[3] Mark Padgham, Bob Rudis, Robin Lovelace, and Maëlle Salmon. osmdata. *The Journal of Open Source Software*, 2(14), jun 2017.

[4] Carlos F. Garrocho Rangel and Juan Campos Alanís. Un indicador de accesibilidad a unidades de servicios clave para ciudades mexicanas: fundamentos, diseño y aplicación. *Economía Sociedad y Territorio*, September 2006.

[5] Robert J. Hijmans. *geosphere: Spherical Trigonometry*, 2021. R package version 1.5-14.

[6] Mirosław Krzyśko, Wojciech Łukaszonek, Waldemar Ratajczak, and Waldemar Wołyński. Analiza nieliniowych składowych głównych dla danych czasowo-przestrzennych geograficznie ważonych. *Acta Universitatis Lodziensis. Folia Oeconomica*, 4(337):169–181, September 2018.

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Desarrollo de una metodología para la evaluación del riesgo de invasión basada en modelos de distribución de especies: el caso de Vespa velutina en Europa

Victoria Formoso-Freire[1], A. Márcia Barbosa[2], Andrés Baselga[3], Carola Gómez-Rodríguez[4]

[1] CRETUS, Dept of Functional Biology and Dept of Zoology, Genetics and Physical (Area of Ecology), Univ. de Santiago de Compostela, Santiago de Compostela, Spain.

[2] CICGE (Centro de Investigação em Ciências Geo- Espaciais), Universidade do Porto, Porto, Portugal

[3] CRETUS, Dept of Zoology, Genetics and Physical Anthropology, Univ. de Santiago de Compostela, Santiago de Compostela, Spain

[4] CRETUS, Dept of Functional Biology (Area of Ecology), Univ. de Santiago de Compostela, Santiago de Compostela, Spain.

RESUMO

Los modelos predictivos son herramientas claves en la gestión del medioambiente, ya que nos permiten comprender la relación entre la biodiversidad y su medio y, con ello, anticipar las respuestas bióticas ante cambios en el equilibrio ecológico. En concreto, los modelos de distribución de especies son una herramienta clave en la lucha preventiva contra la invasión de especies, una de las mayores amenazas para la biodiversidad en la actualidad (Kortz & Magurran, 2021). Actualmente, una de las especies invasoras más preocupantes en Europa es la avispa asiática (Vespa velutina), ya que supone una amenaza para la estabilidad de los ecosistemas debido a su fuerte impacto sobre las poblaciones de polinizadores (Monceau et al., 2014). En este sentido, es esencial generar una estrategia de gestión eficaz que incluya los siguientes aspectos: (i) la detección temprana de la especie en el área invadida, y (ii) la identificación de localidades climáticamente adecuadas para el establecimiento de la especie. Con este objetivo, hemos desarrollado algoritmos que nos permiten identificar las regiones de Europa con mayor probabilidad de establecimiento de esta especie invasora. Para ello hemos integrado algunos de los modelos estadísticos más usados en Biogeografía (presence-only, presence-background y presence-absence). En general, los modelos muestran que casi todo el continente europeo presenta condiciones favorables para el desarrollo de V. velutina, si bien la mayor probabilidad de expansión es hacia las Islas Británicas, Península Itálica, oeste de la Península Balcánica y sur de la Península Báltica. Esta cartografía de riesgo de invasión supondrá una herramienta clave para la priorización de los recursos de seguimiento y la orientación de los esfuerzos de erradicación hacia las zonas de alto riesgo. Además, la metodología que hemos puesto a punto podrá ser aplicada en otros contextos, como la respuesta de las especies ante la degradación de las condiciones abióticas o alteraciones en la composición de las comunidades biológicas.

Palabras e frases chave: Invasión, Conservación biodiversidad, Modelos predictivos de distribución de especies (SDM)

Referencias

Frans, V. F., Augé, A. A., Fyfe, J., Zhang, Y., McNally, N., Edelhoff, H., Balkenhol, N., & Engler, J. O. (2022). Integrated SDM database: Enhancing the relevance and utility of species distribution models in conservation management. Methods in Ecology and Evolution, 13(1), 243–261. https://doi.org/10.1111/2041-210X.13736

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). Habitat suitability and distribution models with applications in R. Cambridge University Press.

Kortz, A. R., & Magurran, A. E. (2021). Complex community responses underpin biodiversity change following invasion. Biological Invasions, 23(10), 3063–3076. https://doi.org/10.1007/s10530-021-02559-8

Monceau, K., Bonnard, O., & Thiéry, D. (2014). Vespa velutina: A new invasive predator of honeybees in Europe. Journal of Pest Science, 87(1), 1–16. https://doi.org/10.1007/s10340-013-0537-3

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

REFREG: UN PAQUETE DE R PARA ESTIMAR REXIÓNS DE REFERENCIA

Óscar Lado-Baleato [1], Javier Roca-Pardiñas [2,3], Carmen Cadarso-Suárez[4,3,†] & Francisco Gude[5]

[1] Plataforma de apoio os ensaios clínicos, Instituto de Salud Carlos III, Instituto de Investigación Sanitaria (IDIS), Santiago de Compostela, Galicia.
[2] Inferencia estadística, decisión e investigación operativa (SiDOR), Universidade de Vigo, Galicia, España.
[3] Centro de investigación e tecnoloxía matemática de Galicia (CITMAga).
[4] Departamento de estatística, análise matemático e optimazación, Universidade de Santiago de Compostela, Galicia, España.
[5] Unidade de epidemioloxía clínica, Complexo Hospitalario de Santiago de Compostela, Galicia, España.

RESUMO

O diagnóstico e control da maioría das enfermidades basease en marcadores de distribución continúa. Os resultados dos mesmos intrepetanse mediante os chamados intervalos de referencia (IR). Este intervalo define os valores de normalidade dentro da poboación san como dous puntos de corte entre os que se atopan o 95 % dos resultados. A extensión do IR para o caso de multiples marcadores denominase rexión de referencia multivariada (RRM). Esta definense por unha envoltura convexa que contén a maioría dos resultados multivariados dos pacientes sans. Propostas fai agora 50 anos a súa aplicación na práctica clínica é anecdótica, a pesar das súas vantaxas; ofrecen unha maior especificidade e sensibilidade no diagnóstico de enfermidades que empregen máis dun marcador continúo (p.ex, diabetes e hipo- ou hipertiroidismo).

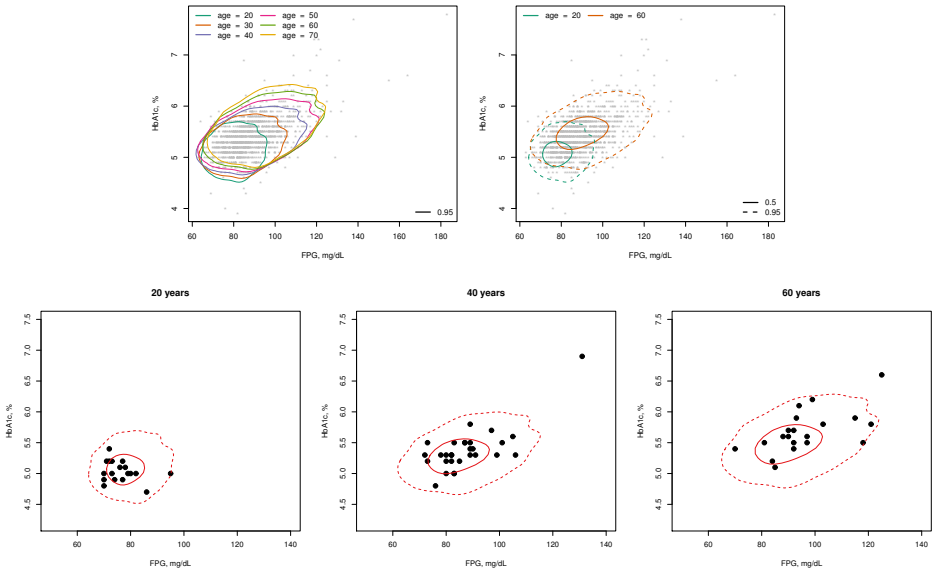


Figura 1: Resumo gráfico da estimación dunha rexiónd de referencia condicional para dous marcadores continúos mediante o paquete `refreg`.

O paquete `refreg` implementa novos métodos de estimación de rexións de referencia. Esta rexión estimase de forma non-paramétrica empregando estimadores de densidade bivariada tipo núcleo, o efecto das covariables sobre esta rexión estimase mediante un modelo de localización-escala bivariado[1, 2]. O paquete permite a investigadores clínicos propoñer regras diagnósticas específicas para cada paciente baseadas na distribución conxunto de marcadores continúos de distribución non-Gausiana (ver Figura 1). Actualmente estamos a traballar tanto na extensión dos métodos implementados no paquete para dimensións > 2 como no desenvolvemento de métodos de selección de variables no contexto das rexións de referencia.

Palabras e frases chave: regresión bivariada, rexións de referencia, estimación non-paramétrica, diabetes, diagnóstico clínico.

AGRADECEMENTOS

Óscar Lado-Baleato está actualmente contratado pola plataforma de apoio os ensaios clínicos do instituto de Saúde Carlos III (ISCIII) PT20/00043. Este estudio foi financiado por ISCIII PI20/01069, ISCIII RD21/0016/0022, cofinanciado pola UE. Dentro do proxecto MTM2017-83513-R. E tamén con financiación da Xunta de Galicia: ED431C 2020/20 e IN607A/2021-2.

Referencias

[1] Lado-Baleato Ó., Roca-Pardiñas J., Cadarso-Suárez C., Gude F. (2021). Modeling conditional reference regions: Application to glycemic markers. *Statistics in Medicine* 40(26), 5926–5946.

[2] Roca-Pardiñas J., Ordoñez C., Lado-Baleato Ó. (2021). Nonparametric location–scale model for the joint forecasting of SO2 and NOx pollution episodes. *Stochastic Environmental Research and Risk Assessment* 35(2), 231-244.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

Do one thing every day that scares you.

