VIII XORNADA DE USUARIOS DE R EN GALICIA

14 de outubro de 2021

LIBRO DE RESUMOS

> ORGANIZA

> PATRONICAN

XUNTA DE GALICIA

VIII XORNADA DE USUARIOS DE R EN GALICIA

PROGRAMA E RESUMOS

14 de outubro de 2021

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

VIII XORNADA DE USUARIOS DE R EN GALICIA

Editora: María José Ginzo Villamayor

ISBN: 978-84-09-34662-2

Presentación

A Asociación de usuarios de software libre da Terra de Melide (MeLiSA) comprácese en presentar a VIII 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 dezasete relatorios ao longo de todo o día. Dos cales oito son convidados e ás outras sete atenderon á chamada de recepción de propostas. Durante a xornada teremos unha mesa redonda adicada a discutir sobre datos, R e COVID-19, presente nas nosas visas desde o 2020.

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 e, das tres universidades galegas, así como doutras nacionais como son á Universidad de Castilla-La Mancha (UCLM), Universidad de Oviedo, Universidad de Valencia, Universidad Complutense de Madrid, e dos seguintes centros de investigación: Centro de Investigación de Tecnoloxías da Información e das Comunicacións (CITIC), Instituto de Biomecánica de Valencia, International Centre for Numerical Methods in Engineering (CIMNE) e de empresas como Nextail ou Kstats e unha profesora do Colexio Manuel Peleteiro. Santiago de Compostela.

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

Santiago de Compostela, outubro de 2021

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

VIII XORNADA DE USUARIOS DE R EN GALICIA

Data

14 de outubro de 2021

Lugar de celebración

Salón de Graos. Facultade de Dereito (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 VIII Xornada.

Programa

14 de outubro de 2021

VIII XORNADA DE USUARIOS DE R EN GALICIA

09:15 – 09:35	Sesión de apertura Salvador Naya Ferández (*Vicerreitor de Política Científica, Investigación e Transferencia – Universidade da Coruña*), Rafael Rodríguez Gayoso (Melisa), María José Ginzo Villamayor (*Universidade de Santiago de Compostela – Comité Científico*)
09:35 – 10:00	HDiR: AN R PACKAGE FOR NONPARAMETRIC PLUG-IN ESTIMATION OF DIRECTIONAL HIGHEST DENSITY REGIONS Paula Saavedra Nieves e Rosa M. Crujeiras Casais. *Universidade de Santiago de Compostela*
10:00 – 10:25	DTDA: AN UPDATED AND EXPANDED R PACKAGE FOR THE STATISTICAL ANALYSIS OF DOUBLY TRUNCATED DATA Jacobo de Uña Álvarez. *Universidade de Vigo*
10:25 – 10:50	CONTROL ESTATÍSTICO DE PROCESOS MEDIANTE O PAQUETE qcr Salvador Naya Fernández (*CITIC e Universidade da Coruña*), Javier Tarrío Saavedra (*CITIC e Universidade da Coruña*), Miguel Flores (*SIGTI e Escuela Politécnica Nacional de Quito (Ecuador)*) e Rubén Fernández Casal (*CITIC e Universidade da Coruña*)
10:50 – 11:15	ANÁLISIS Y VISUALIZACIÓN DEL EXCESO DEL EXCESO DE DEFUNCIONES EN 2020 CON R Virgilio Gómez Rubio. *Universidad de Castilla la Mancha*
11:15 – 12:00	**PAUSA**
12:00 – 12:20	CORRECCIÓN DE EXAMES TIPO TEST, MEDIANTE R, NOS PROCESOS SELECTIVOS DE OPOSICIÓN DA XUNTA DE GALICIA Marcos Fernández Arias. *Xunta de Galicia*
12:20 – 12:40	ALGÚNS USOS DO R NOS PROCESOS DO INSTITUTO GALEGO DE ESTATÍSTICA Mª Esther López Vizcaíno. *Instituto Galego de Estatística*
12:40 – 14:00	MESA REDONDA: R e COVID (Moderador: Miguel A. Rodríguez Muíños, *Saúde Pública-Consellería de Sanidade*) Manuel A. Novo Pérez (*Universidade da Coruña*), Manuel Vaamonde Rivas (*Universidade da Coruña*), Javier Álvarez Liébana (*Universidad Complutense de Madrid*), Javier Kniffki (*Kstats*), María Jesús Hernández Vega (*CaixaBank*)
14:00 – 16:00	**PAUSA**
16:00 – 16:20	R NO BACHARELATO. ESTAMOS TOLOS OU QUE? Beatriz Padín Romero. *Colexio Manuel Peleteiro. Santiago de Compostela*
16:20 – 16:40	EXEMPLOS DE USO DE R NA EMPRESA Antonio Vidal Vidal. *Nextail*
16:40 – 17:00	D2MCS: UN PAQUETE EN R PARA DESENVOLVER E DESPREGAR AUTOMATICAMENTE UN SISTEMA MULTI-CLASIFICADOR Ferreiro Díaz (Universidade de Vigo, CINBIO, SERGAS), David Ruano Ordás (Universidade de Santiago de Compostela, SERGAS), José Ramón Méndez (Universidade de Vigo, CINBIO, SERGAS)
17:00 – 17:20	R-INLA: A FLEXIBLE TOOL FOR IMPLEMENTING BAYESIAN REGRESSION MODELS Alba Fuster Alonso (Universitat de València e Instituto Español de Oceanografía (IEO) de Vigo), S. Cerviño (Instituto Español de Oceanografía (IEO) de Vigo), David Conesa (Universitat de València), Marta Cousido Rocha (Instituto Español de Oceanografía (IEO) de Vigo), F. Izquierdo (Universitat de València) e M.G. Pennino (Universitat de València)
17:20 – 17:40	**PAUSA**
17:40 – 18:00	PRE-PROCESSING OF DAM DATA COUPLED WITH BEHAVIOR ANALYSIS THROUGH MACHINE LEARNING André Conde Vázquez e Fernando Salazar. *International Center for Numerical Methods Engineering*
18:00 – 18:20	ANÁLISIS DE INTERACCIONES CON EL PAQUETE "PHIA" Helios de Rosario Martínez. *Instituto de Biomecánica de Valencia*
18:20 – 18:40	MODELIZACION PARA LA AUTOMATIZACIÓN DEL CRECIMIENTO ECONÓMICO REGIONAL Priscila Espinosa Adamez e José Manuel Pavia. *Universidad de Valencia*
18:40 – 19:00	DIBUJANDO ARBOLES GENEALÓGICOS ACADÉMICOS CON LA LIBRERÍA visNetwork Arís Fanjul Hevia. *Universidad de Oviedo*

Índice

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

PRE-PROCESSING OF DAM DATA COUPLED WITH BEHAVIOR ANALYSIS THROUGH MACHINE LEARNING

Andre Conde[1], Fernando Salazar[2]

[1] International Center for Numerical Methods Engineering, aconde@cimne.upc.edu
[2] International Center for Numerical Methods Engineering, fsalazar@cimne.upc.edu

ABSTRACT

A predictive study framework for dams is proposed using Machine Learning (ML) algorithms programmed in R [1]. This study is capable of estimating the relationship between dam response and internal properties or external forces. For a better efficiency of the analysis with predictive models, a previous step has been developed consisting of database cleaning and expansion processes. An interactive interface for each of the two phases has been developed using the Shiny package [2] with the aim of making these processes accessible also to professionals who do not possess knowledge of laborious programming scripts.

Key words: Data preprocessing; Data interpretation; Machine learning; Dams.

1. INTRODUCTION

The current trend towards the use of hydroelectric power as a clean energy source entails the development of projects to increase the available generation capacity of such energy, which requires, among other things, detailed studies on dam safety. On the other hand, it is well known that climate change has modified the frequency and intensity of rainfall and floods, which affects the integrity of dams and can cause considerable damage. These two realities are reflected in the recent publication of new technical standards for dam safety, and their implementation will generate an increase of activity in the dam safety sector. For these reasons it is necessary to improve the techniques used for dam failure probability analysis. This can be achieved by taking advantage of advances in monitoring devices that allow greater accuracy and frequency of readings but, at the same time, requires advanced tools for the analysis of the amount of data recorded. With these objectives in mind, two applications have been developed for the processing of prey behavior data and their analysis with ML [3]. The developed tools are free to use and can be found at https://cimnetest.shinyapps.io/PREDATOR/ and https://cimnetest.shinyapps.io/SOLDIER/, making them available to dam engineers all over the world, even in less developed countries.

2. PRELIMINARY STEP: PREDATOR

Interpretation and expertise of technicians, based on fast and flexible data visualization, are essential to distinguish between complex relationships involving variables and errors or random noise [4]. Unfortunately, measurements obtained from monitoring networks often present errors due to sensor failures, transmission or storage errors as well as periods of missing data due to system outages. In other cases, changes in the systems or the coexistence of old manual records and more recent digitized data cause the same series to present different reading frequencies or different definitions of the same variable. Once the initial variables and the response to be analyzed have been selected (keeping in mind that ML algorithms allow us to consider any type of dam response), the developed tool allows the options explained later in this section.

-Previsualization in the form of: (i) interactive data tables; (ii) multivariate and multi-axis time series plots that facilitate identifying reading errors or periods of missing

data; (iii) dynamic scatter plots that allow detecting relative changes over time between variables in a family; (iv) quasi-3D and quasi-4D scatter plots using as an extra variable the color of the points. These latter plots include the possibility of setting limits for a variable, which facilitates the analysis of the relationship between variables during specific values of another variable.

-Filling periods of missing data, which is a common issue in auscultation measurement series. Some of the more conventional procedures have been implemented in software such as: (i) linear interpolation; (ii) K-Nearest-Neighbor imputation based on a selection of variables; (iii) parabolic interpolation; (iv) substitution by the mean of values recorded on the same calendar day/time of the previous and subsequent year/day; (v) substitution by a chosen fixed value.

-Data cleaning, which consists of identifying corrupted, incorrect or irrelevant data and then replacing or deleting them. In the current version, in addition to the previously mentioned options for missing data, the user can add a fixed value to the selected period or delete such values.

-Compensation of known alterations when you are aware of an event that affects the data (a maintenance operation, a punctual climatic effect...) and you do not want this event to influence the results. To compensate for this effect it is possible to use the options mentioned in the previous two paragraphs.

-Unification of data acquisition frequency to eliminate irrelevant/redundant information or noise, achieving a more efficient database when building predictive models. In this sense, the application allows the recorded series to be reduced to a new measurement frequency value (day, week, fortnight or month) through the following procedures: (i) taking as value the one recorded at a given hour of the day; (ii) taking maximum or minimum value for each period; (iii) adding values as average or sum of the measurements.

-Generation of new variables derived from the raw logs to enrich the set of initial variables. For example, moving averages help to improve the signal-to-noise ratio of the data series by replacing each value of the series by the average of a given number of previous values [5], while explicitly considering the time variable (year, month) allows to analyze the temporal evolution of the system behavior. Other derived variable options available are cumulative sums, linear combinations or variables indicating the increase/decrease (numerical or categorical).

3. ANALYSIS STEP: SOLDIER

It is common to use statistical models to generate predictions of response variables of hydraulic structures and to analyze the contribution of each of the acting loads. Linear regression approaches remain the most common for the study of dam behavior in both professional practice and research, although their limitations have also been identified [6]. A number of alternatives have been proposed from advanced statistical models to ML models. In this work, we generate ML-based prediction models using Boosted Regression Trees (BRT) [7] an algorithm that proved to be the most advantageous in a comparative study on early anomaly detection [8]. These BRT models are fitted to some of the available data (the training set) and then the model is applied to predict the response for the remaining period (the test or validation set).

Some BRT features that make it appropriate for this problem are: (i) the flexibility of the algorithm allows considering variables of different nature and range of variation without the need for additional transformations or determining a priori how they affect the target value; (ii) automated selection of more relevant variables in the response; (iii) it is robust with respect to training parameters so it does not require a deep knowledge of the algorithm for its use. The developed interface allows the user to select the predictor variables and the target variable as well as the values of the algorithm parameters (although the default values usually give good results). In addition, the main decision in this step is the selection of the training period with the complementary period being reserved for validation, this has proven important to control over-fitting of the results [9].

In addition to the possibility of making predictions for new sets of initial variables for which the response is not known, these models are also used to analyze the dam in two different ways as explained in the following sections.

3.1 Anomaly identification by comparing predictions

Under the predictive modelling approach, the reliability of the conclusions obtained is related to the accuracy of the model predictions. The developed application allows analyzing the results by comparing the obtained predictions with real readings showing the mean absolute error (MAE) and the coefficient of determination (R2) values. Additionally, it also shows the comparison of the predictions with the recorded behavior through its evolution over time in absolute terms and relative error (Figure 1). This comparison locates discrepancies in the performance of the dam. For example, the results in Figure 1 (left) show an increase in the prediction error for the validation period, this could be due to an over-fitting of the model or a change in the behavior of the response for the most recent period compared to the one used for model fitting.

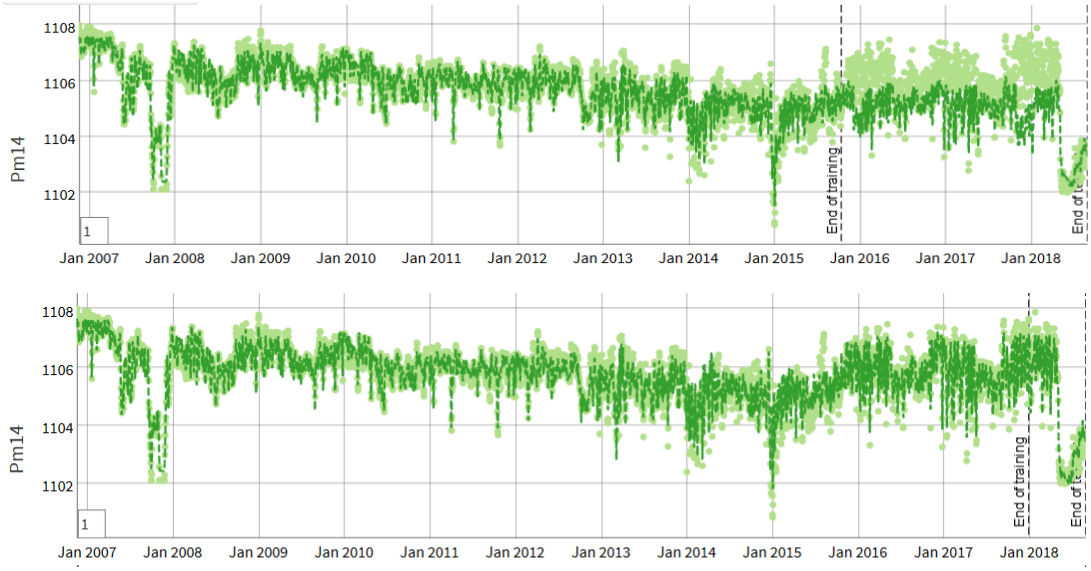


Figure 1: Comparisons between predictions (dark green) and records (light green).

Fitting a model with an extended training period, until the end of 2017, the accuracy of the tests increases (Figure 1 right). This confirms that there was some change in the system between 2015 and 2017 that affects the response of the target variable, such as a cleanup or refitting (as was the case analyzed, as reported).