Ariel Levy[1], Orlando Celso Longo[2] e Luciane F. Alcoforado3[3]

[1] Universidade Federal Fluminense 1
[2] Universidade Federal Fluminense
[3] Academia da Foça Aérea Brasileira

Abstract

SER is a multidisciplinary event, which integrates professionals, students, and practitioners from most diversified knowledge areas of data analysis.

Key Words: Evento SER, R language

1. Do one thing every day that scares you.

This advice from Eleonor Roosevelt changed our lives in 2015 when we decided to study and join people who also learn and practice the R language. To achieve this goal; we started SER - The International Seminar of Statistics with R, for which we have the crucial support of prof. Manoel Febrero Bande, USC, Prof. Maysa Sacramento de Magalhães, ENCE/IBGE, Jorge P Zubelli, IMPA, Celso José da Costa, IME- UFF, Sidney Mello – UFF.

SER was considered an event to give newbies a chance to present in a mix with more experienced users. As a result, students and practitioners can catch up and network.

The event was recognized by the R Foundation (2018) for its pioneering in Latin America in bringing together an expressive number of R users.

Every year, the challenge scares us even more. Anyone who tries to start or maintain themed events will understand the difficulties of putting up the show. And even more in Brazil, with no sponsors except for CAPES (Coordenação de Aprefeiçoamento de Pessoal de Nível Superior) but minimal funding.

From the first edition (2016) with 215 participants from 4 countries to the fiftth edition (2021) with more than 1400, from 27 countries, and VI SER (2022) with more than 1300, it was an apprenticeship journey. SER is the Brazilian most expected R event, where almost every new Brazilian R package is presented.

All this effort paid off as we relied on the R community, which showed aremarkable growth during these years in Brazil. As a result, users have spread all scientific subjects to the public and business administration.

Although we did come to more than expected, this success could be even greater if the big players in this industry had supported us by funding or showing their advances to our community.

Even this lack of support did not stop us from figuring out the spikes and trends. Instead, in SER's first edition, we called for market innovation and professional performance, predicting R as a differential in employability.

In May 2017, the second edition took place, which consolidated the inclusion in the annual calendar for researchers who need to update and disseminate their research in progress. By then, the call was for Big Data analysis. The following year pulled out multidisciplinary capacities of R with the possible integration of fields of science. In 2019, we brought up IV SER – R & Python, and Covid-19 with everybody online; the call was Remotes(). In 2022, the sixth edition was set one year before claimed - R for All; a year ago, we had no idea that Quarto was coming up, but it was somehow already to be expected.

R software (R Core Team, 2022) has been standing out on the world stage, a free software adopted by universities, and public and private sectors, providing resource savings insofar as it avoids the expense of expensive commercial software licenses.

A new breath was brought up with the *tidyverse* package (Wickham et al., 2019) and the pipe from *magrittr* package (Bache; Wickham, 2022). The *ggplot2* package (Wickham, 2016) inspires a lot of other extensions to enhance its outputs. The *Shiny* package (Chang et al., 2022), as the name says, brought the glow from the internet into R possibilities and enhanced the flow of managing and reporting using simple steps to publicize one's work. Machine Learning also presented a new simple approach with the *tidymodels* package (Kuhn et al., 2020).

The R Markdown (2014) package (Xie ; Allaire; Grolemund, 2019) was a breakthrough in bringing together analysis and reporting. Moreover, this year, the new Posit (RStudio) presented the *Quarto* (Allaire, 2022), an open-source scientific and technical publishing system built on Pandoc that made it possible to create dynamic content with Python, R, Julia, and Observable.

On our side, as academics, despite having The R Journal, we lack journals for such a rich field of applied work and are now presenting The SER Journal and hope to contribute to more spaces for researchers and practitioners to present their work.

During the IX Jornada with R, we expect to share experiences (Levy,2020) and discuss best practices in teaching, learning, and doing research in R either in a presentation or at a round table.

References:

Allaire JJ (2022). quarto: R Interface to 'Quarto' Markdown Publishing System. R package version 1.2. https://CRAN.R-project.org/package=quarto

Bache S, Wickham H (2022). magrittr: A Forward-Pipe Operator for R. R package version 2.0.3,https://CRAN.R-project.org/package=magrittr

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022, July 19). shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny

Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Levy, A. (Diretor). (2020, October 21). Como eu aprendo e ensino R em 2020. https://www.youtube.com/watch?v=70TwbqBfInQ

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SER - International Seminar on Statistics with R | SER - International Seminar on Statistics with R. ([s.d.]). Recuperado 25 de setembro de 2022, de https://ser.uff.br/

Xie, Y., Allaire, J. J., & Grolemund, G. (2019). R Markdown: The definitive guide. CRC Press, Taylor and Francis Group.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer Science+Business Media, LLC.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.”_Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 (URL: https://doi.org/10.21105/joss.01686).

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

Utilização do pacote AHP na tomada de decisão

Orlando Celso Longo[1]
Luciane Ferreira Alcoforado[2]
Ariel Levy[3]
[1] Programa de Pós-Graduação em Engenharia Civil/UFF
[2] Academia da Força Aérea/Divisão de Ensino
[3] Programa de Pós-Graduação em Administração/UFF

RESUMO

Neste trabalho apresentamos o pacote AHP implementado em R, trata-se de um pacote para aplicação do método de Análise Hierarquica de Processo (AHP) de autoria de Tomas Saaty para o processo de tomada de decisão. Como ilustração mostraremos um estudo de caso sobre a escolha de uma obra de ligação entre dois pontos geográficos, com base em duas alternativas apuradas em estudos preliminares de viabilidade técnica, em condições normais de utilização.

Palabras e frases chave: AHP, linguagem R, tomada de decisão.

1. Introdução

O método da Análise Hierarquica de Processos, AHP (Analytic Hierachy Process) foi desenvolvido na década de 1970 por Tomaz Saaty. É uma ferramenta que visa apoiar a tomada de decisão quando há múltiplos critérios e alternativas à disposição. O AHP é o método com o maior número de artigos publicados em periódicos científicos. Saaty apresentou alguns dos usos do AHP como em decisões militares, de alta gerência em diversas empresas, estudos sobre economia e mesmo em conflitos entre nações. Apesar de sua popularidade o AHP apresenta dificuldades na construção de julgamentos. Estes julgamentos são feitos par a par comparando um item a seus pares em uma matriz de julgamentos ou também conhecida como matriz paritária.
Em 2019, o pacote AHP foi implementado por Oliveira (2020) na linguagem R durante o desenvolvimento de um projeto de iniciação científica da Universidade Federal Fluminense. Neste trabalho vamos mostrar o uso do pacote através de um estudo de caso de um problema de logística resolvido pela engenharia civil.
O estudo de caso em tela surge da necessidade de uma meio de ligação entre dois pontos geográficos para resolver o problema de minimizar o tempo de deslocamento. A região é caracterizada por montanhas e vales e duas alternativas se mostraram viáveis a partir de um estudo de viabilidade técnica, para uma mesma extenção: 1-ponte; 2-túnel. O processo de tomada de decisão foi pautado em quatro critérios macros, definidos na tabela 1: 1-ciclo de vida, para um periodo minimo de 100 anos de utilização, 2-custo de manutenções periódicas, preventivas, preditivas e corretivas, 3-impactos ambientais no entorno do empreendimento , 4-custo de construção para transporte modal rodoviário.

Tabela 1: Definição dos critérios

Critérios	Definições
1 Ciclo de vida	Tempo estimado de utilização da ponte/túnel
2 Custo de manutenção	Valor monetário para manter o funcionamento da ponte/túnel
3 Impactos ambientais	Medidas para minimizar a destruição ambiental no entorno da ponte/túnel
4 Custo de construção	Custo total da construção somados todas as despesas da ponte/túnel

2. Conceitos Iniciais

Alternativas: representam as diferentes opções de ação disponíveis para o tomador de decisão. Normalmente, o conjunto de alternativas é considerado finito, podendo chegar a centenas. Elas devem ser selecionadas, priorizadas e eventualmente classificadas.
Múltiplos Critérios: Cada problema está associado a vários critérios. Os critérios representam as diferentes dimensões a partir das quais as alternativas podem ser visualizadas.
Matriz de decisão: Uma matriz de decisão A é uma matriz (M × N) na qual o elemento aij indica o desempenho da alternativa Ai quando avaliada em termos do critério de decisão Cj, (para i = 1,2,3,..., M e j = 1,2,3,..., N). Supõe-se também que o tomador de decisão determinou os pesos de desempenho dos critérios de decisão (denominados como Wj, para j = 1,2,3,..., N). Esta informação é melhor resumida na figura1.

	Criteria				
	C_1	C_2	C_3	...	C_N
Alt.	W_1	W_2	W_3	...	W_N
A_1	a_{11}	a_{12}	a_{13}	...	a_{1N}
A_2	a_{21}	a_{22}	a_{23}	...	a_{2N}
A_3	a_{31}	a_{32}	a_{33}	...	a_{3N}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_M	a_{M1}	a_{M2}	a_{M3}	...	a_{MN}

Figura 1: Matriz de Decisão, Triantaphyllou et all (1998).

Escala Fundamental de Saaty: estabelece valores para a matriz de julgamentos, conforme figura 2:

Relação da escala utilizada com a Escala de Saaty

Escala Percentual (%)		Escala Fundamental de Saaty	Grau de Importância
X	Y		Definição
90%	10%	9	Componente X é extremante mais importante que o Componente Y.
80%	20%	7	Componente X é muito importante em relação ao Componente Y.
70%	30%	5	Componente X é importante em relação ao Componente Y.
60%	40%	3	Componente X é um pouco mais importante em relação ao Componente Y.
50%	50%	1	Os dois Componentes têm a mesma importância.
40%	60%	1/3	Componente Y é um pouco mais importante em relação ao Componente X.
30%	70%	1/5	Componente Y é importante em relação ao Componente X.
20%	80%	1/7	Componente Y é muito importante em relação ao Componente X.
10%	90%	1/9	Componente Y é extremante mais importante que o Componente X.

Fonte: Bandeira & Correia (2006).

Figura 2: Escala Fundamental de Saaty para comparação entre pares.

Julgamento Holístico: Fornece pesos a cada critério usando a escala de Saaty: supondo que haja n criterios, estabelecer pesos diferentes para cada um dos criterios de acordó com sua importância, sendo w1 o peso do critério 1; w2 o peso do critério 2 e assim por diante. A matriz paritária será construída fazendo aij = wi – wj +1 se wi > wj (ou seja, critério i possui importância maior que critério j); aij = 1/(wj – wi +1) se wi < wj. Este método foi proposto por Godoi (2014).

Matriz de Julgamentos: É uma matriz que contém o resultado do julgamento pelo comitê decisor após realizar a comparação entre pares (tanto dos critérios como também das alternativas à luz de cada criterio), utilizando a escala fundamental de Saaty. O número de matrizes de julgamento é igual ao número de critérios mais um, ou seja se há 4 critérios, haverão 5 matrizes de julgamento (uma comparando os critérios entre si e as outras quatro matrizes serão associadas a cada critério, comparando as alternativas entre si).

Estrutura hierarquica da AHP: establece a estrutura hierarquica do problema, conforme figura 3.

Figura 3: Estrutura Hieráquica do problema,

A etapa final da AHP trata da estrutura de uma matriz M×N (onde M é o número de alternativas e N é o número de critérios). Esta matriz é construída usando as importâncias relativas das alternativas em termos de cada critério. O vetor (ai1, ai2, ai3, ..., aiN) para cada i é o principal autovetor de uma matriz de julgamento N×N que é determinado por comparações aos pares do impacto das M alternativas no i-ésimo critério.

3. O Problema

O problema consiste em determinar a melhor escolha entre duas alternativas A1 = construção de uma ponte ligando dois pontos, com a técnica construtiva de pre-moldados; A2 = construção de um túnel ligando dois pontos, com caracteristica de material geológico em solo argiloso, com base nos seguintes critérios: C1-ciclo de vida, C2-custo de manutenção, C3-impactos ambientais, C4-custo de construção.

Julgamento Holístico

M1 - Matriz de julgamento dos critérios

Pesos atribuídos pelos avaliadores para cada criterio: w1= 1; w2 = 5; w3 = 3; w4 = 4

aij = wi – wj +1 se wi > wj

	c1	c2	c3	c4
c1	1	1/5	1/3	1/4
c2	5	1	3	2
c3	3	1/3	1	1/2
c4	4	1/2	2	1

Julgamento Holístico:
i=1,2,3,4 e j=1,2,3,4
aij = wi – wj +1 se wi > wj
aij = 1/(wj – wi +1) se wi < wj

M2 - Matriz de julgamento das alternativas em relação ao critério C1 – ciclo de vida

Pesos atribuídos pelos avaliadores para cada Alternativa: w1= 1, w2 = 3

C1=ciclo de vida	A1	A2
A1	1	0,33
A2	3	1

M3 - Matriz de julgamento das alternativas em relação ao critério C2- custo de manutenção

Pesos atribuídos pelos avaliadores para cada Alternativa: w1= 1, w2 = 4

C2=manutenção	A1	A2
A1	1	0,25
A2	4	1

M4 - Matriz de julgamento das alternativas em relação ao critério C3- impactos ambientais

Pesos atribuídos pelos avaliadores para cada Alternativa: w1= 1, w2 = 2

C3=ambiental	A1	A2
A1	1	0,5
A2	2	1

M5 - Matriz de julgamento das alternativas em relação ao critério C4- custo de construção

Pesos atribuídos pelos avaliadores para cada Alternativa: w1= 5, w2 = 3

C3=construção	A1	A2
A1	1	3
A2	0,333	1

4. Utilizando o pacote AHP

Para utilizar o pacote AHP é necessário organizar as matrizes de julgamento em um arquivo contendo n+1 planilhas, sendo n o número de critérios do problema. No caso em tela devemos organizar uma planilha para cada matriz M1, M2, ..., M5, nesta exata ordem e em seguida utilizar a função ***ahp_geral*** do pacote.

A organização da planilha pode ser vista na figura 4. O script e resultado final pode ser visto na figura 5.

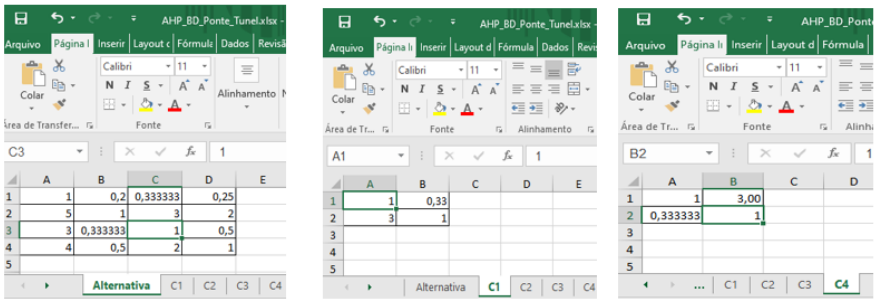


Figura 4:Matrizes organizadas em planilhas: M1em Alternativa, M2 em C1, M3 em C2, M4 em C3, M5 em C4

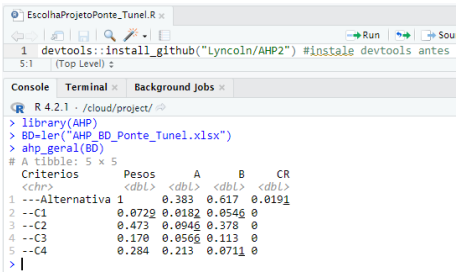


Figura 5:Solução do problema com uso do pacote AHP.

Discussão dos Resultados

A matriz de decisão final informa o peso de cada critério em relação a cada alternativa. Assim os critério 1,2,3 e 4 tiveram peso final respectivamente igual a 0.0729; 0.473; 0.17 e 0.284, portanto o critério de maior peso na decisão final foi o critério 2 (custo de manutenção). É possível ainda visualizar o peso de cada alternativa em relação a cada critério, por exemplo quanto ao critério C2, as alternativas A1 e A2 tiveram peso 0.0946 e 0.378 respectivamente. O peso final de cada alternativa foi 0.383 para a alternativa A1 e 0.617 para a alternativa A2. Com base neste resultado a alternativa preferível é a alternativa A2 ou seja construir um túnel. Note que a soma dos pesos dos critérios (coluna 2) ou das alternativas (linha 2) somam sempre 1. A última coluna informe o índice de consistência dos julgamentos (CR), sendo considerável aceitável se for menor do que 0,1 como foi o caso deste exemplo (CR = 0.0191)

Referencias

[1] Alcoforado, L.F. (2021) *Utilizando a Linguagem R: conceitos, manipulação, visualização, Modelagem e Elaboração de Relatório*, Alta Books, Rio de Janeiro.
Unicamp, Campinas.

[2] Godoi, W.C. (2014). Método de construção das matrizes de julgamento paritário no AHP – método de julgamento holístico. Revista Gestão Industrial, ISSN 1808-0448 / v. 10, n. 03: p.474- 493, D.O.I: 10.3895/gi.v10i3.1970

[3] Oliveira, L.S., *AHP*, Github.com. (2020) URL = https://github.com/Lyncoln/AHP, Acesso em 20/09/2022.

[4] Oliveira, L.S., Alcoforado, L.F., Ross, S.D., Simão, A.S. (2019). Implementando a AHP com R. *Anais do SER*, ISSN 2526-7299, v.4, n.2. URL: https://periodicos.uff.br/anaisdoser/article/view/29331

[4] Triantaphyllou, E., Shu, B., Nieto Sanchez, S., Ray, T. (1998). Multi-Criteria Decision Making: An Operations Research Approach. *Encyclopedia of Electrical and Electronics Engineering*, (J.G. Webster, Ed.), John Wiley & Sons, New York, NY, Vol. 15, pp. 175-186.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

El paquete biosensors.usc

Marcos Matabuena[1]

[1]CiTIUS, Centro Singular en Tecnoloxías Intelixentes, Universidad de Santiago de Compostela

RESUMO

El análisis de datos de biosensores es uno de los retos metodológicos más importantes en el desarrollo de los nuevos paradigmas clínicos de la medicina de precisión y digital, en especial cuando los pacientes se encuentran monitorizados en condiciones libres de vida. Ejemplos paradigmáticos de esta situación pueden venir datos por los datos de aceleración de los dispositivos móviles que permiten estimar nuestro gasto energético, o el análisis a lo largo del tiempo de los valores de glucosa de un paciente. Hasta la fecha, el enfoque más popular en la literatura consiste en resumir la información temporal proveniente de los biosensores en representaciones composicionales de naturaleza vectorial. En una serie de trabajos recientes, hemos generalizado esta idea a un contexto funcional [1, 2], y hemos mostrado que el nuevo método: i) consigue una mayor ganancia de información en tareas predictivas; ii) permite comparar y visualizar diferencias en los perfiles de actividad de los sujetos a lo largo de un continuo de intensidades registradas por el dispositivo a diferencia de los métodos existentes; iii) no necesita categorizar la información en intervalos, lo que puede introducir subjetividad en el análisis y ser altamente dependiente de la población de estudio analizada. El objetivo de esta charla es introducir brevemente los nuevas metricas composicionales funcionales de biosensores, y ilustrar como poder realizar distantas tareas de modelado estadístico (test de hipótesis, análisis de conglomerados, modelos de regresión) con el nuevo paquete biosensors.usc, mostrando las ventajas de nuestra propuesta frente a los métodos existentes. El ejemplo que usaremos consistirá en analizar datos de la monitorización continua de la glucosa de una base de datos pública de pacientes no-diabéticos.

Palabras e frases chave: biosensores; datos funcionales; datos composicionales; dispositives vestibles.