3.2 Analysis of the effect of external actions on response

The conventional approach to identifying the variables with the strongest association with response assumes that the loadings are independent, which is not strictly true. In contrast, BRT measures relative influence on the basis that if a variable not associated with response is removed, the accuracy of the model should not change, and vice versa. Numerous papers have shown that BRT models allow consideration of nonlinear effects.

Figure 2 (left) shows the result in an example application where it can be noticed that the displacement is influenced by the reservoir level and the 90-day moving average temperature. It is interesting to see that the ambient temperature in a day presents a very low influence, which would reflect the thermal inertia of the dam. The ability of the algorithm to deal with highly correlated inputs is demonstrated by the fact that the small difference between Tair_90 and Tair_60 results in a large difference in significance. In the same way, the partial influence of each of the variables with respect to the response can be studied. It is easy to see in Figure 2 (right) how alterations of these two variables among low values of their range have little influence on the result, while a small variation in the upper values of their ranges causes a major change in the response.

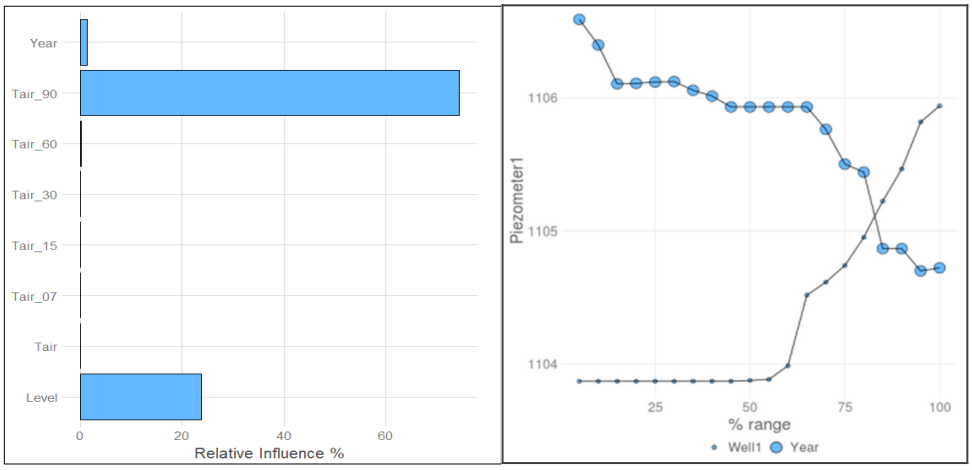


Figure 2: Analysis of global influence (left) and partial dependence (right).

3. CONCLUSIONS

The tools presented can be used both for the maintenance and control of existing dams and for the validation of designs for dam construction or improvement through: (i) identifying changes in the past behavior of the system; (ii) detecting deviations from normal operation in real time; (iii) determining the degree of association between input variables and response; (iv) predicting dam response to a set of actions; (v) locating areas of higher sensitivity. This last option is carried out by calculating the relative influence of the monitoring devices, so that the most relevant ones can be selected for new sensor installation or specific maintenance.

These applications can also be used for other types of problems and variables, as long as monitoring data are available.

However, the implemented model suffers from a limitation: in general, predictions outside the training data range are unreliable. This implies that the model prediction will be poor for a reservoir level higher than the historical maximum recorded. To control for this extrapolation effect, the test data are checked to see if they are within the training range, and a warning is issued in case of extrapolation.

References

[1] R Development Core Team, R. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (Vol. 1). Vienna.

[2] Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J. 2018. Shiny: Web Application Framework for R. R package version 1.1.0

[3] Salazar F., Kohler R., Conde A., Landstorfer F. (2021) Interpretation of Dam Monitoring Data Combining Visualisation Tools and Machine Learning. Eberlaste Dam Case Study. In: Bolzon G., Sterpi D., Mazzà G., Frigerio A. (eds) Numerical Analysis of Dams. ICOLD-BW 2019. Lecture Notes in Civil Engineering, vol 91. Springer, Cham.

[4] Guyon, I., Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR),* 3(3), 1157–1182.

[5] Dilawari, M. 2018. Forecasting models for the displacements and the piezometer levels in a concrete arch dam. Thesis. McGill University

[6] Salazar, F., Morán, R., Toledo, M. Á., & Oñate, E. Data-based models for the prediction of dam behaviour: a review and some methodological considerations. *Archives of computational methods in engineering,* 24(1), 1-21 (2017).

[7] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology,* 77(4), 802-813.

[8] Salazar, F., Toledo, M. A., Oñate, E., & Morán, R. 2015. An empirical comparison of machine learning techniques for dam behaviour modelling. *Structural Safety,* 56.

[9] Lever, J., Krzywinski, M. & Altman, N. Points of significance: Model selection and overfitting. *Nat. Methods* 13, 703–704 (2016).

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

ANÁLISIS DE INTERACCIONES CON EL PAQUETE "PHIA"

Helios De Rosario Martínez[1]

[1] Universitat Politècnica de València

RESUMEN

El paquete de R *phia* proporciona una serie de funciones para realizar el análisis *post-hoc* de las interacciones con factores en múltiples tipos de modelos estadísticos, según diversos métodos disponibles en la literatura.

Palabras e frases chave: Post-hoc, interacciones, modelos

1. INTRODUCIÓN

Uno de los objetivos habituales al realizar ajustes de modelos estadísticos es investigar qué influencia tienen los distintos factores que se introducen en los modelos. En los casos más sencillos, como los modelos lineales o variantes de los mismos, esto suele hacerse mediante análisis de la varianza (ANOVA) o de la covarianza (ANCOVA), a menudo complementados con análisis *post-hoc* cuando hay factores de más de dos niveles (por ejemplo en una comparación entre varias terapias más un placebo), para indagar en el detalle de las diferencias entre niveles.

Las técnicas de análisis *post-hoc* se basan esencialmente en comparar la respuesta del modelo aplicando variaciones en los niveles de los factores, con algún tipo de corrección para reducir el riesgo de errores de tipo I debidos a la realización de comparaciones múltiples. Pero la generalización de ese concepto a modelos con interacciones entre varios factores, o con interacciones entre factores y covariables (por ejemplo la interacción entre la terapia y el género o la edad de los pacientes), no es trivial.

Existen múltiples propuestas sobre tests *post-hoc* para las interacciones, desde las populares pero muy criticadas comparaciones de efectos principales simples [1], hasta el método de los contrastes residuales [2], o el de contrastes entre interacciones [3], entre otros muchos. Sin embargo no hay un criterio universalmente aceptado, y son habituales los errores de interpretación debido en parte al limitado soporte de las herramientas de software tradicionales para un análisis apropiado de los efectos de las interacciones [4].

Como respuesta a esa limitación, el paquete *phia* [5] proporciona unas utilidades para hacer el análisis *post-hoc* en modelos con interacciones complejas, o en modelos distintos del modelo lineal general, con opciones para adaptarse a distintos criterios publicados en la literatura, o para hacer análisis personalizados. En esencia, *phia* es una interfaz para construir tablas y tests de hipótesis mediante transformaciones lineales de modelos previamente ajustados, para lo que se usa la función "`linearHypothesis`" del paquete *car* [6]. Las dos herramientas principales que se proporcionan son las funciones "`interactionMeans`" y "`testInteractions`".

2. TABLAS Y GRÁFICOS DE INTERACCIONES

La función "interactionMeans" calcula una tabla de valores medios de la respuesta de un modelo (junto a sus errores estándar), para todas las combinaciones posibles de los factores involucrados en una interacción dada. A continuación se muestra un ejemplo, con un modelo lineal basado en el conjunto de datos de Boik, que viene incluido en el paquete *phia*, y modela la respuesta electrodérmica (EDR) de una muestra pacientes sujetos a distintas terapias y medicaciones [3]:

```
> mod <- lm(edr ~ therapy * medication, data=Boik)
> (boik.means <- interactionMeans(mod, c("therapy", "medication")))

  therapy medication adjusted mean std. error
1 control    placebo      50.20043   1.786533
2      T1    placebo      49.89963   1.786533
3      T2    placebo      45.69925   1.786533
4 control         D1      47.49899   1.786533
5      T1         D1      38.20065   1.786533
...
```

El objeto devuelto por "interactionMeans" se presenta por defecto como una tabla, pero también es graficable mediante la función "plot", que como se muestra en la Figura 1, representa las medias ajustadas y sus intervalos de confianza para factores simples (medias marginales, en los paneles diagonales) y para las parejas de factores incluidas en la tabla.

```
> plot(boik.means)
```

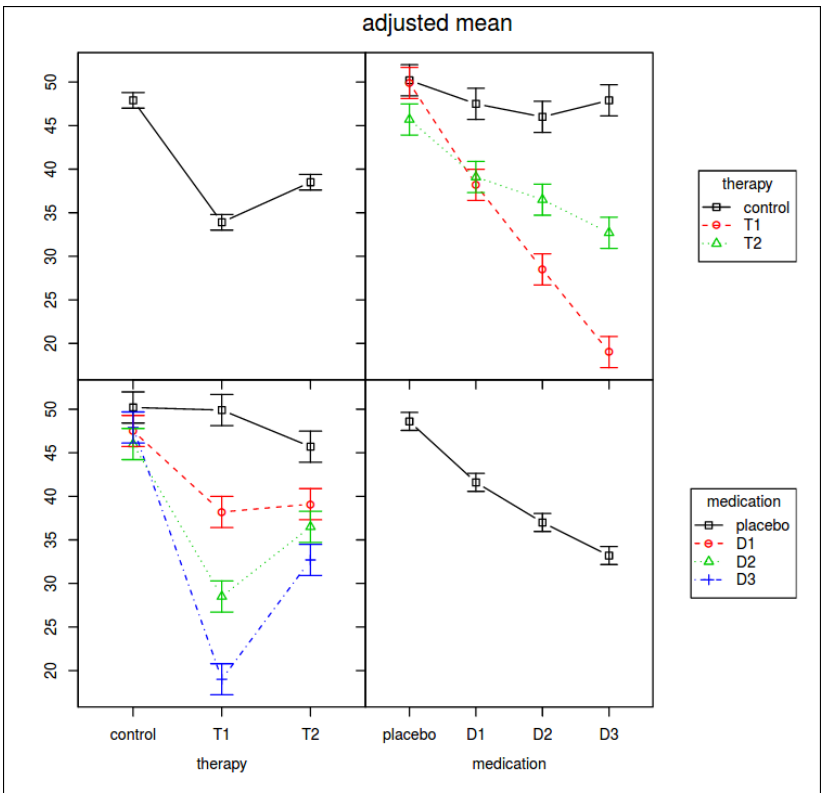


Figura 1: Medias ajustadas para un modelo con interacción de dos factores.

3. TESTS DE HIPÓTESIS SOBRE INTERACCIONES

Para hacer tests de hipótesis sobre las interacciones de un modelo se puede utilizar la función "testInteractions", cuya interfaz es semejante a la de "interactionMeans", pero en la que se puede indicar métodos específicos de análisis de interacciones. Por defecto se emplean los contrastes pareados, que representan "diferencias de diferencias":

```
> testInteractions(mod, c("therapy", "medication"))

F Test:
P-value adjustment method: holm
                          Value Df Sum of Sq       F    Pr(>F)
control-T1 : placebo-D1  -8.9975  1    121.43  6.3411 0.1448291
control-T2 : placebo-D1  -3.8985  1     22.80  1.1905 0.6951907
     T1-T2 : placebo-D1   5.0990  1     39.00  2.0365 0.6951907
control-T1 : placebo-D2 -17.1985  1    443.68 23.1687 0.0001563 ***
control-T2 : placebo-D2  -4.9984  1     37.48  1.9569 0.6951907
     T1-T2 : placebo-D2  12.2002  1    223.27 11.6587 0.0149653 *
...
```

En la tabla resultante podemos ver detalles de la interacción entre la terapia y la medicación, al comparar los cambios en la respuesta del modelo debido a usar distintas terapias, para distintas variaciones de la medicación. Por ejemplo en la cuarta línea, que muestra un contraste altamente significativo: se observa que la diferencia debido a usar la terapia de control en lugar de "T1" es 17.2 unidades menor cuando se administra un placebo, comparada con la diferencia que se da con medicación "D2".

La cantidad de contrastes pareados puede llegar a ser muy grande (en el ejemplo anterior la tabla completa contiene 18 contrastes), pues aumenta geométricamente con el número de factores cruzados y de niveles de los mismos. Para controlar el error de tipo I en esos tests de hipótesis, los valores *p* de las tablas se dan con una corrección por comparaciones múltiples (por defecto la corrección de Holm [7], que puede modificarse con el argumento "p.adjust").

Por otro lado, también se pueden analizar los contrastes residuales, efectos simples, contrastes ortogonales, y hasta contrastes personalizados mediante fórmulas arbitrarias, pasando los factores del modelo a argumentos específicos.

4. MODELOS CON COVARIABLES

Cuando hay covariables en los modelos, su influencia se caracteriza como una relación de proporcionalidad entre el valor de la covariable y la respuesta media del modelo. Esto suele representarse gráficamente con líneas cuya "pendiente" indica la fuerza de esa influencia. Ese tipo de representación, aunque intuitiva, es difícil de adaptar para mostrar las interacciones entre covariables y factores, por lo que en *phia* se opta por mostrar los valores de esas pendientes en lugar de las medias, cuando se usa el argumento "slope" con el nombre de la covariable en "interactionMeans" o "testInteractions". Por ejemplo, en la Figura 2 se presenta la interacción entre el tipo de ocupación (factor "type") y los ingresos (covariable "income") en un modelo lineal que trata de estimar el prestigio asociado a distintas profesiones [6]. En la gráfica superior, obtenida con el paquete *effects* [8], esa interacción se aprecia en las distintas pendientes de las rectas que representan el efecto de "income", según el valor de "type". En la gráfica inferior el valor de esas pendientes se muestra numéricamente en el eje Y.

5. ANÁLISIS DE OTROS MODELOS

Aunque los ejemplos mostrados se basan en modelos lineales generales, no existe una limitación a priori sobre qué modelos pueden analizarse con el paquete *phia*, mientras tengan una interfaz para consultar sus factores y coeficientes semejante a la de los modelos lineales. Concretamente, está comprobada la compatibilidad con modelos lineales generalizados ("glm"), no lineales ("lme", con el paquete *nlme*), y mixtos (con el paquete *lme4* [9]), y también pueden analizarse estudios de medidas repetidas ajustadas con modelos lineales multivariantes [10].