Referencias

[1] Matabuena, M., Petersen, A., Vidal, J. C., Gude, F. (2021). Glucodensities: a new representation of glucose profiles using distributional data analysis. Statistical methods in medical research, 30(6), 1445-1464.

[2] Matabuena, M., Petersen, A. (2021). Distributional data analysis of accelerometer data from the NHANES database using nonparametric survey regression models. Arxiv

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

TUGlabR, un paquete de R para juegos coalicionales

Iago Núñez Lugilde[1], Miguel Ángel Mirás Calvo[2], Carmen Quinteiro Sandomingo[3] y Estela Sánchez Rodríguez[1]

[1] Universidade de Vigo, SIDOR. Departamento de Estatística e Investigación Operativa
[2]Universidade de Vigo, RGEAF, Departamento de Matemáticas
[3]Universidade de Vigo, Departamento de Matemáticas

RESUMEN

La teoría de juegos se ocupa de modelar y estudiar procesos de decisión en los que intervienen varios agentes o individuos que se comportan de modo estratégico (competitivo o no cooperativo) o de modo cooperativo. Desde la publicación del libro *The Theory of Games and Economic Behavior* ([13]), esta disciplina ha ido creciendo y se han desarrollado diversos modelos y aplicaciones tanto para comprender las acciones estratégicas de los "jugadores" como para proponer repartos o soluciones en contextos en los que la cooperación es determinante. [5] es un libro de referencia que cubre gran parte de los modelos y soluciones clásicas.

El proyecto TUGlab (Transferable Utility Games laboratory, [7, 8]) nace en el año 2006, tratando fundamentalmente de resaltar los aspectos geométricos de la teoría de juegos cooperativos para 3 y 4 jugadores, sin preocuparse de la complejidad matemática de los cálculos. Más adelante, en [6], se comienza a desarrollar **TUGlabExtended**, en donde se tratan funciones para juegos con cualquier número de jugadores. **TUGlab-Web** (`http://TUGlabweb.uvigo.es/TUGlabWEB2/index.php`, [3]) es una plataforma online en la que se encuentran implementadas las funciones básicas de **TUGlab**, de modo que el usuario puede experimentar con juegos cooperativos simplemente introduciendo la función característica del juego a analizar. Esta plataforma la utilizan usuarios de diferentes nacionalidades en cursos de doctorado, másteres o en tareas de búsqueda de contraejemplos en sus investigaciones. Teniendo en cuenta el uso y la aceptación en la comunidad internacional, presentamos un paquete de teoría de juegos cooperaticos en **R**, **TUGlabR**, que es una extensión de **TUGlab**, en la que el usuario puede trabajar con juegos generales teniendo siempre presente la limitación inexcusable del *input* de la función característica (vector en $\mathbb{R}^{2^n-1}$ siendo n el número de jugadores).

[11] es un paquete exclusivamente dedicado a repartir recursos escasos, una clase de juego aplicable a problemas de bancarrota, reparto de impuestos, ... [1] y [12] son otros paquetes de **R** centrados en la teoría de juegos cooperativos. Una diferencia importante entre estos dos últimos paquetes y **TUGlabR** está por ejemplo en el número de jugadores que se pueden considerar y en los tiempos de computación. Además, **TUGlabR** contiene diversas funciones relacionadas con las investigaciones más recientes de los autores, como por ejemplo el cálculo del core-center ([2]) para determinadas clases de juegos, los juegos de las caras ([9]) o los valores ponderados ([4] y [10]). La incorporación de estas funciones permite analizar particularidades que pueden ser añadidas al juego, como situaciones en las que existen prioridades o asimetrías entre jugadores o coaliciones que deben ser consideradas.

Palabras y frases clave: juegos coalicionales, soluciones puntuales, soluciones de conjunto, comparación de métodos de reparto, visualización gráfica.

AGRADECIMIENTOS

Este trabajo está financiado por FEDER/Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, SPAIN, projects PID2021-124030NB-C33 y PID2019-106281GB-I00, (AEI / FEDER, UE), y por el Programa de axudas á etapa predoutoral da Xunta de Galicia, Consellería de Educación, Universidade e Formación Profesional, número ED481A 2021/325.

Referencias

[1] Cano-Berlanga S., Gimenez-Gomez J. M. y Vilella C. (2017). Enjoying cooperative games: The R package GameTheory. *Applied Mathematics and Computation* 305, 381–393.

[2] González-Díaz J. y Sánchez-Rodríguez, E. (2007). A natural selection from the core of a TU game: the core-center. *International Journal of Game Theory.* 36, 27–46.

[3] Grande Cougil, R.P., Mosquera Rodríguez, M.A. y Ramos Valcarcel, D. (2011). TUGlabWeb: Interfaz Web para juegos TU. *X Congreso Galego de Estatística e Investigación de Operacións.*

[4] Kalai, E. y Samet, D. (1987). On weighted Shapley values. *International journal of game theory.* 16(3), 205–222.

[5] Maschler, M., Zamir, S., y Solan, E. (2020). Game theory. Cambridge *University Press.*

[6] Mirás Calvo, D. (2008). Programas informáticos orientados a juegos TU. Proyecto fin de máster

[7] Mirás Calvo, M. A. y Sánchez Rodríguez, E. (2006). **TUGlab** users guide. http://mmiras.webs.uvigo.es//TUGlabGUIDE.pdf

[8] Mirás Calvo, M. A. y Sánchez Rodríguez, E. (2010). Herramientas informáticas de cálculo y representación gráfica para juegos TU. *La Gaceta de la RSME* 13 (1), 89–108.

[9] Mirás Calvo, M. A., Quinteiro Sandomingo, C. y Sánchez Rodríguez, E. (2020). The boundary of the core of a balanced game: face games. *International Journal of Game Theory.* 49, 579—599.

[10] Mirás Calvo, M. A., Núñez Lugilde, I., Quinteiro Sandomingo, C. y Sánchez Rodríguez, E. (2022). Coalitional-weighted Shapley values. *Preprint.*

[11] Núñez Lugilde, I., Mirás Calvo, M. A., Quinteiro Sandomingo, C. y Sánchez Rodríguez, E. (2022). ClaimsProblems: Analysis of Conflicting Claims, **R** package version 0.2.0.

[12] Staudacher, J. y Anwander, J. (2019). Using the R package CoopGame for the analysis, solution and visualization of cooperative games with transferable utility; R Vignette.

[13] Von Neumann, J. y Morgenstern, O. (2007). Theory of games and economic behavior. *Princeton university press.*

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

`knobi`: an R package implementing Known-Biomass Production Models

Anxo Paz[1], Marta Cousido-Rocha[1], Santiago Cerviño[1] and Maria Grazia Pennino[1]

[1]Instituto Español de Oceanografía (IEO-CSIC). Centro Oceanográfico de Vigo. Subida a Radio Faro, 50-52. 36390 Vigo (Pontevedra).

ABSTRACT

Understanding the assessment of fish stocks to make recommendations for their sustainable exploitation has become an essential part of the management of fisheries resources. Traditional Surplus Production Models (SPMs) are one of the most used assessment models for data-limited marine populations, which relates historical series of catch to historical fishing effort or indexes of relative biomass such as CPUE (catch per unit effort). An alternative line of research based on surplus production models named Known-Biomass Production Models (KBPMs) was developed (MacCall, 2002 [1]) based on the idea that the annual surplus production in an unfished stock is equal to differences in biomass between two consecutive years, and that, for a fished stock, the calculation of surplus production depends on catch.

In contrast to the traditional SPMs, KBPMs use as input data a biomass time series produced by other stock assessment model instead of a biomass index. The simplicity of the model allows us to consider, for example, multispecific approaches or environmental effects.

In spite of the useful applications of KBPMs, very few studies implemented this approach in practice. Perhaps this is because the scientific community has not yet observed the potential of these models. For this reason, we implement them in the R package `knobi`, available at `https://github.com/MERVEX-group/knobi`, highlighting their advantages and illustrating their use. This package allows: (1) the KBPM fit to the stock; (2) the retrospective analysis; (3) the estimation of the effects of environmental variability; and (4) the estimation of future projections for the stock.

Therefore, `knobi` package implements for the first time the KBPMs including tools for the environmental effect analysis and the projections. The package is user-friendly and it is hoped that it will serve the scientific community by providing a simple and powerful tool for the KBPM analysis.

Key words: Assessment model, surplus production, climate change.

INTRODUCTION

The sustainable exploitation of the fish stocks has become essential for the fisheries resources management. For the impact of fisheries and environmental factors on fish stocks understanding, mathematical and statistical techniques (termed assessment methods) can be applied. Depending the available data the options goes from simple data-limited methods to more complex age or length-structured methods.

In recent years, there has been an increasing research effort on developing methods that can generally improve the reliability of stock assessments in data-limited situations. Surplus production models (SPMs) are among the assessment methodologies recommended for this purpose, and derived from this, the Known Biomass Production Models (KBPMs), which are the objective of

this work, are being developed to improve fisheries management. Their simplicity facilitates the consideration of crucial aspects of fisheries management as the environmental variability or the multispecific approach.

METHODS

For the correct understanding of the Known biomass production models (KBPMs), implemented in `knobi`, knowledge of the surplus production models (SPMs) framework is required. Then, SPMs are introduced for then focus in the KBPMs formulation.

Traditional SPMs are one of the most used assessment models for data-limited marine populations and they have the following general structure,

$$B_{t+1} = B_t + f(B_t) - C_t$$

where B_t is the stock biomass at the beginning of the year t or at the end of the previous one, C_t is the biomass caught among the year t and $f(B_t)$ is the biomass production function of year t.

These SPMs relates historical series of catch to historical fishing effort or indexes of relative biomass such as CPUE (catch-per-unit-effort). An alternative line of research based on surplus production models named known-biomass production models (KBPMs) was introduced by MacCall (2002 [1]), based on the idea that the annual surplus production in an unfished stock is equal to $B_{t+1} - B_t$, and that, for a fished stock, the calculation of surplus production depends on catch.

$$SP_t = \overline{B}_{t+1} - \overline{B}_t + C_t \tag{1}$$

where SP_t is the surplus production, $\overline{B}_t$ is the average biomass or spawn stock biomass (SSB), $\overline{B}_t = (B_t + B_{t+1})/2$, and C_t represent the catch. The subscript t denotes time (years).

In contrast to the traditional SPMs, KBPMs use as input data a biomass time series produced by other stock assessment model instead of a biomass index. Then, surplus production is calculated from the known average biomass (of two consecutive years) and the observed catch using equation 1. Then, for the KBPM fit,

$$SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) \tag{2}$$

where r is the intrinsic population grown rate, K is the virgin biomass and p is the asymmetry parameter, used so that the production curve is not always symmetrical.

As mentioned above, KBPMs offer the possibility of considering environmental variability. In `knobi` these effects are included as additive and multiplicative effects in the general KBPM formulation 2. Then, the additive and multiplicative models will be, respectively,

$$SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) + cX_t\overline{B}_t; \qquad SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) exp^{cX_t}$$

being c the parameter that represents the effect of the environmental variable X_t, where t represents time (years).

`knobi` PACKAGE

In this section, `knobi` use and habilities are illustrated through the case study of northern hake (*Merluccius merluccius*) in the International Council for the Exploration of the Sea (ICES).

The first step is the model fit through the `knobi_fit` using the biomass or the spawning stock biomass (SSB) time series, the catch time series and the corresponding years (see Figure 1).

Once the KBPM fit is carried out, its robustness to the deletion of data is tested using the `knobi_retro` function (see Figure 2).

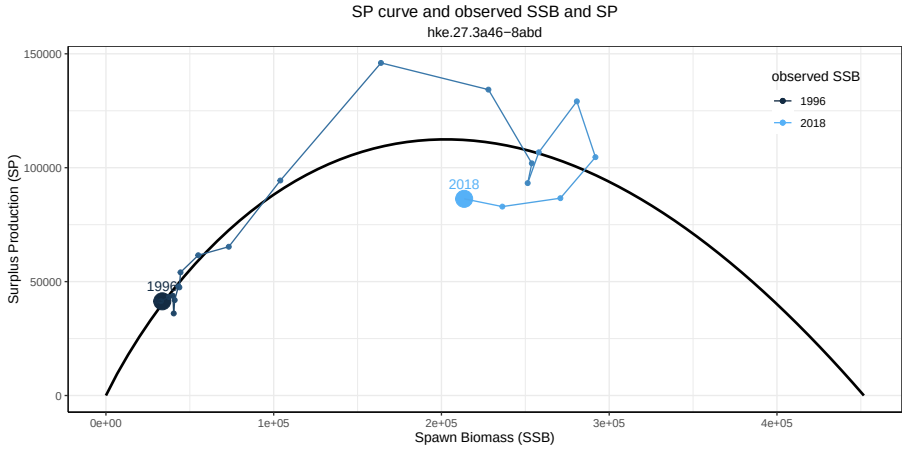


Figure 1: `knobi_fit` example. Estimated production curve derived from the KBPM fit for the Northern hake.

Figure 2: `knobi_retro` example. Production curves from the retrospective analysis for Northern hake.

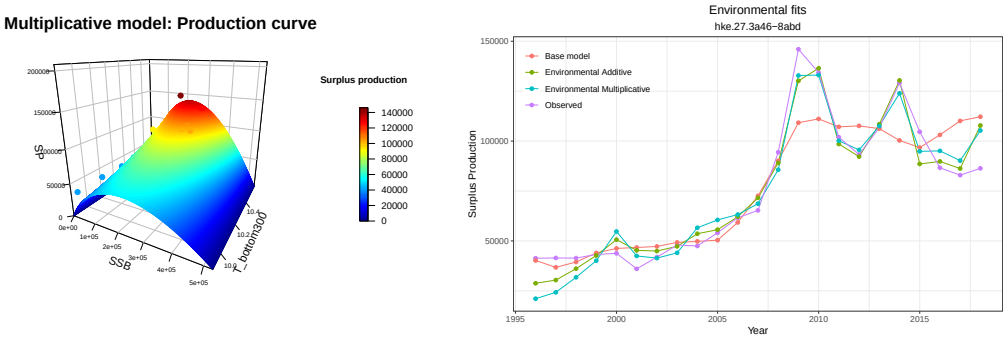
The environmental effects over the surplus production can be addressed through the `knobi_env` function. More precisely, the function analyse and model the relationships between surplus production and environmental variables (considering the possibility of testing different environmental lags according to its correlation with the surplus production) testing if changes in productivity are the response to environmental fluctuations (see Figure 3).

Finally, `knobi_proj` function allows us to project the stock biomass (or stock spawning biomass) time series and then the surplus production assuming certain catch or fishing mortality values for the projected years. Then, it allows us to analyze the future status of the stock under different possible settings of fishing pressure (see Figure 4).

CONCLUSIONS

`knobi` package implements for the first time the Known Biomass Production Models, providing a powerful tool for the stock status analysis from a surplus production point of view. Additionally, the package illustrates KBPMs potential and use, and highlights their advantages. As mentioned before, KBPMs simplicity facilitates the consideration of important aspects that influence the stocks dynamics as the climate change.

For the correct understanding of the package use, please check the available `vignettes` in the package help at `https://github.com/MERVEX-group/knobi`, where illustrative examples are available.



(a) Multiplicative environmental KBPM (b) Models fit and observations

Figure 3: `knobi_env` example. (a) Production curves according to the environmental variable values for multiplicative model in the Northern hake case study. (b) Predicted values of the different models (with and without environmental effects) compared with the observed values. The environmental variable is bottom temperature with two years of lag.

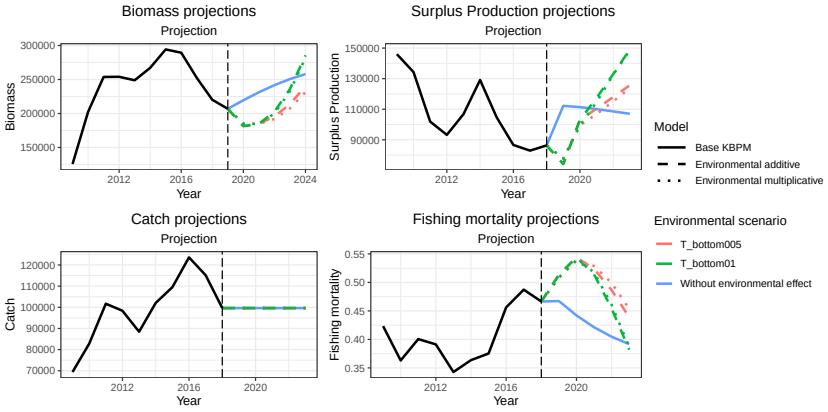


Figure 4: `knobi_proj` example. Projections for Northern hake for different future values of the environmental variable (bottom temperature) scenarios and for each model.

ACKNOWLEDGEMENTS

Proyecto financiado por la Unión Europea-NextGenerationEU. Componente 3. Inversión 7. Convenio entre el Ministerio de Agricultura, Pesca, Y Alimentación y la Agencia Estatal Consejo Superior de Investigaciones Científicas M.P. -A Través del Instituto Español de Oceanografía- Para impulsar la investigación pesquera como base para la gestión pesquera sostenible. Eje4, FishClim: Conocimiento científico para la adaptación al cambio climático del sector pesquero español

References

[1] MacCall, A. (2002). Use of Known-Biomass Production Models to Determine Productivity of West Coast Groundfish Stocks. North American Journal of Fisheries Management. 22. 272-279.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

Prototipado rápido con R: RAD de aplicacións Shiny e AEDA de datos

Mª Teresa Seoane-Pillado[1], Miguel Ángel Rodríguez-Muíños[2]

[1] Área de Medicina Preventiva e Saúde Pública, Departamento de Ciencias da Saúde, Universidade de A Coruña - INIBIC
[2] Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia

RESUMO

RAD (Rapid Application Delevopment): Hai uns anos, avalado polo rápido ascenso de R[1] como contorna de traballo no campo da análise de datos, apareceu un complemento/paquete de R que permitía desenvolver interfaces web en forma de cadros de mando e/ou visores interactivos. Este paquete, chamado Shiny[2], viño para quedar e elevou á súa máxima expresión a capacidade para dotar, a R, de contornas gráficas autónomas. O usuario pode interactuar directamente desde un navegador cunha aplicación, desenvolta en R, que non necesita do propio R no equipo cliente para ser executada.

Estas aplicacións desenvolvidas con Shiny constan, fundamentalmente, de dous compoñentes: ún na parte do servidor, que contén o código funcional propiamente dito da aplicación, e outro na parte do cliente que xenera a interfaz do usuario.

O proceso de xenerar a GUI (Graphical User Interface) pode resultar mais óptimo usando ferramentas RAD (Rapid Application Design) ou de prototipado rápido. Este tipo de aplicacións permiten aforrar tempo de desenvolvemento nos compoñentes gráficos da interface e consiguen código mais limpo e optimizado sobre o que comenzar a desenvolver a nosa aplicación.

No caso concreto do Shiny, existen varios proxectos de recente creación (algún deles todavía en fase Alpha) que se poden usar para esbozar prototipos das nosas aplicacións Shiny. Falaremos sobre dous deles: Designer[3] e ShinyUIEditor[4].

Estes dous proxectos permitennos deseñar os nosos wireframes/mockups/prototipos do GUI da aplicación Shiny que esteamos a desenvolver, na que poderemos insertar obxectos facendo drag-and-drop (arrastrar e soltar) e modificar as súas propiedades. As imaxes [figura 1] e [figura 2] son exemplos das contornas de desenvovemento dos paquetes {designer} e {shinyuieditor}, respectivamente.