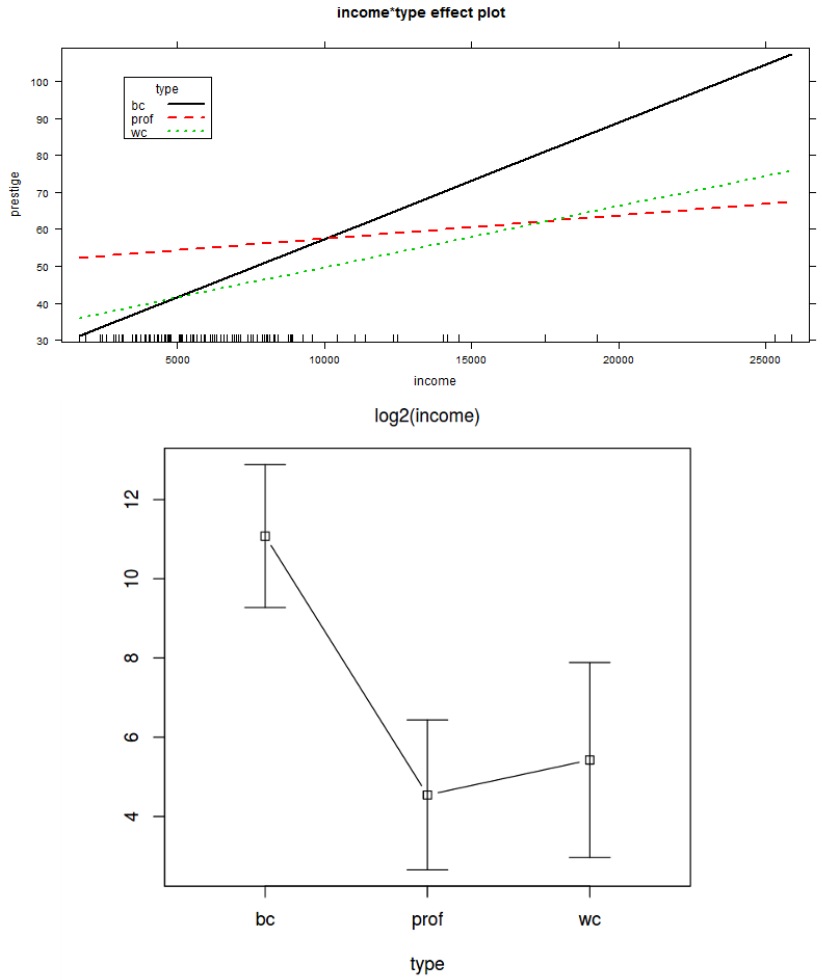


Figura 2: Gráfico de interacción entre factores y covariables con *effects* (arriba) y *phia* (abajo).

Referencias

[1] Games P.A. (1973). Type IV errors revisited. *Psychological Bulletin* 80(4), 304–307.

[2] Rosnow R.L., Rosenthal R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science* 7(4), 253–257.

[3] Boik R.J. (1979). Interactions, partial interactions, and interaction contrasts in the analysis of variance. *Psychological Bulletin* 86(5), 1084–1089.

[4] Pardo A., Garrido García J., Ruiz M.A., San Martín Castellanos R. (2007). La interacción entre factores en el análisis de la varianza: errores de interpretación. *Psicothema* 19(2), 343–349.

[5] De Rosario Martínez H. (2015). Post-hoc interaction Analysis. R package version 0.2-1. https://CRAN.R-project.org/package=phia

[6] Fox J. Weisberg S. (2011). An R Companion to Applied Regression. Sage, 2nd ed.

[7] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 6(2), 65–70.

[8] Fox S., Hong J. (2009). Effect displays in R for multinomial and proportional-odds logit models:Extensions to the effects package. *Journal of Statistical Software* 32(1), 1–24.

[9] Bates D., Mächler M., Bolker, B., Walker S. (2015). Fitting Linear Mixed-Effects Models Using *lme4*. *Journal of Statistical Software* 67(1), 1-48.

[10] Fox J. Weisberg S. (2011). Multivariate linear models in R. An appendix to *An R Companion to Applied Regression*. Sage, 2nd ed.

VIII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 14 de outubro do 2021

DTDA: AN UPDATED AND EXPANDED R PACKAGE FOR THE STATISTICAL ANALYSIS OF DOUBLY TRUNCATED DATA

Jacobo de Uña-Álvarez[1]

[1]Departament of Statistics and OR & CINBIO, Universidade de Vigo

ABSTRACT

Doubly truncated data [1] are often encountered in Survival Analysis and Epidemiology, among many other fields. They appear in particular with interval sampling, where the lifetimes are restricted to events within two specific dates [2]. In this talk I will introduce a recent update of DTDA package (version 3.0, April 11 2021) [3], joint work with Carla Moreira and Rosa M. Crujeiras. Compared to the original package launched more than one decade ago, DTDA v3.0 brings computational savings as well as new functions and real datasets. Specifically, DTDA v3.0 uses parallel computing to reduce the lengthy waiting times when bootstrapping under double truncation, and implements for the first time smoothing methods to estimate the density and hazard rate functions, including automatic bandwidth selection procedures [4]. Applications to proportional hazards regression [5, 6], accelerated failure time models [6], and competing risks [7, 8] will be given. Clinical data on AIDS incubation times and age at onset of Parkinson's disease, among others, will serve to motivate and illustrate the implemented techniques. Benefits of DTDA relative to other existing R packages to analyze doubly truncated data will be discussed.

Keywords: Epidemiology, Interval sampling, Nonparametric statistics, Random truncation, Survival Analysis, Time-to-event data

Referencias

[1] Efron, Bradley, and Vahe Petrosian. "Nonparametric Methods for Doubly Truncated Data". Journal of the American Statistical Association 94 (1999): 824-34.

[2] Zhu, Hong, and Mei-Cheng Wang. "Analysing Bivariate Survival Data with Interval Sampling and Application to Cancer Epidemiology". Biometrika 99 (2012): 345-61.

[3] Moreira, Carla, Jacobo de Uña-Álvarez, and Rosa M. Crujeiras. "DTDA: Doubly Truncated Data Analysis". R package version 3.0 (2021).

[4] Moreira, Carla, and Ingrid Van Keilegom I. "Bandwidth Selection for Kernel Density Estimation with Doubly Truncated Data". Computational Statistics and Data Analysis 61 (2013): 107-23.

[5] Mandel, Micha, Jacobo de Uña-Álvarez, David K. Simon and Rebecca A. Betensky. "Inverse Probability Weighted Cox Regression for Doubly Truncated Data". Biometrics 74 (2018): 481-7.

[6] de Uña-Álvarez, Jacobo, and Ingrid Van Keilegom. “Efron-Petrosian Integrals for Doubly Truncated Data with Covariates: an Asymptotic Analysis”. Bernoulli 27 (2021): 249-73.

[7] de Uña-Álvarez, Jacobo. “Nonparametric Estimation of the Cumulative Incidences of Competing Risks Under Double Truncation”. Biometrical Journal 62 (2020): 852-67.

[8] de Uña-Álvarez, Jacobo, Carla Moreira, and Rosa M. Crujeiras. The Statistical Analysis of Doubly Truncated Data: With Applications in R. Hoboken, NJ: Wiley, forthcoming.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

MODELIZACIÓN PARA LA AUTOMATIZACIÓN DEL CRECIMIENTO ECONÓMICO REGIONAL

Priscila Espinosa[1] y José Manuel Pavía[2]

[1] Departamento de Análisis Económico, Facultad de Economía, Universidad de Valencia
[2] Departamento de Economía Aplicada, Facultad de Economía, Universidad de Valencia

RESUMEN

Al igual que muchos otros agentes, los agentes económicos deben tomar decisiones en ambientes donde las condiciones son inciertas. Desde la última crisis y la actual pandemia, la COVID-19, el número y las fuentes de incertidumbre han aumentado en magnitud e intensidad. Los agentes regionales, más cercanos a la realidad local, buscan mecanismos capaces de mostrar sintéticamente la situación económica por la que atraviesa la economía regional. Este trabajo presenta la experiencia de la Comunitat Valenciana (España) en la generación de un modelo de conjuntos dinámicos de pronósticos económicos de diferentes organizaciones implementados en una Aplicación web desarrollada en Shiny. La aplicación permite la automatización del proceso anual de previsión económica, facilitando así su gestión por parte de los principales agentes económicos.

Palabras y frases clave: predicción, crecimiento económico, ensamble de modelos, modelos dinámicos, shiny, R.

1. INTRODUCCIÓN

Uno de los principales problemas a los que se enfrentan los distintos organismos regionales en el campo macroeconómico es la toma de decisiones en entornos con incertidumbre. La utilización de métodos cuantitativos de predicción [1] que utilizan las relaciones históricas entre las variables de interés permite disponer de información útil para la toma de decisiones en el corto y largo plazo, calibrando el grado de incertidumbre asociado a la misma. Los niveles de sofisticación técnica y el tiempo de que disponen los decisores para abordar esta tarea no siempre es óptimo, por lo que disponer de una aplicación interactiva, con un alto nivel de personalización y adaptada a sus necesidades y desde la que los usuarios pueden interactuar a través de diferentes widgets, facilitaría notablemente su trabajo. Este trabajo describe la interfaz gráfica de usuario desarrollada para tal fin en la Comunitat Valenciana, realizada con Shiny [2].

2. DATOS

La herramienta es flexible y puede adaptarse a distintos escenarios de disposición de información. Los datos empleados en la aplicación se dividen en dos grandes conjuntos. Por un lado, la información histórica de crecimiento económico correspondiente a los años comprendidos entre el año 2000 a 2020, ambos inclusive. Los datos de crecimiento económico se obtienen de forma automática de la web del INE y de la AiREF. Se emplean series en índices de volumen de producto interior bruto a precios de mercado nacional y regional del INE (Instituto Nacional de Estadística) a través de las Contabilidades Nacional de España (CNE), Regional de España (CRE) y Trimestral de España (CTR). Las cifras correspondientes a los instantes más actuales se obtienen de AiREF, cuyas estimaciones son congruentes con las cifras oficiales. Por otro lado, el segundo conjunto de datos,

Figura 1: Visualización de la aplicación del crecimiento económico regional.

relativo al benchmark de predicción para el crecimiento económico, lo constituyen las predicciones publicadas sobre la economía española realizadas por distintos organismos y centros de investigación. Al igual que en el primer conjunto, esta información va actualizándose en función de la disponibilidad del dato y debe ser facilitada por el usuario.

3. MODELIZACIÓN

Para la depuración de los datos y modelización de los mismos se ha utilizado el software estadístico `R` [3]. En primer lugar, se obtienen los índices de volumen de las series de crecimiento y se verifica si es necesario la corrección de ajuste de calendario y desestacionalización. Posteriormente, se equilibran al mismo horizonte temporal. Además, se genera el fichero con la predicción económica de los organismos/instituciones nacionales dedicados a la predicción económica.
Por último, la estimación final se logra mediante un ensamble de predicciones, obtenidas cada una de ellas mediante modelos dinámicos [4]. La especificación del modelo dinámico viene dada por la expresión 1:

$$y_t = \alpha + \phi y_{t-1} + \beta_{x_t} + \epsilon_t \tag{1}$$

Una vez estimados los parámetros del modelo, la ecuación se utiliza para estimar, a partir de los datos disponibles de las variables explicativas valores futuros de las variables a predecir $\hat{y}_{j,t+h|t}$, donde **j** es un índice que identifica el predictor y **h** el número de periodos hacia delante en que se basa la predicción.

4. DESARROLLO DE LA APLICACIÓN

El modelo anterior, así como todos los pasos previos para su implementación, han sido automatizados en una app desarrollada a media. Esto ofrece autonomía al usuario y le permite experimentar con diferentes conjuntos de datos, valorando cuál sería el impacto de la previsión de crecimiento regional. El desarrollo de la aplicación ha sido posible gracias a la utilización del software R a través de Shiny. La librería Shiny ha permitido desarrollar la aplicación de manera interactiva con un alto nivel de personalización y adaptarla a las necesidades requeridas. La página de inicio de la aplicación se ofrece en la Figura 1.

5. CONCLUSIONES

En este trabajo hemos mostrado como es posible la automatización de la predicción económica regional anual para la Comunitat Valenciana a partir de una aplicación web realizada con Shiny. Todo ello permite dotar a analistas y decisores regionales de flexibilidad a la hora de elaborar los presupuestos sin tener que ser expertos en matemáticas, estadística o en lenguajes de programación.

AGRADECIMIENTOS

El desarrollo de la aplicación web interactiva ha sido posible gracias la financiación de la Dirección General de Economía Sostenible (Consellería de Economía Sostenible, Sectores Productivos, Comercio y Trabajo; Generalitat Valenciana), a través de un convenio de colaboración (OTR2019-19290SUBDI) firmado con la Universitat de Valencia para el "desarrollo de las previsiones macroeconómicas de la economía valenciana" y sus perspectivas en el corto, medio y largo plazo y también gracias al Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2017-2020 en la PTA2018-015997-I, financiada por la Agencia Estatal de Investigación.

Referencias

[1] Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; Ljung, G. M., (2015) *Time series analysis: Forecasting and control*. John Wiley & Sons.

[2] Chang, W., Cheng, J., Allaire, J.J., Xie, Y. and McPherson, J. (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL http://www.R-project.org/.

[4] Petris, G.; Petrone, S.; Campagnoli, P., (2009) *Dynamic linear models. Dynamic linear models with R.* Springer, pp. 31–84.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

DIBUJANDO ÁRBOLES GENEALÓGICOS ACADÉMICOS CON LA LIBRERÍA visNetwork

Arís Fanjul Hevia

Departamento de Estadística e Investigación Operativa y Didáctica de la Matemática, Universidad de Oviedo

RESUMO

A la hora de analizar unos datos siempre ayuda el poder representarlos de forma ordenada y visual. La librería visNetwork de R permite dibujar redes con mucha flexibilidad, lo que puede ser útil en muchas situaciones. En este trabajo se explora esta herramienta y se utiliza para dibujar árboles genealógicos académicos.

Palabras e frases chave: árbol genealógico, redes, visNetwork.

1. LA LIBRERÍA VISNETWORK

La librería *visNetwork* [1] es una herramienta creada para la visualización de redes o grafos. Se puede exportar como una página web, lo que permite que se creen gráficos interactivos y compatibles con shiny o R Markdown. Esto hace que sean posibles de modificar fácilmente, aumentando y disminuyendo el tamaño o cambiando los nodos de sitio, y que se puedan poner etiquetas que ofrezcan más información al usuario al pasar el cursor por encima de los nodos.
Este paquete ofrece muchas opciones para el diseño del grafo, como el color, la forma, el tamaño o el sombreado de los nodos o como el color, el grosor, el sentido o la elasticidad de ejes. Permite destacar, cambiando su apariencia, determinados nodos (con sus ejes adyacentes) al seleccionarlos manualmente o según un criterio especificado previamente. También es posible establecer cierta jerarquía en el grafo, así como utilizar iconos o imágenes en lugar de los nodos.
Se puede encontrar una guía para esta librería en https://datastorm-open.github.io/visNetwork.

2. DIBUJANDO ÁRBOLES GENEALÓGICOS ACADÉMICOS

Se llamará árbol genealógico académico a aquel formado por individuos que han obtenido su doctorado y en el que existe una jerarquía que toma por antecesores a los directores o directoras de un determinado individuo y por descendientes a los estudiantes que se haya podido tener. La representación de este tipo de árboles en dos dimensiones puede resultar complicada, ya que puede haber relaciones académicas entre distintos niveles del árbol. En este trabajo se utiliza la librería visNetwork para representar dos árboles genealógicos académicos; en el primero se incluyen únicamente los ascendientes directos de la autora de este trabajo; en el segundo, mostrado en la Figura 1, se incluyen las relaciones académicas entre todos los más de cien egresados del Programa de Doctorado interuniversitario de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela, de la Universidade de A Coruña y de la Universidade de Vigo .
El proyecto *Mathematics Genealogy Project* (https://www.genealogy.math.ndsu.nodak.edu) se dedica a recoger la información de matemáticos de todas las partes del mundo, de forma que cada uno pueda construir su propio árbol genealógico matemático. *The Academic Family Tree* (https://academictree.org) es otro proyecto del mismo estilo que engloba cualquier disciplina académica.

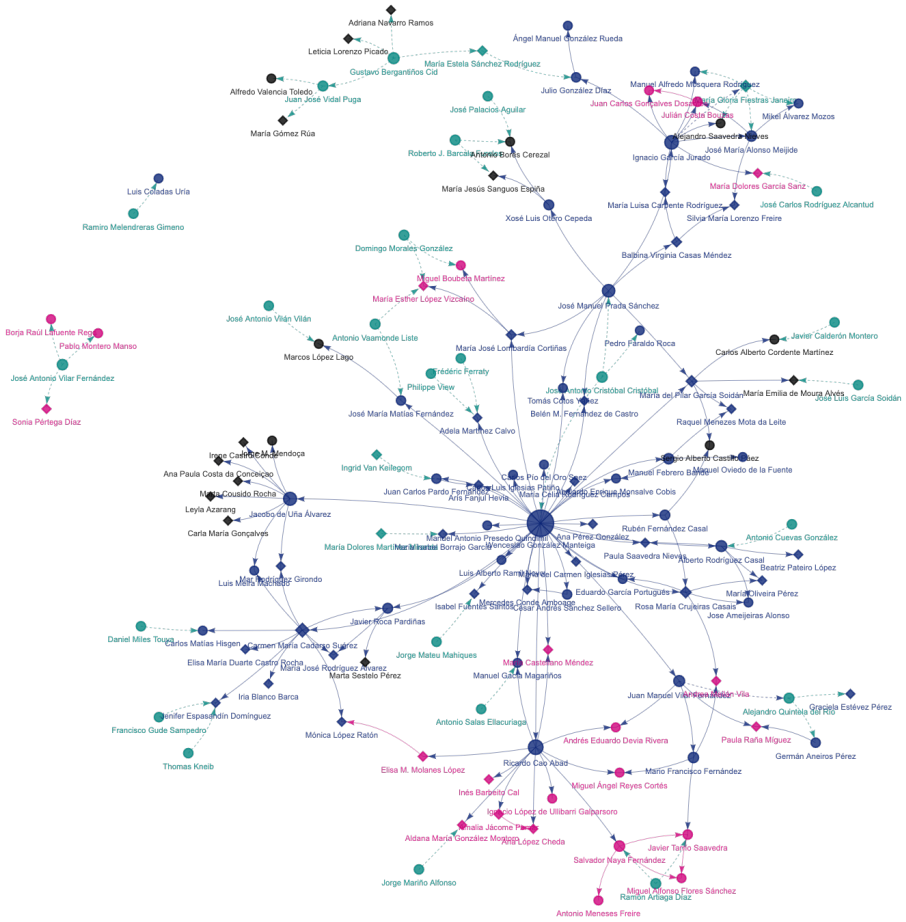


Figura 1: Árbol genealógico del Programa de Doctorado interuniversitario de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela (en azul), de la Universidade de A Coruña (en rosa) y de la Universidade de Vigo (en negro). En verde están representados aquellos directores que no fueron miembros de dicho Programa de Doctorado. Los círculos representan a los hombres y los rombos a las mujeres.

Referencias

[1] Almende B.V., Benoit Thieurmel and Titouan Robert (2019). visNetwork: Network Visualization using 'vis.js'Library. R package version 2.0.9. https://CRAN.R-project.org/package=visNetwork

VIII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 14 de outubro do 2021

Corrección de exames tipo test, mediante R, nos procesos selectivos de oposición da Xunta de Galicia

Marcos Fernández Arias
Axencia de Modernización Tecnolóxica de Galicia (AMTEGA)

RESUMO

Explicación técnica dos procesos de cálculo da corrección e asignación da puntuación dos exames tipo test das convocatorias de procesos selectivos (oposicións) da Xunta de Galicia, dende o ano 2021.

Para os cálculos utilizamos certas fórmulas, redactadas principalmente en R, con algo de Python e VBA.

Palabras e frases chave: OMR, corrección exames tipo test, stringr, purrr, reticulate

1. Lectura mediante scanner OMR

Mediante un scanner OMR (optical mark recognition) realizamos a lectura das follas de exame (figura 1) en papel para obter certos ficheiros (similares a CSV) do contido (figura 2).
Para certos detalles, resulta necesario utilizar Python, que o integramos no fluxo mediante o paquete *reticulate*.

2. Cálculos

Para a convocatoria creamos un ficheiro de configuración específico, no que constan os criterios de corte (cortes parciais, corte global, cupo de aspirantes...) e outros detalles de códigos identificativos da convocatoria.
As principais fórmulas dos cálculo son xenéricas: son aplicables a todas as convocatorias.
Utilizamos amplamente:

- o paquete *stringr* para manexo de cadeas de texto
- a sintaxe de programación funcional (mediante o paquete *purrr*)
- e a programación orientada a fluxo (mediante o paquete *tidyverse*)

3. Xeración de informes

Xéramos ficheiros en formato Excel, aos que mediante certo código VBA, aplicamos formato, para producir ficheiros PDF listos para publicación.
(figuras 4 e 5)

4. Análise de resultados

Realizamos certas análises dos resultados dos aspirantes en conxunto e tamén da efectividade das preguntas.
(figura 6)

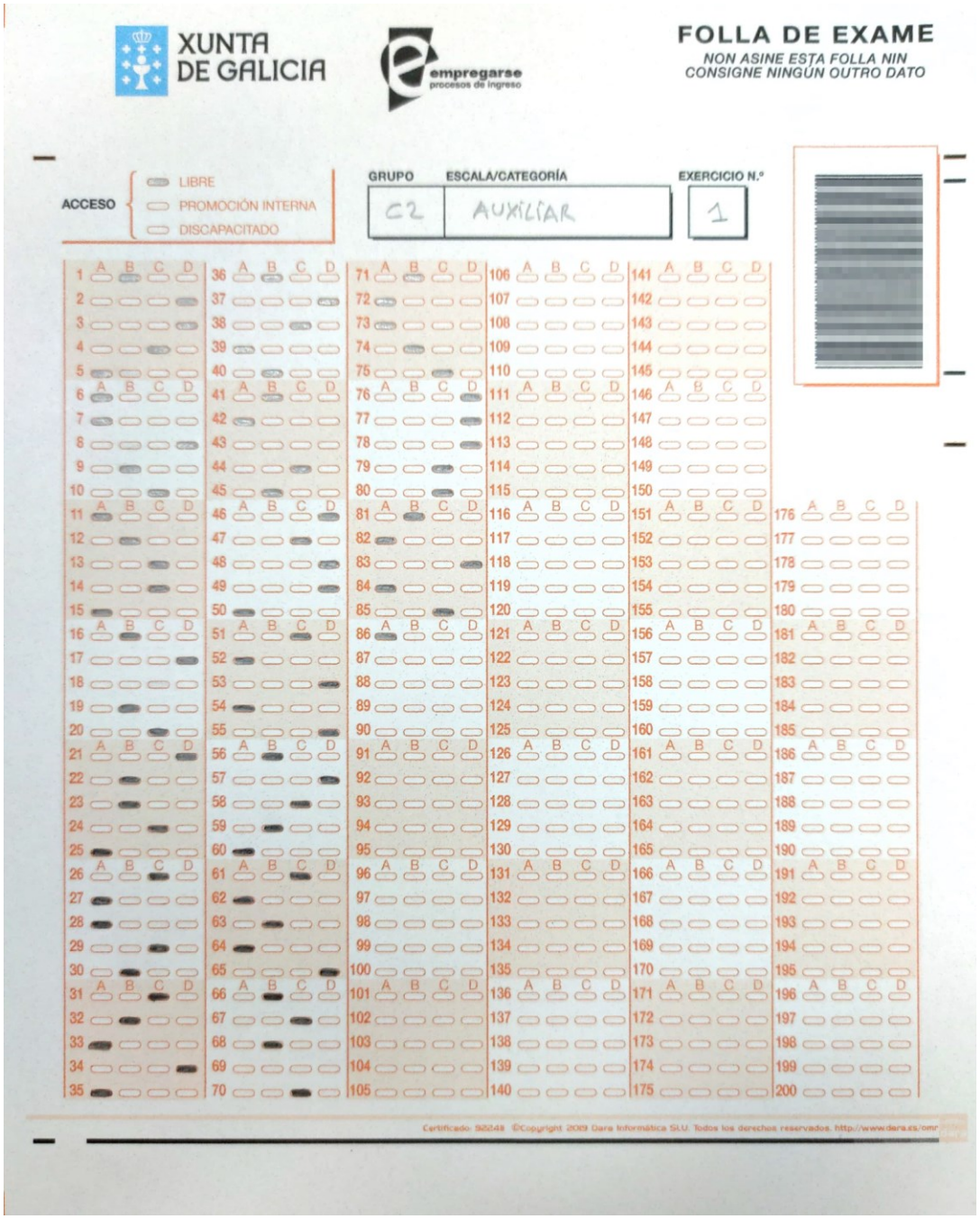
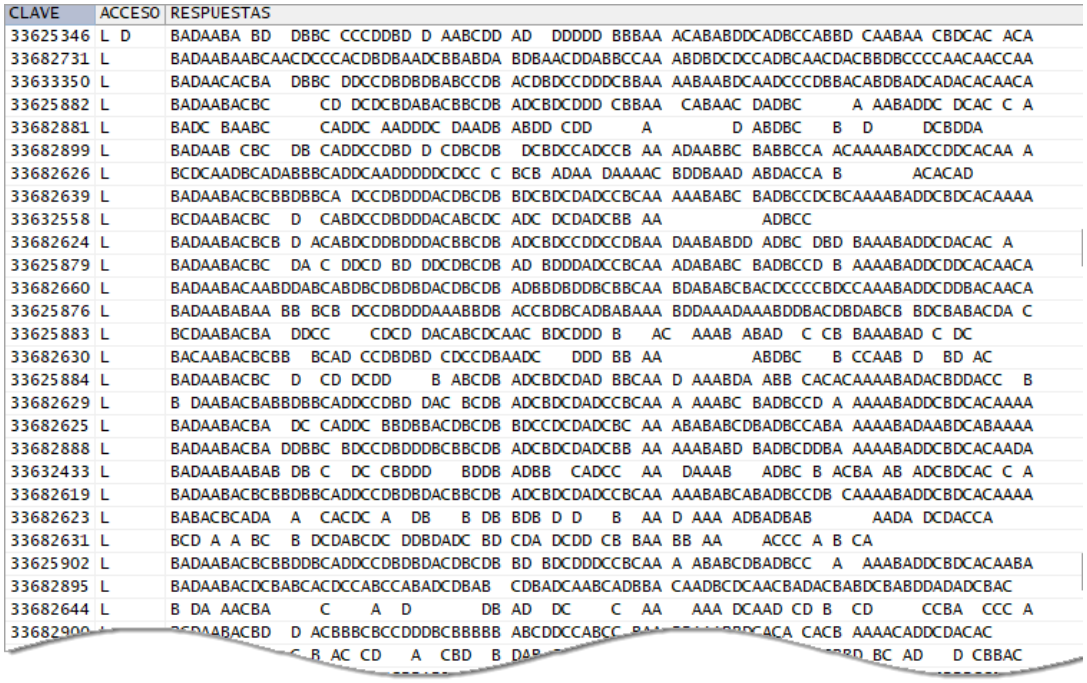
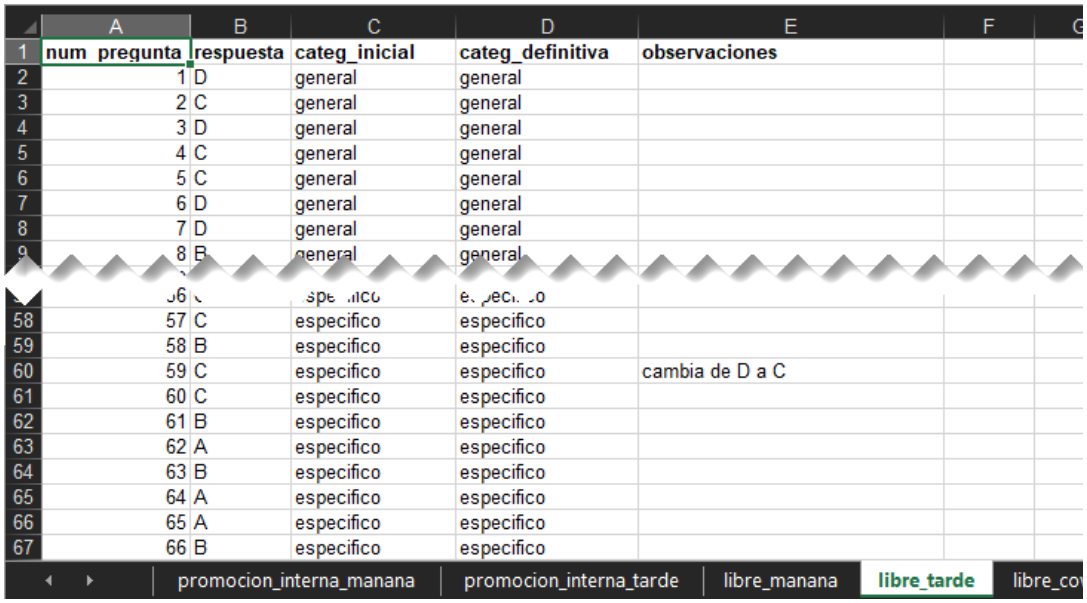