Finalmente, o código R xenerado poderemos copialo ou descargalo para usar como base do módulo ui.R do noso proxecto. O resto do desenvolvemento (como a lóxica do servidor, aloxada no módulo server.r) deberemos programalo seguindo os métodos "tradicionais".

AEDA (Automated Exploratory Data Analysis): A Análise Exploratoria de Datos (AED) ten como obxectivo realizar unha investigación inicial sobre os datos resumindo as súas características a través de técnicas estatísticas e de visualización, e é un paso inicial crítico en calquera fluxo de traballo de Ciencia de Datos.

O paquete base de R dispón de "summary()" , unha función xenérica utilizada para producir resumos de resultados a partir de diferentes obxectos de entrada, como os conxuntos de datos. En particular, cando se proporciona un conxunto de datos (df) como

entrada, summary(df), devolve diferentes métricas (como a media, a mediana, o mínimo, o máximo, ...) para as columnas numéricas e a distribución (recontos) para as columnas categóricas. Tamén devolve información sobre os datos que faltan.

Falaremos sobre dous paquetes de R que facilitan/automatizan esta tarefa inicial e proporcionan un apoio substancial no manexo de datos, a visualización e a elaboración de informes. Estes paquetes son DataExplorer[5] e SmartEDA[6].

Estes paquetes aportan un potente conxunto de ferramentas que automatiza a maioría das tarefas de EDA proporcionando funcións para estatísticas descritivas, visualización de datos, táboas personalizadas e/ou Informes HTML. Cabe destacar que o paquete {DataExplorer} permite xerar un informe HTML completo invocando a función create_report() sobre un conxunto de datos e o paquete {SmartEDA}, análogamente, utilizando a función ExpReport().

As imaxes [figura 3] e [figura 4] son exemplos dos informes xenerados cos paquetes {DataExplorer} e {SmartEDA}, respectivamente.

Palabras e frases chave: RAD, AEDA, Mockup, Wireframe, prototipo, Shiny

Referencias

[1] R: https://www.r-project.org/

[2] Shiny: https://shiny.rstudio.com/

[3] Designer: https://ashbaldry.github.io/designer/

[4] ShinyUIeditor: https://rstudio.github.io/shinyuieditor/index.html

[5] DataExplorer: https://cran.r-project.org/web/packages/DataExplorer/

[6] SmartEDA: https://cran.r-project.org/web/packages/SmartEDA/

IMAXES

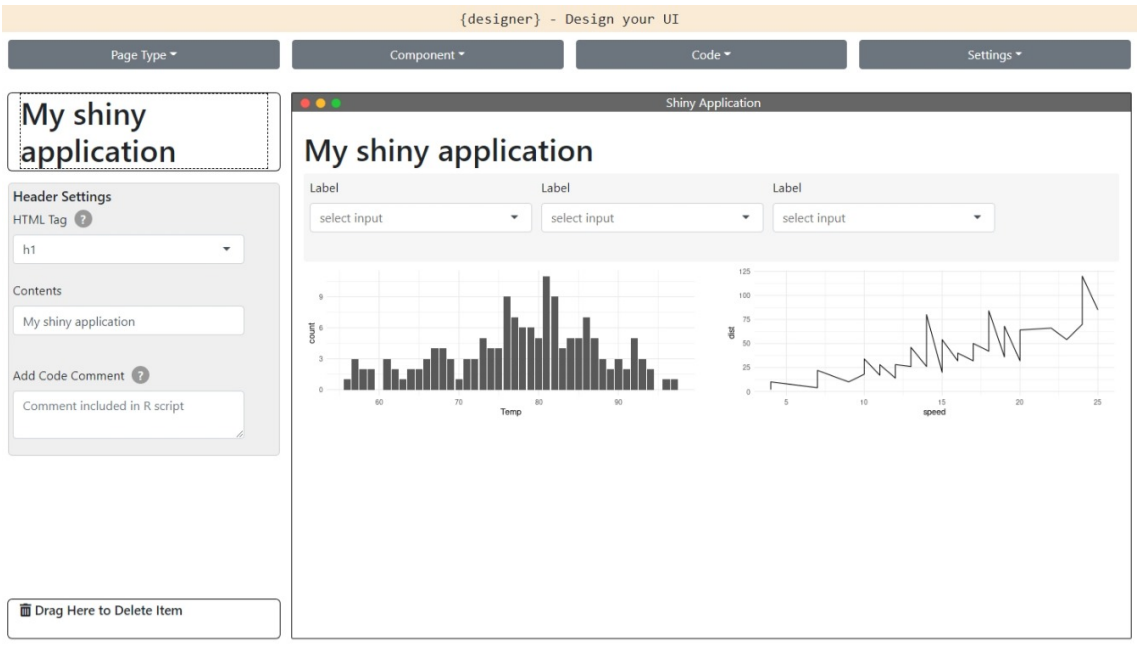


Figura 1: Interface do paquete {designer}

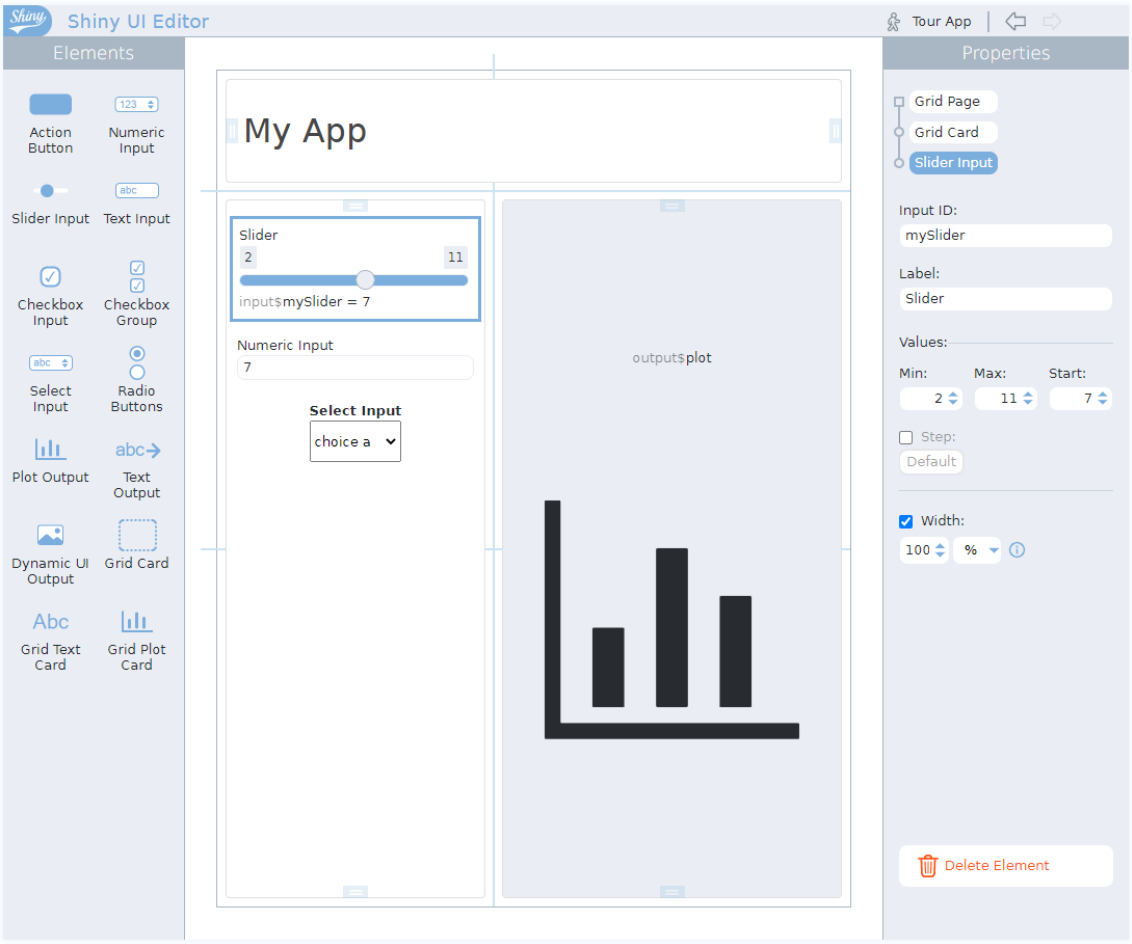


Figura 2: Interface do paquete {shinyuieditor}

Data Profiling Report

- Basic Statistics
  - Raw Counts
  - Percentages
- Data Structure
- Missing Data Profile
- Univariate Distribution
  - Histogram
  - QQ Plot
- Correlation Analysis
- Principal Component Analysis

Basic Statistics

Raw Counts

Name	Value
Rows	70,000
Columns	13
Discrete columns	0
Continuous columns	13
All missing columns	0
Missing observations	0
Complete Rows	70,000
Total observations	910,000
Memory allocation	3.7 Mb

Figura 3: Detalle do informe HTML xenerado co paquete {DataExplorer}

Exploratory Data Analysis Report

2022-09-16

Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

Overview of the data

Descriptions <chr>	Value <chr>
Sample size (nrow)	70000
No. of variables (ncol)	13
No. of numeric/interger variables	13
No. of factor variables	0
No. of text variables	0
No. of logical variables	0
No. of identifier variables	1
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	100% (13)

1-10 of 13 rows Previous 1 2 Next

Figura 4: Detalle do informe HTML xenerado co paquete {SmartEDA}

IX Xornada de Usuarios de R en Galicia

Santiago de Compostela, 20 de outubro do 2022

APLICACIÓNS DOS GRÁFICOS DE CONTROL PARA ESTIMAR A CARGA VIRAL DA COVID-19 NAS AUGAS RESIDUAIS

Claudia Torviso Rodríguez[1], Salvador Naya[2] e Javier Tarrío-Saavedra[2]

[1] Universidade da Coruña

[2] Grupo Modes, CITIC, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña

RESUMO

Neste traballo realízase unha análise dun caso de estudo con datos do número de contaxios e da carga vírica de COVID-19 nas augas residuais da área metropolitana de A Coruña. O obxectivo é avaliar se o uso de ferramentas como os gráficos de control, tanto os gráficos Shewhart, CUSUM e MEWMA, como aqueles máis específicos de uso en epidemioloxía, permite a detección temperá de brotes do virus SARS-CoV-2.