Figura 1: Folla de exame



CLAVE	ACCESO	RESPUESTAS
33625346	L	D BADAABA BD DBBC CCCDDBD D AABCDD AD DDDDD BBBAA ACABABDDCADBCCABBD CAABAA CBDCAC ACA
33682731	L	BADAABAABCAACDCCCACDBDBAADCBBABDA BDBAACDDABBCCAA ABDBDCDCCADBCAACDACBBDBCCCCAACAACCAA
33633350	L	BADAACACBA DBBC DDCCDBDBDBABCCDB ACDBDCCDDDCBBAA AABAABDCAADCCCDBBACABDBADCADACACAACA
33625882	L	BADAABACBC CD DCDCBDABACBBCDB ADCBDCDDD CBBAA CABAAC DADBC A AABADDC DCAC C A
33682881	L	BADC BAABC CADDC AADDDC DAADB ABDD CDD A D ABDBC B D DCBDDA
33682899	L	BADAAB CBC DB CADDCCDBD D CDBCDB DCBDCCADCCB AA ADAABBC BABBCCA ACAAAABADCCDDCACAA A
33682626	L	BCDCAADBCADABBBCADDCAADDDDDCDCC C BCB ADAA DAAAAC BDDBAAD ABDACCA B ACACAD
33682639	L	BADAABACBCBBDBBCA DCCDBDDDACDBCDB BDCBDCDADCCBCAA AAABABC BADBCCDCBCAAAABADDCBDCACAAAA
33632558	L	BCDAABACBC D CABDCCDBDDDACABCDC ADC DCDADCBB AA ADBCC
33682624	L	BADAABACBCB D ACABDCDDBDDDACBBCDB ADCBDCCDDCCDBAA DAABABDD ADBC DBD BAAABADDCDACAC A
33625879	L	BADAABACBC DA C DDCD BD DDCDBCDB AD BDDDADCCBCAA ADABABC BADBCCD B AAAABADDCDDCACAACA
33682660	L	BADAABACAABDDABCABDBCDBDBDACDBCDB ADBBDBDDBCBBCAA BDABABCBACDCCCCBDCCAAABADDCDDBACAACA
33625876	L	BADAABABAA BB BCB DCCDBDDDAAABBDB ACCBDBCADBABAAA BDDAAADAAABDDBACDBDABCB BDCBABACDA C
33625883	L	BCDAABACBA DDCC CDCD DACABCDCAAC BDCDDD B AC AAAB ABAD C CB BAAABAD C DC
33682630	L	BACAABACBCBB BCAD CCDBDBD CDCCDBAADC DDD BB AA ABDBC B CCAAB D BD AC
33625884	L	BADAABACBC D CD DCDD B ABCDB ADCBDCDAD BBCAA D AAABDA ABB CACACAAAABADACBDDACC B
33682629	L	B DAABACBABBDBBCADDCCDBD DAC BCDB ADCBDCDADCCBCAA A AAABC BADBCCD A AAAABADDCBDCACAAAA
33682625	L	BADAABACBA DC CADDC BBDBBACDBCDB BDCCDCDADCBC AA ABABABCDBADBCCABA AAAABADAABDCABAAAA
33682888	L	BADAABACBA DDBBC BDCCDBDDDBCBBCDB ADCBDCDADCBB AA AAABABD BADBCDDBA AAAABADDCBDCACAADA
33632433	L	BADAABAABAB DB C DC CBDDD BDDB ADBB CADCC AA DAAAB ADBC B ACBA AB ADCBDCAC C A
33682619	L	BADAABACBCBBDBBCADDCCDBDBDACBBCDB ADCBDCDADCCBCAA AAABABCABADBCCDB CAAAABADDCBDCACAAAA
33682623	L	BABACBCADA A CACDC A DB B DB BDB D D B AA D AAA ADBADBAB AADA DCDACCA
33682631	L	BCD A A BC B DCDABCDC DDBDADC BD CDA DCDD CB BAA BB AA ACCC A B CA
33625902	L	BADAABACBCBBDDBCADDCCDBDBDACDBCDB BD BDCDDDCCBCAA A ABABCDBADBCC A AAABADDCBDCACAABA
33682895	L	BADAABACDCBABCACDCCABCCABADCDBAB CDBADCAABCADBBA CAADBCDCAACBADACBABDCBABDDADADCBAC
33682644	L	B DA AACBA C A D DB AD DC C AA AAA DCAAD CD B CD CCBA CCC A
33682909	L	DAABACBD D ACBBBCBCCDDDBCBBBBB ABCDDCCABCC CACB AAAACADDCDACAC
		C B AC CD A CBD B DAB BC AD D CBBAC

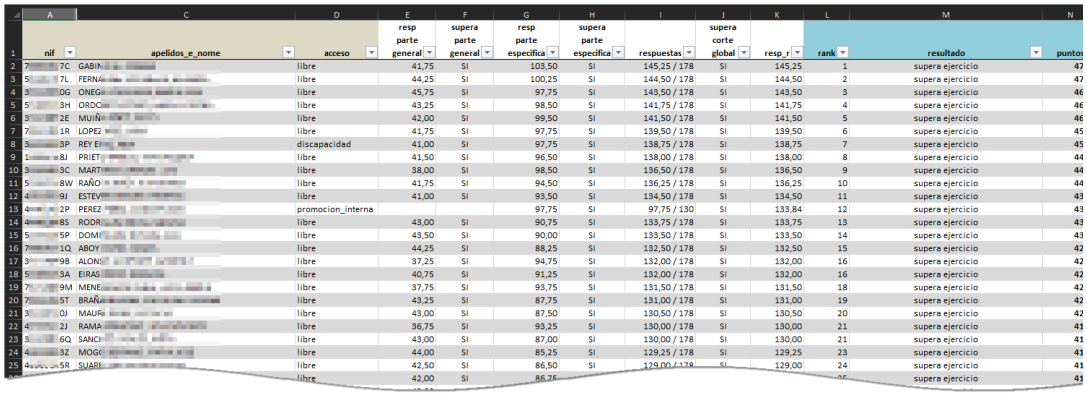
Figura 2: Datos lidos polo scanner OMR. Cada liña do ficheiro corresponde a un exame.



	A	B	C	D	E	F
1	num_pregunta	respuesta	categ_inicial	categ_definitiva	observaciones	
2	1	D	general	general		
3	2	C	general	general		
4	3	D	general	general		
5	4	C	general	general		
6	5	C	general	general		
7	6	D	general	general		
8	7	D	general	general		
9	8	B	general	general		
57	56	C	especifico	especifico		
58	57	C	especifico	especifico		
59	58	B	especifico	especifico		
60	59	C	especifico	especifico	cambia de D a C	
61	60	C	especifico	especifico		
62	61	B	especifico	especifico		
63	62	A	especifico	especifico		
64	63	B	especifico	especifico		
65	64	A	especifico	especifico		
66	65	A	especifico	especifico		
67	66	B	especifico	especifico		

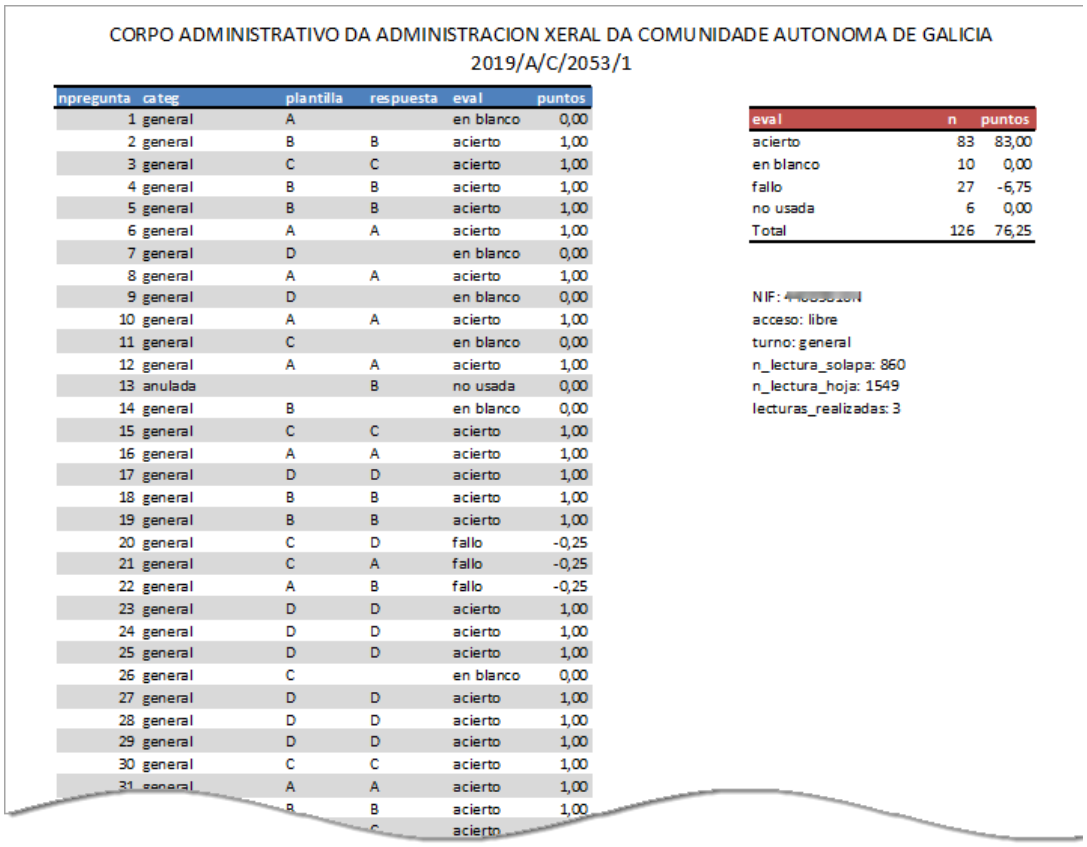
promocion_interna_manana | promocion_interna_tarde | libre_manana | libre_tarde | libre_co

Figura 3: Plantillas de corrección. (Poden ser varias para unha convocatoria: se tivo varias quendas.)



	A	C	D	E	F	G	H	I	J	K	L	M	N
1	nif	apelidos_e_nome	acceso	resp parte general	supera parte general	resp parte especifica	supera parte especifica	respuestas	supera corte global	resp_r	rank	resultado	puntos
2	7[illegible]7C	GABIN[illegible]	libre	41,75	SI	103,50	SI	145,25 / 178	SI	145,25	1	supera ejercicio	47
3	5[illegible]7L	FERNA[illegible]	libre	44,25	SI	100,25	SI	144,50 / 178	SI	144,50	2	supera ejercicio	47
4	3[illegible]0G	ONEG[illegible]	libre	45,75	SI	97,75	SI	143,50 / 178	SI	143,50	3	supera ejercicio	46
5	5[illegible]3H	ORDO[illegible]	libre	43,25	SI	98,50	SI	141,75 / 178	SI	141,75	4	supera ejercicio	46
6	3[illegible]2E	MUIÑ[illegible]	libre	42,00	SI	99,50	SI	141,50 / 178	SI	141,50	5	supera ejercicio	46
7	7[illegible]1R	LOPEZ[illegible]	libre	41,75	SI	97,75	SI	139,50 / 178	SI	139,50	6	supera ejercicio	45
8	3[illegible]3P	REY E[illegible]	discapacidad	41,00	SI	97,75	SI	138,75 / 178	SI	138,75	7	supera ejercicio	45
9	1[illegible]8J	PRIET[illegible]	libre	41,50	SI	96,50	SI	138,00 / 178	SI	138,00	8	supera ejercicio	44
10	3[illegible]3C	MART[illegible]	libre	38,00	SI	98,50	SI	136,50 / 178	SI	136,50	9	supera ejercicio	44
11	5[illegible]8W	RAÑO[illegible]	libre	41,75	SI	94,50	SI	136,25 / 178	SI	136,25	10	supera ejercicio	44
12	4[illegible]9J	ESTEV[illegible]	libre	41,00	SI	93,50	SI	134,50 / 178	SI	134,50	11	supera ejercicio	43
13	4[illegible]2P	PEREZ[illegible]	promocion_interna			97,75	SI	97,75 / 130	SI	133,84	12	supera ejercicio	43
14	4[illegible]8S	RODR[illegible]	libre	43,00	SI	90,75	SI	133,75 / 178	SI	133,75	13	supera ejercicio	43
15	5[illegible]5P	DOMI[illegible]	libre	43,50	SI	90,00	SI	133,50 / 178	SI	133,50	14	supera ejercicio	43
16	7[illegible]1Q	ABOY[illegible]	libre	44,25	SI	88,25	SI	132,50 / 178	SI	132,50	15	supera ejercicio	42
17	3[illegible]9B	ALONS[illegible]	libre	37,25	SI	94,75	SI	132,00 / 178	SI	132,00	16	supera ejercicio	42
18	5[illegible]3A	EIRAS[illegible]	libre	40,75	SI	91,25	SI	132,00 / 178	SI	132,00	16	supera ejercicio	42
19	7[illegible]9M	MENE[illegible]	libre	37,75	SI	93,75	SI	131,50 / 178	SI	131,50	18	supera ejercicio	42
20	7[illegible]5T	BRAÑ[illegible]	libre	43,25	SI	87,75	SI	131,00 / 178	SI	131,00	19	supera ejercicio	42
21	5[illegible]0J	MAURI[illegible]	libre	43,00	SI	87,50	SI	130,50 / 178	SI	130,50	20	supera ejercicio	42
22	4[illegible]2J	RAMA[illegible]	libre	36,75	SI	93,25	SI	130,00 / 178	SI	130,00	21	supera ejercicio	41
23	3[illegible]6Q	SANC[illegible]	libre	43,00	SI	87,00	SI	130,00 / 178	SI	130,00	21	supera ejercicio	41
24	4[illegible]3Z	MOGO[illegible]	libre	44,00	SI	85,25	SI	129,25 / 178	SI	129,25	23	supera ejercicio	41
25	4[illegible]5R	SUARI[illegible]	libre	42,50	SI	86,50	SI	129,00 / 178	SI	129,00	24	supera ejercicio	41
			libre	42,00	SI	86,75						supera ejercicio	41

Figura 4: Informe de resultados da convocatoria.



CORPO ADMINISTRATIVO DA ADMINISTRACION XERAL DA COMUNIDADE AUTONOMA DE GALICIA

2019/A/C/2053/1

npregunta	categ	plantilla	respuesta	eval	puntos
1	general	A		en blanco	0,00
2	general	B	B	acierto	1,00
3	general	C	C	acierto	1,00
4	general	B	B	acierto	1,00
5	general	B	B	acierto	1,00
6	general	A	A	acierto	1,00
7	general	D		en blanco	0,00
8	general	A	A	acierto	1,00
9	general	D		en blanco	0,00
10	general	A	A	acierto	1,00
11	general	C		en blanco	0,00
12	general	A	A	acierto	1,00
13	anulada		B	no usada	0,00
14	general	B		en blanco	0,00
15	general	C	C	acierto	1,00
16	general	A	A	acierto	1,00
17	general	D	D	acierto	1,00
18	general	B	B	acierto	1,00
19	general	B	B	acierto	1,00
20	general	C	D	fallo	-0,25
21	general	C	A	fallo	-0,25
22	general	A	B	fallo	-0,25
23	general	D	D	acierto	1,00
24	general	D	D	acierto	1,00
25	general	D	D	acierto	1,00
26	general	C		en blanco	0,00
27	general	D	D	acierto	1,00
28	general	D	D	acierto	1,00
29	general	D	D	acierto	1,00
30	general	C	C	acierto	1,00
31	general	A	A	acierto	1,00
		B	B	acierto	1,00
		C		acierto	

eval	n	puntos
acierto	83	83,00
en blanco	10	0,00
fallo	27	-6,75
no usada	6	0,00
Total	126	76,25

NIF: [illegible]

acceso: libre

turno: general

n_lectura_solapa: 860

n_lectura_hoja: 1549

lecturas_realizadas: 3

Figura 5: Informe individual (por aspirante).

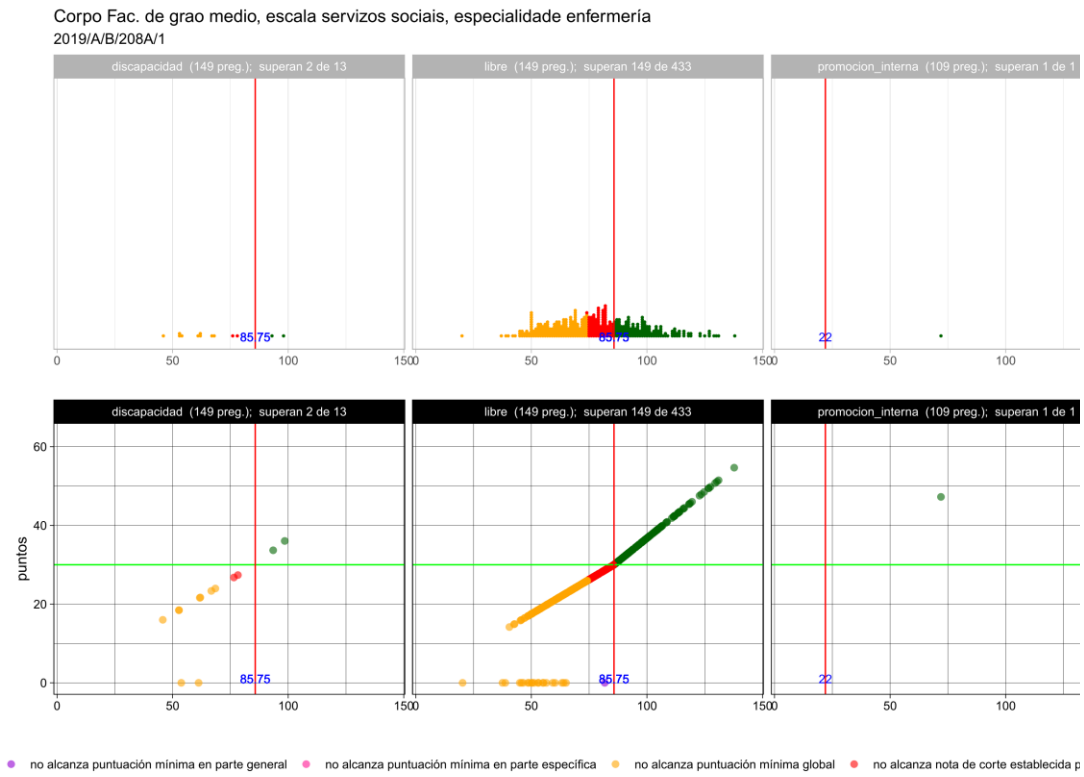


Figura 6: Gráficas de análise: histogramas e gráficas de proporción entre puntuacións.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

D2MCS: UN PAQUETE EN R PARA DESENVOLVER E DESPREGAR AUTOMATICAMENTE UN SISTEMA MULTI-CLASIFICADOR

Miguel Ferreiro-Díaz[1,2,3], David Ruano-Ordás[4,5] e José Ramón Méndez[1,2,3]

[1] Department of Computer Science, University of Vigo, ESEI - Escuela Superior deIngeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

[2] CINBIO-Biomedical Research Centre, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain

[3] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur). SERGAS-UVIGO, 36312 Vigo, Spain

[4] Department of Electronics and Computing, University of Santiago de Compostela, EPSI - Escuela Politécnia Superior de Ingeniería, Edificio Politécnico, Campus Terra, 27002, Lugo, Spain

[5] Computer Graphics and Data Engineering (COGRADE) Research Group

RESUMO

Nas últimas decadas, o expoñencial crecemento da cantidade de datos provocou tanto o aumento no número de observacións como na cantidade de características. Non obstante, este feito non implica un impacto positivo no adestramento dos modelos de Machine Learning debido principalmente á presenza dunha maior proporción de información redudante. Neste senso, o uso de Sistemas Multi-Clasificador (SMC) converteuse nunha opción moi adecuada para lidiar con problemas de alta dimensionalidade. Para iso, este traballo introduce D2MCS, un novo paquete capaz de diseñar automaticamente o máis adecuado SMC para un conxunto de datos de entrada. Ofrece mecanismos (empregando o paradigma OOP proporcionado polas clases R6) para automatizar a formación dos grupos de características máis apropiados e para identificar os modelos de Machine Learning máis precisos (desde un punto de vista de rendemento) para cada un dos grupos. Ademáis, para dotar de maior flexibilidade, D2MCS dispón da capacidade de deseñar e implementar facilmente novas estratexias de agrupamento de características e novos sistemas de votación.

Palabras e frases chave: paquete de R, R6, SMC, machine learning.

1. INTRODUCIÓN

Os avances tecnolóxicos das últimas décadas incrementaron considerablemente a xeración, almacenamento e manipulación dunha gran cantidade de datos procedentes de diferentes oríxenes (Ekbia

et al. 2015; Gandomi *et al.* 2015; Chen *et al.* 2014). Esto radica tanto no aumento do número de observacións (instancias) como no número de características (variables). Mentres que o primeiro permite mellorar a estimación da distribución da variable de interés e consecuentemente reducir o impacto negativo dos posibles datos atípicos (Burmeister *et al.* 2012), o segundo céntrase en aumentar o nivel de detalle dos datos a través dun incremento do número de dimensións. Non obstante, aínda que desde un punto de vista teórico o aumento de número de características debería incrementar a calidade dos datos, nun entorno real este feito non adoita a mellorar o rendemiento dos modelos de aprendizaxe automático xa que aumenta a probabilidade de dispor, tanto de datos con ruido (inconsistencias), como redundantes.

Durante a última década, os SMC consolidáronse como unha alternativa adecuada para manexar grandes volumes de información sen necesidade de recurrir aos enfoques de redución de dimensionalidade. Concretamente os SMCs están compostos por un conxunto de modelos cuxas saídas individuais (predicións) combínanse coa finalidade de obter unha única solución máis precisa. Este proceso componse de tres etapas principais: (i) división do espazo orixinal de dimensións en grupos de características relacionadas entre sí (clustering), (ii) adestramento de modelos de ML sobre os grupos de características previamente creados e (iii) obtención da predición final mediante a combinación dos resultados obtidos por cada modelo individual (votación). En (Woźniak *et al.* 2014; Roli, 2009; Giancinto *et al.* 2001) demostrouse cómo o uso de SMCs tende a superar o rendimento frente ao uso de clasificadores individuais. Non obstante, a principal desvantaxe deste enfoque radica na ausencia de técnicas adecuadas para (i) selecionar os grupos de características máis apropiados en función das particularidades de cada característica e as dependencias existentes entre elas, e (ii) determinar os modelos ML máis axeitados para cada grupo de características.

Para resolver esta situación, se iniciou o desenvolvemento dun framework en R denominado orientado á creación dun SMC para o ao cribado de medicamentos (*in silico screening*) no ámbito farmacolóxico D2MCS (Ruano-Ordás *et al.* 2019; Ruano-Ordás *et al.* 2019). Concretamente, o funcionamento de D2MCS divídese en catro etapas: (i) creación de grupos de características, (ii) construción de modelos de entrenamento, (iii) identificación dos modelos más axeitados e (iv) cálculo da predicción final.

Este artigo ofrece unha descrición completa das principais funcionalidades e recursos do paquete D2MCS (Ferreiro-Díaz *et al.* 2021). A versión actual é 1.0.0 e contén unha lista actualizada (con toda a colección de recursos) no documento do vignette e no manual de referencia. A sección seguinte ofrece unha descrición completa da estrutura e funcionalidades do paquete.

2. ESTRUCTURA DO PAQUETE

O paquete *D2MCS* basease na interacción de catro compoñentes asociados a cada unha das fases do proceso de construcción e execución do SMC: (i) manipulación de datos, (ii) agrupamento de características, (iii) creación do SMC e (iv) clasificación. A Figura 1 amosa en detalle cada unha das catro etapas, así como o fluxo da información para cada unha delas.

A primeira etapa céntrase en xestionar a carga dos datos de entrada (desde ficheiros CSV) así como habilitar a posibilidade de aplicar transformacións sobre os datos (como borrar columnas específicas ou con valores constantes, crear particións do conxunto inicial, etc). Os datos procesados almacénanse en dous tipos de estruturas en función da sua finalidade.

Durante a segunda etapa indentifícanse o número de grupos (k) que aseguren unha distribución óptima das características. Ademáis, co fin de garantir a compatibilidade da ferramenta independentemente dos tipos de relacións entre funcións, D2MCS proporciona unha interface capaz de: (i) invocar as estratexias de agrupamento incluídas por defecto, (ii) implementar e despregar novas estratexias personalizadas mediante un sinxelo esquema de herdanza, (iii) indicar as heurísticas que guían o proceso de creación de grupos de características en base a súa relación e (iv) personalizar a configuración base das diferentes estratexias. Por outro lado, can finalidade de estimar o rendemento da estratexia, incorporouse un avaliador que permite (i) identificar visualmente a

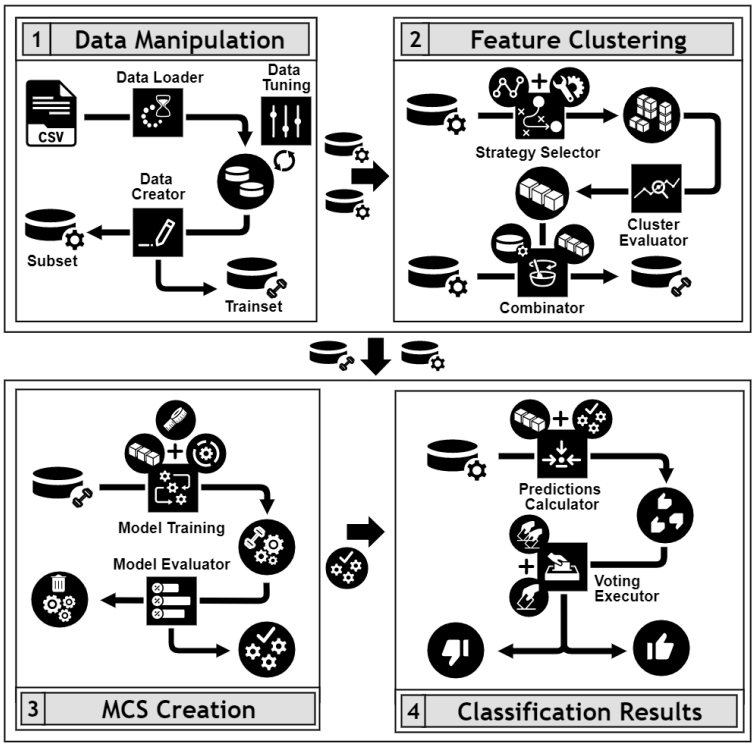


Figura 1: Diagrama de fluxo do funcionamento do paquete D2MCS.

composición de cada grupo de características e (ii) medir a calidade de cada grupo. Para mellorar a flexibilidade e eficiencia do proceso, D2MCS dispón dun mecanismo que permite (i) seleccionar automaticamente a distribución de características máis óptimas para a seguinte etapa ou (ii) personalizar manualmente a configuración dos grupos de características que se utilizarán nas siguientes fases. Mentres que o primeiro permite a elección da mellor distribución de características á ferramenta, o segundo engade tres niveis de personalización para que os usuarios especifiquen e definan detidamente a distribución de clústeres máis axeitada ás súas necesidades. Para iso, permítese (i) determinar o número de clústers desexados, (ii) seleccionar os grupos de características a usar e (iii) decidir a relevancia das características sen agrupadas (no caso de que existan).

A seguinte etapa (*MCS Creation*) ten como obxectivo realizar o adestramento de modelos de Machine Learning e identificar aqueles que ofrecen o mellor rendemento para cada un dos grupos de características previamente seleccionados. Para iso, esta fase incorpora tres funcionalidades que permiten (i) escoller os modelos para adestrar, (ii) definir as heurísticas que guiarán o proceso de optimización dos hiperparámetros dos modelos durante a fase de adestramento e (iii) visualizar o rendemento final obtido por cada modelo. Por outra banda, ademaís de incorporar heurísticas de adestramento por defecto, D2MCS incorpora un mecanismo de herdanza que permite definir, desenvolver e cargar dunha manera sinxela novas heurísticas que se adapten ás necesidades do usuario. Por outro lado, co fin de garantir un uso correcto dos recursos de hardware do equipo, D2MCS permite almacenar en disco só o mellor modelo para cada grupo de características (aforro de espazo almacenamento) e manter un rexistro dos modelos que xa se executaron (evita as reexecucións innecesarias).

Finalmente, na última etapa, os modelos co mellor rendemento úsanse para calcular o resultado final. Para iso, as predicciónss individuais obtidas tras aplicar os modelos sobre os grupos de características identificados na primeira fase (distribución máis adecuada) comínanse nun único resultado. En concreto, D2MCS incorpora dous mecanismos de votación que permiten combinar as saídas dos clasificadores internos e proporciona unha API que permite aos usuarios definir fácilmente outros novos. Os métodos implementados son: (i) un sistema de votación simple no que o

resultado final se obtén da suma das predicións individuais de cada modelo, e (ii) un sistema de votación combinado no que o resultado final se obtén a partir das solucións intermedias obtidas tras aplicar múltiples esquemas de votación simple. Cabe destacar que D2MCS proporciona dous esquemas de votación simple (i) por maioría simple no que todos os modelos teñen o mesmo peso e a clase final se obtén a partir da suma das prediccións individuais de cada modelo e (ii) por maioría ponderada, no que cada modelo adestado ten asociado un peso específico e o resultado final veñen determinado pola clase que alcanza o valor total máis alto. Por outro lado, o esquema de votación combinado despreg a un meta-modelo capaz de calcular o resultado final a partires da execución de sistemas de votación simples con diferentes tipos de configuracións.

3. CONCLUSIÓNS E TRABALLO FUTURO

Neste traballo, introdúcese D2MCS (Ferrreiro-Díaz *et al.* 2021), un paquete de R orientado a desenvolver e despregar automáticamente un SMC preciso baseado na distribución de agrupamento de características lograda a partir dun conxunto de datos de entrada. D2MCS céntrase en catro aspectos principais: (i) a capacidade de determinar un método eficaz para evaluar a independencia das características, (ii) a identificación do número óptimo de clusters de características, (iii) o adestramento e axuste dos modelos Machine Learning e (iv) a execución de esquemas de votación para combinar as saidas de cada clasificador que compoñen o SMC.

O traballo futuro céntrase en tres aspectos principais: (i) o redeseño das estructuras que manexan a variable obxectivo para habilitar a posibilidade de dar soporte a problemas multiobxectivo, (ii) a ampliación da variedade técnicas de clustering supervisado e non supervisado ofrecidas polo paquete e (iii) mellorar e ofrecer mecanismos de visualización tanto do estado actual da execución como da información obtida na saida das etapas de agrupamento de características e de clasificación.

AGRADECEMENTOS

O grupo SING agradece ao CITI (Centro de Investigación, Transferencia e Innovación) da Universidade de Vigo o aloxamento da súa infraestructura informática.

Referencias

[1] H. Ekbia et al., "Big data, bigger dilemmas: A critical review," J. Assoc. Inf. Sci. Technol., vol. 66, no. 8, pp. 1523–1545, Aug. 2015.

[2] A. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.

[3] M. Chen, S. Mao, Y. Liu, "Big Data: A Survey," Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014.