Palabras e frases chave: COVID-19, augas residuais, Gráficos de Control.

1. INTRODUCIÓN

Durante os primeiros meses do comezo da pandemia provocada pola COVID-19 e, sobre todo, durante o confinamento domiciliario e a "desescalada", a poboación estivo moi pendente das novidades que publicaban sobre o virus as autoridades políticas e sanitarias. Estes datos (número de contaxios, mortes, ingresos en UCI...) difundíanse de forma errónea, informando soamente da evolución de números que non eran entendibles para a poboación xeral.

O exceso de información, sumado á falta de rigor na toma e verificación dos datos, provocou un aumento da preocupación, niveis altos de ansiedade e problemas de sono [1].

Neste contexto, son fundamentais os procedementos que permitan identificar os sinais de alarma ante un potencial aumento dos contaxios de forma rápida. Por iso, propoñemos o uso de gráficos de control, co obxectivo de detectar sinais de alarma cando se produce un cambio significativo na pandemia. Para avaliar o funcionamento destes gráficos, estudouse a súa aplicación a datos reais recollidos na área metropolitana da Coruña, no marco do proxecto de referencia COVIDBENS.

2. GRÁFICOS DE CONTROL EN EPIDEMIOLOXÍA

Un gráfico de control representa o comportamento dun proceso ou dalgunha variable do mesmo rexistrando os datos ordenados no tempo. O obxectivo destes gráficos é detectar o máis rápido posible calquera cambio ou tendencia que poida afectar á calidade do proceso [2].

Existen grandes diferenzas no Control Estatístico de Procesos na industria e na medicina. No ámbito da saúde, a diferenza da industria, é moi común o uso de atributos (tipo de sangue, sexo, presenza de factores de risco...). Ademais, é necesario ter en conta este risco inherente aos datos, non podemos comparar dous pacientes con diferente idade ou con diferentes hábitos. Ao aplicar un proceso de eliminación de observacións fóra de control (o cal é moi común na industria), é probable que sigan aparecendo máis observacións fóra de control ao tratarse dun proceso que estea por completo fóra de control.
Os gráficos máis utilizados en medicina son os gráficos de control de Shewhart, pola súa variedade, fácil aplicación e intuitiva interpretación. Sen embargo, estes non son os máis adecuados para observacións individuais ou para detectar cambios pequenos, polo que será común utilizar gráficos CUSUM ou EWMA. Estes gráficos de control están dispoñibles en R dentro de varios paquetes, como os paquetes qcr [3] e qcc [4]. Se o que queremos é controlar a actividade hospitais ou médicos, aparte dos gráficos comparativos e dos gráficos de embude. Existe unha serie de gráficos, baseados nos gráficos Shewhart e CUSUM, especialmente deseñados para controlar procesos en epidemioloxía. Estas novas alternativas, utilizadas neste traballo, están dispoñibles no paquete "Surveillance" [5].

3. CASO DE ESTUDO DA ZONA METROPOLITANA DE A CORUÑA

O proxecto de COVIDBENS, impulsado por Edar Bens, en colaboración con investigadores da UDC e outros centros de investigación asociados, estudáronse as augas residuais da área metropolitana de A Coruña (municipios de A Coruña, Arteixo, Cambre, Culleredo e Oleiros) co obxectivo de analizar a cantidade de copias do virus presentes nesta área. O éxito do proxecto recaeu na alerta temperá ao conseguir detectar novos brotes ata 18 días antes de que se detectara unha alarma. Ademais, creáronse modelos estatísticos para estimar o número de portadores do virus e determinar a frecuencia das variantes do mesmo [5].
Con datos da carga vírica (número de copias do virus por litro de auga residual) e o número de casos activos na área metropolitana de A Coruña, estudamos algúns gráficos de control moi utilizados na industria e outros específicos para datos epidemiolóxicos.
En concreto, estudamos o funcionamento dos gráficos de control de Shewhart para medidas individuais e os gráficos CUSUM e EWMA. Estes gráficos non funcionan ben para datos autocorrelados, como é o caso da maioría das variables epidemiolóxicas, as cales dependen de datos anteriores no tempo. Para que estes gráficos funcionen ben é necesario corrixir esta autocorrelación. Neste traballo, corriximos este fenómeno a través da construción dun modelo ARIMA e do uso dos seus residuos. Unha vez corrixida a autocorrelación das variables, podemos construír gráficos de control que detecten olas e brotes da enfermidade, como podemos ver na Figura 1. Nesta figura, conséguese detectar coma fóra de control 8 residuos, correspondentes a incrementos grandes no número de casos activos de COVID-19. Porén, neste gráfico non se están a detectar tódolos brotes de COVID-19 e os que se detectan non parecen detectarse a tempo.
No paquete "surveillance" existen diferentes gráficos e métodos para o estudo e modelización de datos de tipo epidemiolóxico e das enfermidades infecciosas. De entre eses métodos, para este traballo, utilizamos o Sistema de Detección Temperá de Anomalías (do inglés, método EARS) e as súas 3 variantes. Este método detecta sinais de alarma cando a variable que esteamos estudando sobrepase un límite de control superior. A primeira variante (EARS C1) pódese interpretar coma un gráfico de Shewhart cun límite superior de control representado en azul e as sinais de alarma representadas cun triángulo vermello.
A segunda variante (EARS C2) constrúese de forma similar á primeira, pero cun pequeno desfase á hora de construír o límite de control, polo que tamén se pode interpretar coma un gráfico de control de Shewhart. Este gráfico é máis sensible aos cambios que a primeira variante, polo que detectará un maior número de brotes.

A terceira variante (EARS C3) difire máis das dúas, aínda que utiliza o estatístico da EARS C2 para a construción do límite superior. Este último gráfico pódese interpretar coma un gráfico CUSUM. Na Figura 2 obsérvase o control do número de casos activos (esquerda) e da carga vírica (panel dereito). Sendo máis sensible que as demais alternativas, se identifican de forma temperá practicamente tódolos brotes mediante un triángulo vermello.

4. CONCLUSIÓNS

Vimos que os gráficos de control poden aplicarse a datos de tipo epidemiolóxico sempre que teñamos en conta a posible autocorrelación dos datos. Estes gráficos, non obstante, poden non ser eficientes ao non dispoñer dun intervalo de tempo no cal os datos se manteñan estables e baixo control. Por iso, propoñémolo uso de ferramentas estatísticas baseadas no CEP como o método EARS. Das tres alternativas estudadas, a C3 compórtase mellor ao detectar máis brotes reais, ademais de detectalos na carga vírica antes de que aumente o número de casos.

AGRADECEMENTOS

Os autores agradecen o soporte e axuda do proxecto COVIDBENS, ao abeiro do cal foron tomados todos os datos usados neste traballo, así como aos seus coordinadores, Margarita Poza, Carlos Lamora, Ricardo Cao e Susana Ladra, e a todos os seus participantes.

Referencias

[1] Roy, D., Tripathy, S., Kar, S. K., Sharma, N., Verma, S. K., & Kaushal, V. (2020). Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic. Asian Journal of Psychiatry, 51, 102083. https://doi.org/10.1016/j.ajp.2020.102083

[2] Prat Bartés, A., Tort-Martorell Llabrés, X., & Grima Cintas, P. (2015). Métodos estadísticos: Control y mejora de la calidad. Universitat Politècnica de Catalunya.

[3] Flores, M., Fernández-Casal, R., Naya, S., Tarrío-Saavedra, J. (2021). Statistical Quality Control with the qcr Package. R Journal, 13(1), 194-217.

[4] Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. R News 4/1, 11-17.

[5] Salmon M, Schumacher D, Höhle M (2016). "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance." Journal of Statistical Software, 70(10), 1–35. doi:10.18637/jss.v070.i10.

[6] COVIDBENS. (2022). COVIDBENS. Detección del COVID-19 en las aguas residuales. Edar Bens. https://edarbens.es/covid19/

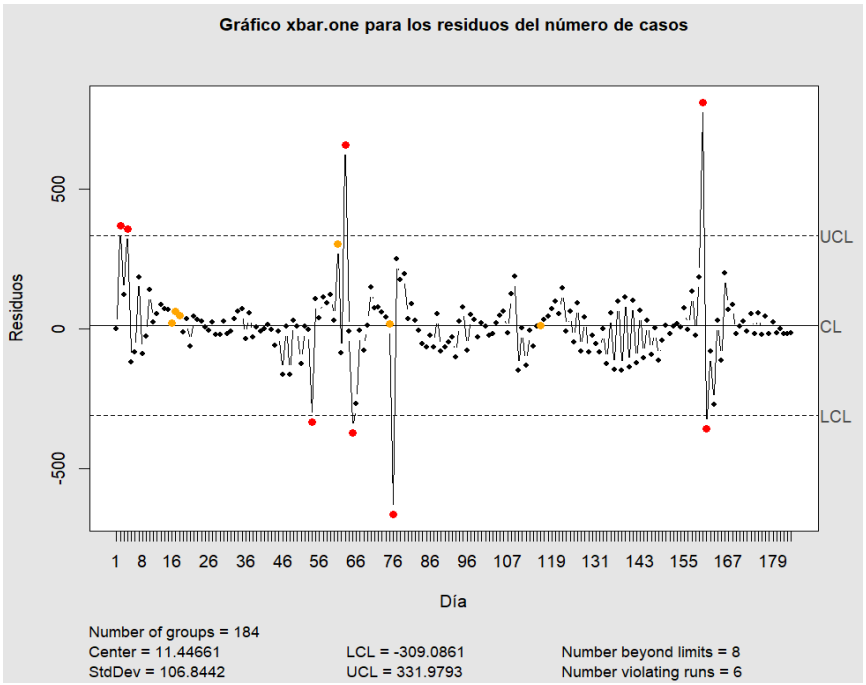


Figura 1: Gráfico de medidas individuais para os residuos da variable "número de casos".