[4] E. Burmeister, L. M. Aitken, "Sample size: How many is enough?," Aust. Crit. Care, vol. 25, no. 4, pp. 271–274, Nov. 2012.

[5] M. Woźniak, M. Graña, E. Corchado, "A survey of multiple classifier systems as hybrid systems," Inf. Fusion, vol. 16, pp. 3–17, Mar. 2014.

[6] F. Roli, "Multiple Classifier Systems," in Encyclopedia of Biometrics, Springer US, Boston, MA, pp. 981–986. 2009.

[7] G. Giacinto, F. Roli, "An approach to the automatic design of multiple classifier systems," Pattern Recognit. Lett., vol. 22, no. 1, pp. 25–33, Jan. 2001.

[8] D. Ruano-Ordás, I. Yevseyeva, V. B. Fernandes, J. R. Méndez, M. T. M. Emmerich, "Improving the drug discovery process by using multiple classifier systems,.Expert Syst. Appl., vol. 121, pp. 292–303, May 2019.

[9] D. Ruano-Ordás et al., "A multiple classifier system identifies novel cannabinoid CB2 receptor ligands," J. Cheminform., vol. 11, no. 1, p. 66, Dec. 2019.

[10] Ferreiro-Díaz, Ruano-Ordas, D., Méndez, J. R. (2021). D2MCS: Data Driving Multiple Classifier System. R package version 1.0.0. `https://CRAN.R-project.org/package=D2MCS`.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

R-INLA: a flexible tool for implementing Bayesian regression models

A. Fuster-Alonso[1], S. Cerviño[2], D. Conesa[1], M. Cousido-Rocha[2], F. Izquierdo[2], M.G. Pennino[2]

[1]Universitat de Valéncia
[2]Instituto Español de Oceanografía (IEO), Vigo

ABSTRACT

Bayesian methodology has been widely used for fitting regression models due to some advantages: the large amount of information provided by estimations, how simple are to interpret those estimations, and the ease with which prior information can be incorporated. But lately, the most relevant advantage may be the fact that it can handle complex models such as those including spatial and spatio-temporal effects.

Nevertheless, the posterior distributions of this kind of complex models do not yield to analytical expressions and so computational approaches are needed. Among them, in this work we review the use of the Integrated Nested Laplace Approximation (INLA) methodology (Rue et al. 2009) and software (`http://www.r-inla.org`) as an alternative to the popular Markov chain Monte Carlo (MCMC) methods, the main reason being the speed of calculation. In particular, we here provide an illustration of the application and performance of different type of models (generalized additive models, spatial models and hurdle models) in `R-INLA` using simulated data from oceanographic trawl surveys.

Key words: Data simulation; Delta models; Generalized additive models; Geostatistical models; Random sampling; `R-INLA`.

1. INTRODUCTION

In general, when performing statistical inference there are two approaches: the frequentist and the Bayesian one. The main difference between them is how they interpret probability. When estimating a parameter, e.g. β, the frecuentist and Bayesian methods deal with the probability as follows:

1. The frequentist approach focuses on the probability of data given the parameter β, P(data$|\beta$). Therefore, we get a fixed estimate of β with a 95 % confidence interval.

2. The Bayesian approach focuses on the probability of β given the data P($\beta|$data). As a consequence, the distribution of the parameter is obtained using the data. The term P($\beta|$data) is called the posterior distribution of β and Bayes' theorem states that it is proportional to the likelihood of the data P(data$|\beta$) given the parameters and the prior knowledge P(β) (equation 1):

$$\mathrm{P}(\beta|\text{data}) \propto \mathrm{P}(\text{data}|\beta) \times \mathrm{P}(\beta)\,. \tag{1}$$

Traditionally, the frequentist approach has been the most popular, but recently the Bayesian approach has gained more interest. Some of the reasons underneath are:

- The posterior distributions P($\beta|$data) provide all the information about the parameters, e.g., mean, median, quantiles, nonzero probabilities, etc. and are so easy to interpret.

- The ease with which prior information could be incorporated P(β). When this knowledge is poor, vague priors are assumed so that the posterior distribution is based on the observed data
- An extensive number of complex models could be fitted efficiently and faster comparing to another classic techniques.

In brief, Bayesian inference on complex models results in analytical expressions to obtain a posterior distributions, so computational approaches and numerical approximations are needed. Traditionally, some computational approaches have been proposed to estimate the posterior distributions with Markov chain Monte Carlo (MCMC) methods, implemented in softwares such as WinBUGS (Lunn et al., 2000). However, MCMC provides simulations from the ensemble of model parameters so it requires a lot of simulation for valid inference and maybe we are interested in a single parameter or subset of the parameters. For this reason, in this work we review the use of the Integrated Nested Laplace Approximation (INLA) methodology (Rue et al. 2009) and software (`http://www.r-inla.org`) as an alternative to the popular Markov chain Monte Carlo (MCMC) methods. Likewise, `R-INLA` is computationally faster and accurate, due to numerical integration techniques are used instead of Monte Carlo sampling. INLA framework focuses on estimating the marginal posterior distribution of the models effect and hyperparameters. The single assumption needed is that the statistical model is a Latent Gaussian Model (LGM) (Gómez-Rubio, 2020). In fact, most statistical models are LGM's, for example, spatial models, spline models, generalized linear/nonlinear models, survival models, joint models etc. Among all the models mentioned, the following work will focus mainly on the fitting of spatial models and hurdle models with simulated georeferenced data.

2. SIMULATION

In order to provide examples of how to use `R-INLA` to fit different regression models, an oceanographic trawl survey has been simulated. Firstly, an oceanographic survey is based on randomly georeferenced sampling a space to obtain relative biomass indices of a target species. Consequently, it is necessary to simulate a pattern of points, which represents the coordinates where the target species has been fished. This simulation is performed through a LGCP (Log Gaussian Cox Process). Also, it is necessary to simulated few different variables:

1. **CPUE**: it refers to the response variable and represent the catches per unit effort (2). It is a quantitative, continuous and positive variable, whose domain is $(0, +\infty)$:

$$\text{CPUE} = \frac{\text{Catches}}{\text{Effort}} \tag{2}$$

2. **Bathymetry**: it refers to the depth measured in meters. It is a quantitative, continuous and positive covariate. In addition, a nonlinear relationship between the response variable and the bathymetry has been established.
3. **Effort**: it refers to the minutes in which the gear remains active on the seabed. It is a quantitative, continuous and positive covariate.
4. **Coordinates**: coordinates have been simulated on the x and y axis over a 1x1 window, resulting in two variables x and y that represent the location of fishing points. It could be added to the models as a spatial random effect.

3. MODELS IN R-INLA

In this section, the review of three types of models implemented in R-INLA is discussed: **GAM, spatial model and hurdle model**.

GAM: generalized additive models

The GAM proposed follows the equation 3, where the response variable is CPUE:

$$\begin{aligned} \text{CPUE}_i \sim \text{lognormal}(\mu_i, \sigma) \quad i = 1, .., n\,, \\ \mu_i = \beta_0 + f(B_i)\,, \end{aligned} \tag{3}$$

where, μ_i represents the response variable mean, the link function is the identity and $f(\cdot)$ is a rw1 (random walk model of order 1) applied to the covariate bathymetry (B_i). R-INLA has implemented two default smoothing models rw1 (random walk model of order 1) and rw2 (random walk model of order 2).
The R code is almost identical to that of fitting a linear regression model with glm() or lm() function. The synthesis to fit the model in R-INLA would be using the function inla(formula, data = data), the argument formula includes the response variable and the linear predictor proposed in the model.

Spatial models

The equation is nearly identical to the equations without a spatial term, except that it contains an extra term $u(\text{s}_i)$ at the end. It refers to the spatial correlated random term, which is estimated with the Stochastic Partial Differential Equations (SPDE) approach by R-INLA (Krainski et al., 2018):

$$\begin{aligned} \text{CPUE}(\text{s}_i) \sim \text{lognormal}(\mu(\text{s}_i), \sigma) \quad \text{i} = 1, .., \text{n}\,, \\ \mu(\text{s}_i) = \beta_0 + f(B(\text{s}_i)) + u(\text{s}_i)\,, \\ u(\text{s}_i) \sim \text{GMRF}(0, \Sigma)\,, \end{aligned} \tag{4}$$

where, f is a rw1 model, B_i refers to the covariate bathymetry, $u = (u(\text{s}_1), ..., u(\text{s}_\text{N}))$ is a random effect distributed as a Gaussian Markovian Random Field (GMRF) and Σ is the Matérn covariance of dimension N x N. In order to quantify the covariance matrix it is necessary to estimate the Matérn correlation function, which is a mathematical source to define correlation as a function of distance. Attempt to estimate the parameters related to the Matérn correlation we need to use the SPDE and the finite element approach. In brief, the spatial effect $u(\text{s}_i)$ could be approach solving the equation 5:

$$u(\text{s}_i) = \sum_{k=1}^{G} a_k(\text{s}_i)\,\text{x}\,w_k, \tag{5}$$

where, $a_k(\text{s}_i)$ is a know value that indicate the position at the mesh and w_k is the spatial field.
Zuur et al. (2017) explain how to implement a spatial model in R-INLA:

1. Elaborate a Mesh: the study area is divided into a large number of triangles, it is equivalent to the elaboration of an irregular grid.

2. Define the projector matrix a_{ik}.

3. Define the SPDE: it is use to simplify the GMRF, because we assume that only neighbouring sites have non-zero covariance values.

4. Define the spatial field $w_k(s)$.

5. Elaborate a Stack: the term stack refers to the union of the response variable and its linear predictor with the space in which the study is located (coordinates).

6. Define the model formula (the response variable and the linear predictor).

7. Finally run the model in R-INLA and explore the results obtain.

Hurdle models

Hurdle model for continuous data consists of modeling two process: (1) a binary part to fit the presence/absence of the target specie and (2) a continuous part to model the intensity when the response is non-zero (Izquierdo et al., 2021). The distribution for the binary part will be a bernoulli distribution and for the continuous part we use a gamma distribution (equation 6).

$$\begin{aligned}
\text{YCPUE}(\text{s}_i) &\sim \text{Bernoulli}(\pi(\text{s}_i)) \quad \text{i} = 1, .., \text{n}\,, \\
\text{ZCPUE}(\text{s}_i) &\sim \text{Gamma}(\mu(\text{s}_i), \sigma) \quad \text{i} = 1, .., \text{n}\,, \\
\text{logit}(\pi(\text{s}_i)) &= \beta_{0\text{YCPUE}} + f(B(\text{s}_i)) + u_{\text{YCPUE}}(\text{s}_i)\,, \\
u_{\text{YCPUE}}(s_i) &\sim \text{GMRF}(0, \Sigma_{\text{YCPUE}})\,, \\
\log(\mu(\text{s}_i)) &= \beta_{0\text{ZCPUE}} + f(B(\text{s}_i)) + u_{\text{ZCPUE}}(\text{s}_i)\,, \\
u_{\text{ZCPUE}}(s_i) &\sim \text{GMRF}(0, \Sigma_{\text{ZCPUE}})\,,
\end{aligned} \tag{6}$$

where $\text{YCPUE}(\text{s}_i)$ is a first process of presence/absence and the parameter of interest is $\pi(\text{s}_i)$, which represents the probability of presence or absence of the species, it is linked to the linear predictor by link logit. The second process model all non-zeros $\text{ZCPUE}(\text{s}_i)$, the parameter of interest is $\mu(s_i)$ and the link for the linear predictor is the logarithm. The effect of bathymetry is a rw1 $f(B_i)$ and it is the same for each processes. The spatial terms are different for both processes $u_{\text{YCPUE}}(\text{s}_i)$ and $u_{\text{ZCPUE}}(\text{s}_i)$ such as the Matérn correlation Σ_{YCPUE} and Σ_{ZCPUE}.
The steps to follow to implement a hurdle model in `R-INLA` are very similar to the spatial model. The most remarkable difference is the Stack, it will be composed of two parts: (I) one stack is defined for the $\text{YCPUE}(\text{s}_i)$ process and (II) a second stack is defined for the continuous process $\text{ZCPUE}(\text{s}_i)$.

4. CONCLUSIONS

In conclusion, `R-INLA` is a fast, robust and accurate tool for approximate Bayesian inference, making it a good alternative to MCMC methods. Therefore, considering the advantages of `R-INLA`: the low computational cost, since it is not based on sampling, but on numerical integration and the great variety of statistical models available, this work review the use of `R-INLA` through several examples (GAM, spatial model and hurdle model).

5. FUTHER INFORMATION

Github repository: AlbaFuster: Spatial simulation and models in R-INLA

6. ACKNOWLEDGMENTS

This study is a contribution to the project IMPRESS (RTI2018-099868-B-I00), ERDF, Ministry of Science, Innovation and Universities - State Research Agency and also of GAIN (Xunta de Galicia), GRC MERVEX (nº IN607-A 2018-4). AF thanks the University of Valencia for awarding the grant "Ayudas para el inicio a la investigación" (UV-SEDI_II-1577224).

References

[1] Gómez-Rubio, Virgilio. 2020. Bayesian Inference with INLA. Boca Raton, FL: Chapman & Hall/CRC Press. `https://becarioprecario.bitbucket.io/inla-gitbook`.

[2] Izquierdo, F., Paradinas, I., Cerviño, S., Conesa, D., Alonso-Fernández, A., Velasco, F., Preciad I., Punzón A., Saborido-Rey F. and Pennino, M. G. (2021). *Spatio-temporal assessment of the European hake (Merluccius merluccius) recruits in the northern Iberian Peninsula* Frontiers in Marine Science, 8, 1.

[3] Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D. and Rue, H. (2018). Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC.

[4] Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). *WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility.* Statistics and computing, 10(4), 325-337.

[5] Rue, Havard, Sara Martino, and Nicolas Chopin. 2009. *Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations*. Journal of the Royal Statistical Society, Series B 71 (2): 319–92.

[6] Zuur, A. F., Ieno, E. N. and Saveliev, A. A. (2017). *Spatial, temporal and spatial-temporal ecological data analysis with r-inla*. Highland Statistics Ltd, 1.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

ANÁLISIS Y VISUALIZACIÓN DEL EXCESO DEL EXCESO DE DEFUNCIONES EN 2020 CON R

Virgilio Gómez Rubio[1]

[1] Departamento de Matemáticas, ETS Ingenieros Industriales- Albacete, Universidad de Castilla-La Mancha

RESUMO

La pandemia de COVID-19 produjo un exceso de defunciones en muchos países en 2020. En mi charla ilustraré el uso que hemos hecho de R para el análisis de datos de defunciones totales y población, combinnados con datos ambientales, para el estudio del exceso de defunciones en 2020. El estudio analiza las defunciones totales por grupos de edad y sexo en 5 países europeos, y proporciona estimaciones del exceso de mortalidad a distintitos niveles de agregación. Tanto para el análisis de datos como para la visualización de los resultados hemos utilizado el sfotware estadístico R. Además, para la exploración de los resultados hemos desarrollado una aplicación en Shiny.

Palabras e frases chave: INLA, R, Shiny,

Referencias

[1] Konstantinoudis, G., M. Cameletti, V. Gómez-Rubio, I. León-Gómez, M. Pirani, G. Baio, A. Larrauri, J. Riou, M. Egger, P. Vineis, M. Blangiardo (2021). Regional excess mortality during the 2020 COVID-19 pandemic: a study of five European countries. En revisión.