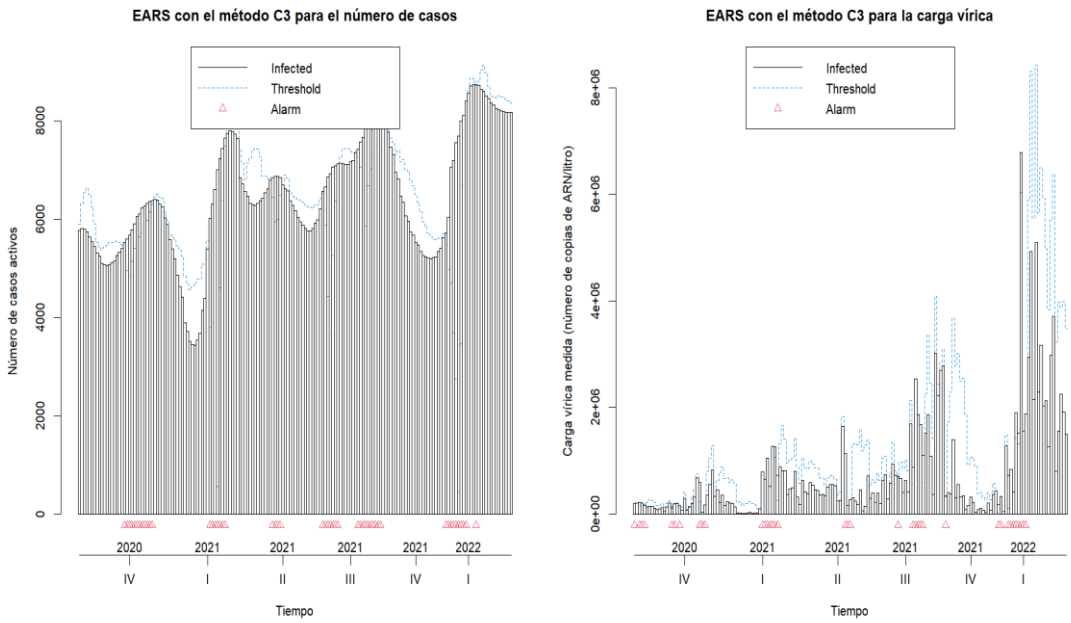


Figura 2: Gráficos EARS C2 para o número de casos (esquerda) e para a carga vírica (dereita).

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

LEARNINGSTATS: UN PAQUETE EN R PARA A DOCENCIA

Sabela Varela-Rey, María Isabel Borrajo[1,2], Mercedes Conde-Amboage[1,2] e Alejandra López-Pérez[2]

[1] CITMAga, 15782 Santiago de Compostela, España.
[2] Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.

RESUMO

A importancia da Estatística obriga ás/aos estudantes (bacharelato, grao, mestrado e doutoramento), e a investigadoras/es de diferentes ámbitos a aprender ferramentas estatísticas, e moitos empregan R, xa que é a linguaxe de programación por excelencia para a ciencia de datos. Porén, a necesidade de codificación, as saídas difíciles de entender e os cálculos escuros poden amedrentar aos usuarios. Co paquete *LearningStats*, tentamos democratizar a aprendizaxe e o uso da Estatística, simplificando a necesidade de habilidades de codificación e guiando ás/aos usuarias/os en cada paso. Este ambiente amigable, pero rigoroso, está deseñado para aplicar diferentes métodos estatísticos permitindo que a/o usuaria/o poida aprender a escoller a mellor opción en función do escenario no que está a traballar, e sacar así conclusións correctas.

O desenvolvemento do paquete *LearningStats*, xorde da constatación de dificultades comúns á hora de realizar unha análise de datos de estudantes de diferentes niveis e titulacións. Este paquete foi creado cunha finalidade didáctica, centrándose na sinxeleza de uso, na explicación reflexiva do cálculo estatístico, en proporcionar un bo soporte gráfico que axude á comprensión dos métodos e mellorar a extracción de conclusións rigorosas. Seguindo esta idea, incluso as posibles mensaxes de erro ou aviso que recibe a/o usuaria/o cando un método non se utiliza correctamente adoitan ser autoexplicativas e fáciles de entender (o que non sempre ocorre cos paquetes de R xerais).

O paquete cobre cinco eidos principais dentro da Estatística:

- Conxuntos de datos. O paquete inclúe conxuntos de datos útiles para implementar diferentes técnicas estatísticas e unha interface sinxela para ler datos, con soporte para numerosas extensións. Unha idea neste aspecto é ir incrementando os datos dispoñibles tanto con conxuntos propios como con datos de calquera que estea disposta/o a compartir.
- Estatística Descritiva. Incorpóranse funcións descritivas ata agora non presentadas en R, con explicacións moi detalladas nos ficheiros de axuda sobre o procedemento implementado en cada caso.
- Modelos de Probabilidade. Implementación didáctica de distribucións de modelos en función dos seus parámetros poboacionais, representando as súas principais funcións características.
- Inferencia Estatística. Achéganse ferramentas de inferencia básicas (intervalos e contrastes), moi utilizadas en diferentes eidos, focalizándose na intuitividade tanto para o cálculo como para os resultados.
- Regresión. O paquete integra, nestes momentos, o modelo de regresión lineal simple e o modelo ANOVA, sobre os que ademais de poder realizar certas tarefas de inferencia, van acompañados de representacións gráficas intuitivas que facilitan a comprensión dos mesmos.

Nesta charla presentaremos algunhas das ferramentas que inclúe o paquete *LearningStats* e que exemplifican o traballo desenvolto en cada un dos eidos arriba indicados. Ilustraremos as distintas funcións de xeito práctico coa fin de amosar o potencial didáctico das mesmas.

Palabras e frases chave: Paquete de R, docencia, Estatística.

Referencias

[1] Borrajo, M. I., Conde-Amboage, M. e López-Pérez, A. (2021). LearningStats: Elemental Descriptive and Inferential Statistics. R package version 0.1.0, `https://CRAN.R-project.org/package=LearningStats`.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

A new measure of dependence: distance correlation

María Vidal-García [1]

[1] CITMAga, 15782 Santiago de Compostela, España.
Departamento de Estatística, Análise Matemática e Optimización.
Universidade de Santiago de Compostela (USC).

RESUMO

When it comes to study the dependence between two univariate variables X and Y, Pearson correlation coefficient is usually the answer. This measure of linear dependence has many advantages, including its simplicity, how easy it is to interpret or its capacity to discriminate between positive and negative correlation. However, this coefficient has also important limitations when dealing with multivariate data or with general forms of dependence in non-gaussian variables. A wide range of alternative dependence measures has been proposed over the years to address these problems , we will focus in a recent approach which preserves some of the desirable properties of the Pearson correlation while being more flexible: the distance correlation.

In this talk we will present this new tool and assess its performance in different scenarios using the R package `energy`. Firstly we will explore how to compute the distance correlation between variables comparing its values with classical correlation. Then, we will further explore some applications such as distance-correlation-based tests of independence or goodness-of-fit tests. We will explain how to easily apply these methods in R using available functions of this package.

Graphical representations, comparisons and applications to data will be provided with illustration purposes.

Palabras e frases chave: dependence, independence, distance-correlation, correlation.

Referencias

[1] Rizzo, M. L. & Székely, G. J. (2022). energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-10, https://CRAN.R-project.org/package=energy.

[2] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6), 2769-2794.

[3] Tjøstheim, D., Otneim, H., & Støve, B. (2022). Statistical Dependence: Beyond Pearson's ρ. *Statistical Science*, 37(1), 90-109.

IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

clustcurv: Clustering of nonparametric curves

Nora M. Villanueva[1,2], Marta Sestelo[3,1], Javier Roca-Pardiñas[3,1] e Luís Meira-Machado[4]

[1]Dep. Statistics and O.R. & SiDOR Group, University of Vigo.
[2]CINBIO, Vigo, 36310, Spain.
[3]CITMAga, Santiago de Compostela, 15782, Spain.
[4]Centre of Mathematics & Department of Mathematics, University of Minho, Portugal.

ABSTRACT

One basic but important goal in the statistics field is the comparison of curves between groups. Several nonparametric methods have been proposed in the literature to test for the equality of nonparametric curves. In this framework, when the null hypothesis of equality of curves is rejected, it can be interesting to ascertain whether curves can be grouped or if all these curves are different from each other. Software in the form of an R package (clustcurv) has been developed in order to allow determining groups with an automatic selection of their number. The package can be used for determining groups in multiple survival curves as well as for multiple regression curves. The applicability of the proposed methods is illustrated using real data.

Keywords: Multiple Regression Curves; Multiple Survival Curves; Number of Groups; Cluster; R package chave

Referencias

[1] Villanueva, N. M., Sestelo, M., and Meira-Machado, L. (2019). A method for determining groups in multiple survival curves. *Statistics in Medicine*, 38:866 – 877.

[2] Villanueva, N. M., Sestelo, M., Meira-Machado, L., and Roca-Pardiñas, J. (2021a). clustcurv: An r package for determining groups in multiple curves. *The R Journal*, 13:164.

[3] Villanueva, N. M., Sestelo, M., Ordóñez, C., and Roca-Pardiñas, J. (2021b). An automatic procedure to determine groups of nonparametric regression curves. *arXiv*

AUTORES

IX XORNADA DE USUARIOS DE R EN GALICIA

```
y<-rnorm(12)
x<-1:12
plot(x,y,xaxt="n",cex.axis=0.8,pch=23,bg="gray"
col="black",cex=1.1,main="Uso de 'lines'
para dibujar una serie",cex.main=0.9)
axis(1,at=1:12,lab=month.abb,las=2,cex.axis=0.8
lines(x,y,lwd=1.5)
```

> ORGANIZA

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

ISBN 9 788409 448524