VIII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 14 de outubro do 2021

Análisis de la concentración de CO_2 en espacios interiores para la prevención del contagio de la COVID-19

María Jesús Hernández[1], Víctor Teodoro[1], Carlos Escudero[1], Manuel Oviedo[1], Óscar Fontenla[1]

[1] Centro de Investigación en Tecnologías de la Información y la Comunicación (CITIC), Universidade da Coruña.

RESUMEN

La concentración de CO_2 en espacios cerrados es un buen indicador de la tasa de ventilación del lugar, un factor fundamental para la prevención de la transmisión de COVID-19 mediante aerosoles. Por esta razón, hemos analizado los niveles de CO_2 en las clases de la Universidade da Coruña durante los tres días en los que se desarrollaron los exámenes de la ABAU 2021. En este estudio, se han utilizado distintas herramientas en R que permiten el análisis y la visualización de datos funcionales, pues la concentración de CO_2 es una variable continua medida a lo largo del tiempo. Además, se ha desarrollado una función en R que extrae de la API del MeteoSIX datos de predicción meteorológica de MeteoGalicia, ya que las condiciones ambientales pueden afectar a los niveles de CO_2.

Palabras y frases clave: concentración de CO_2, COVID-19, datos funcionales, API, MeteoGalicia.

1. VISUALIZACIÓN DE DATOS

Para medir la concentración de CO_2 (ppm), se instaló un sensor en cada una de las aulas de las distintas facultades y campus de la Universidade da Coruña donde se realizaron los exámenes de la ABAU. Veamos en la siguiente figura la evolución de la concentración de CO_2 a lo largo de los tres días. Cada curva representa una clase.

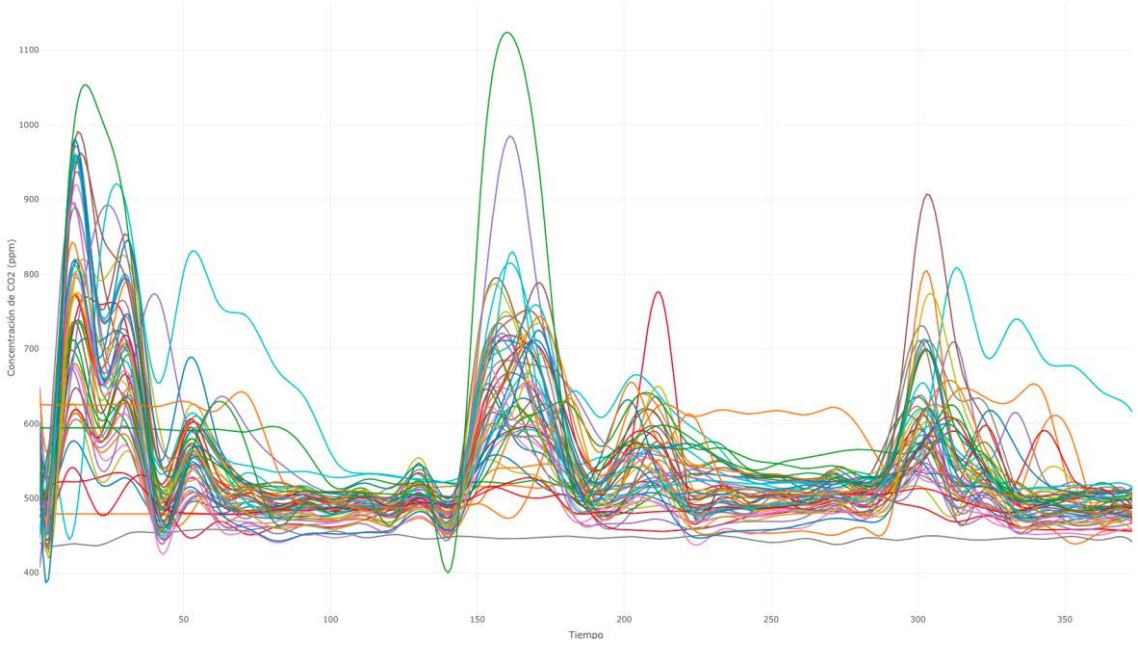


Figura 1. Evolución de la concentración de CO_2 durante la ABAU.

En la siguiente figura vemos la profundidad de Fraiman-Muniz para nuestros datos funcionales, que actúa como una medida de distribucion central. Las curvas con mayor profundidad están en color amarillo, mientras que aquellas con profundidad más baja están en colores oscuros. La función de color azul es la más profunda y puede interpretarse como la función mediana, mientras que la roja representa la recortada al 25%.

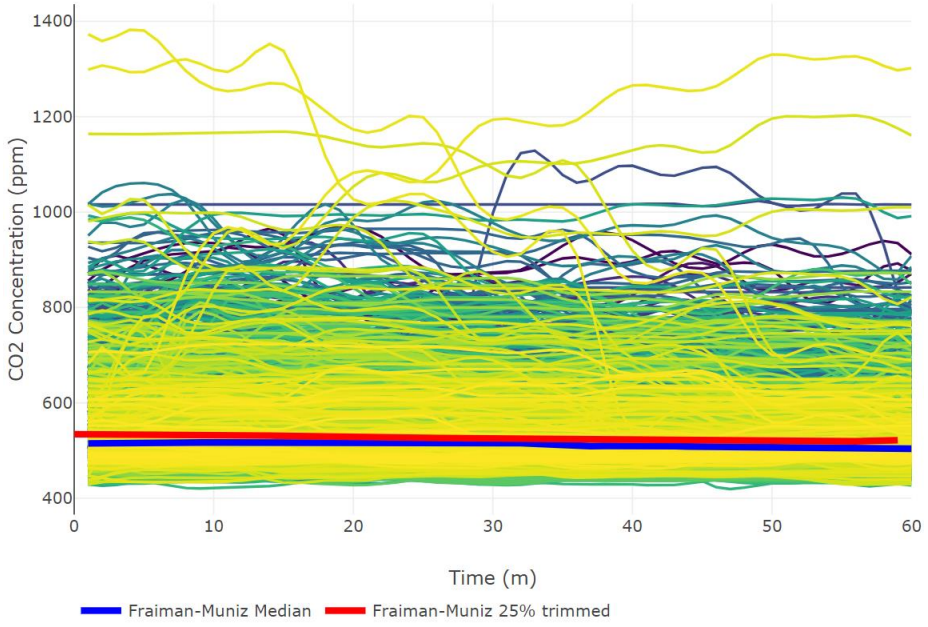


Figura 2. Profundidad de Fraiman-Muniz.

Para el análisis de los datos funcionales se utilizó el paquete *fda.usc* (Febrero & Oviedo, 2012), y para la visualización se usó el paquete *plotly* (Sievert, 2020).

2. CONSULTA DE DATOS METEOROLÓGICOS DE METEOGALICIA MEDIANTE API

La API del MeteoSIX es un servicio web que da acceso a los resultados de los distintos modelos de predicción numérica ejecutados diariamente por MeteoGalicia. Haciendo uso de la versión v4 de esta API, hemos desarrollado una función que permite al usuario consultar el estado del cielo, la temperatura, el viento, las precipitaciones, la humedad relativa, la cobertura nubosa y la presión al nivel del mar, según el modelo WRF (Weather Research Forecast).

Se puede realizar la consulta para cualquier entidad de población o playa de Galicia. Además, es posible seleccionar un rango temporal para los datos, que puede ir desde el día de hoy a una hora determinada hasta un máximo de 7 días, debido a las características de los modelos de predicción.

Veamos un ejemplo de la función, a la que hemos llamado "meteogalicia":

meteogalicia(variables=c("temperature", "relative_humidity"), ciudad="Vigo", fecha_ini= "2021-09-17", hora_ini="12", fecha_fin="2021-09-17", hora_fin="15")

Esta función devolvería un DataFrame con la siguiente información:

ciudad	datetime	temperature	relative_humidity
Vigo	2021-09-17 12:00:00	19	84.12
Vigo	2021-09-17 13:00:00	21	82.20
Vigo	2021-09-17 14:00:00	23	78.22
Vigo	2021-09-17 15:00:00	19	78.22

Tabla 1. Datos meteorológicos obtenidos con la función "meteogalicia".

Para la elaboración de esta función se emplearon los paquetes *httr* (Wickham, 2020) y *jsonlite* (Ooms, 2014).

AGRADECIMIENTOS

Este estudio ha sido apoyado por GAIN (Axencia Galega de Innovación) y la Consellería de Economía, Emprego e Industria de la Xunta de Galicia; subvención COV20/00604 por medio de FEDER.

Referencias

[1] Febrero-Bande, M. & Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51(4), 1-28. URL http://www.jstatsoft.org/v51/i04/.

[2] Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

[3] Sievert, C. (2020) Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC, Florida.

[4] Wickham, H. (2020). httr: Tools for Working with URLs and HTTP. R package version 1.4.2. https://CRAN.R-project.org/package=httr.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

ALGÚNS USOS DO R NOS PROCESOS DO INSTITUTO GALEGO DE ESTATÍSTICA

Mª Esther López Vizcaíno[1]

[1] Instituto Galego de Estatística

RESUMO

R é un paquete estatístico de elevada e crecente importancia para a implementación de técnicas estatísticas en diversas disciplinas científicas aplicadas. O seu carácter gratuíto, a multitude de recursos dispoñibles para o programa e a súa elevada calidade, tanto analítica como gráfica, fan que gradualmente se vaia convertendo nunha especie de lingua franca para a análise estatística.
Ademais das xa mencionadas, as razóns polas que a comunidade da estatística oficial está a adoptar rapidamente R son claras: ten unha comunidade activa de usuarios e usuarias en todo o mundo, hai un amplo apoio da industria e combina unha gran cantidade de funcionalidades para a preparación de datos, metodoloxía, visualización e creación de aplicacións.
Neste relatorio faise unha revisión do uso do software R nos distintos procesos estatísticos levados a cabo no Instituto Galego de Estatística (IGE). Estes usos van desde a captura de datos de fontes externas ou internas para gravar nas nosas bases de datos, pasando polo uso deste software para procesar información estatística, ata a publicación de aplicacións web dinámicas.

No IGE utilízase de forma moi activa o software R para a captura de datos procedentes de fontes externas de información, como pode ser a Dirección Xeral de Tráfico, o SEPE, a Tesourería Xeral da Seguridade Social, Ministerio de Fomento, etc. Esta información, unha vez capturada, almacénase nas bases de datos dispoñibles no IGE en MySql ou SQLServer, utilizando para iso, entre outros, os paquetes RMySQL, RODBC, DBI ou openxlsx.

Por outra banda, tamén se utiliza R en diferentes procesos de manipulación de datos, como pode ser o emparellamento estatístico (Statistical Matching) e a vinculación de rexistros. Nesta tarefa utilízanse paquetes como o RecordLinkage, stringdist ou o MatchIt. E, por último, outro uso intenso que se fai do software R é para a elaboración de informes automatizados de carácter conxuntural ou estrutural. Para isto utilízanse os paquetes rmarkdown e knitr, entre outros. Ademais, tamén se usa para a elaboración de aplicacións web dinámicas co paquete shiny

Palabras e frases chave: estatística oficial, R, R-shiny, R-Markdown.

Referencias

[1] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] 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] JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2021). rmarkdown: Dynamic Documents for R. R package version 2.7. URL https://rmarkdown.rstudio.com.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

R NO BACHARELATO. ESTAMOS TOLOS OU QUE?

Beatriz Padín Romero[1]

[1] Colexio M. Peleteiro (Santiago de Compostela)

RESUMO

O software estatístico R ofrece amplísimas posibilidades nas prácticas de laboratorio de Fisica en Bacharelato. Con el os alumnos poden facer unha análise estatística dos datos experimentais, representalos gráficamente, facer o seu tratamento e escribir pequenos programas que fagan os cálculos pertinentes para obter diversas magnitudes. Ademais da súa utilidade como ferramenta, co seu uso promóvese a competencia dixital nos alumnos, especialmente no importante eido da programación científica.

Palabras e frases chave: R, Física, Bacharelato, laboratorio, programación científica.

1. INTRODUCIÓN

O traballo experimental é un elemento fundamental no estudo da Física en Bacharelato. No currículum desta asignatura establécese como un dos estándares de aprendizaxe "a elaboración e a interpretación de representacións gráficas a partir de datos experimentais, relacionándoas coas ecuacións matemáticas que representan as leis e os principios físicos subxacentes". Neste mesmo documento recollése así mesmo a importancia que ten a competencia dixital na a formación dos alumnos, que "merece un tratamento específico no estudo desta materia".

É moi habitual unir estes dous aspectos do currículum –o traballo experimental e a formación nas competencias dixitais– no desenvolvemento das prácticas de laboratorio de Física, xa que normalmente a análise dos datos se fai utilizando ferramentas dixitais como as follas de cálculo (Excel, Google Sheets, Calc...). Sen embargo, cremos que se pode ir un paso máis alá e propoñemos o uso do software estatístico R para o tratamento dos datos experimentais nas prácticas de Física en Bacharelato. Entre as vantaxes que ofrece R está o feito de que é extremadamente completo e versátil o que, unido a que é tamén unha linguaxe de programación, permite un enfoque moito máis global do tratamento dos datos. Por outro lado R é unha ferramenta de software libre e gratuito, e consideramos que é moi importante promover o uso do software libre como parte da alfabetización dixital dos nosos alumnos. Aínda que o seu uso non é tan inmediato e intuitivo como unha folla de cálculo, a dificultade inicial vese compensada con creces polo feito de que coa súa aprendizaxe estamos capacitando aos alumnos para ser competentes no uso de ferramentas computacionais que lles serán útiles nos seus futuros estudos universitarios.

2. EXEMPLOS DO USO DE R NO LABORATORIO DE BACHARELATO

Por que usar R cando existen outras ferramentas que fan un traballo similar? Neste apartado mostramos unha serie de exemplos que ilustran as vantaxes que ofrece o uso de R sobre a maneira "tradicional" (con papel, bolígrafo e calculadora, ou usando unha folla de cálculo) de analizar os datos obtidos no laboratorio. Todos os problemas recollidos se corresponden co currículo oficial das asignaturas de Física e Química de primeiro de Bacharelato ou de Física de segundo de Bacharelato.

Estatística descritiva

Unha maneira de determinar a aceleración da gravidade é facendo un estudo do movemento dun péndulo. Nesta práctica o primeiro paso é medir o período do péndulo, para o cal se usa un cronómetro co que se repite varias veces, nas mesmas condicións, a medida do período. Unha vez recollidos estes datos o habitual é que os alumnos se limiten a obter a súa media, xa que esta será a mellor estimación do valor do período do péndulo. Pero con R é extremadamente doado coñecer máis cousas sobre os datos: ¿como se distribúen?, ¿están moi dispersos arredor da media?, ¿hai valores atípicos? A análise destas preguntas, aínda que normalmente non se leva a cabo, é de grande axuda para entender conceptos tan importantes como precisión e exactitude da medida, normalidade dunha distribución ou incerteza asociada a unha medida.

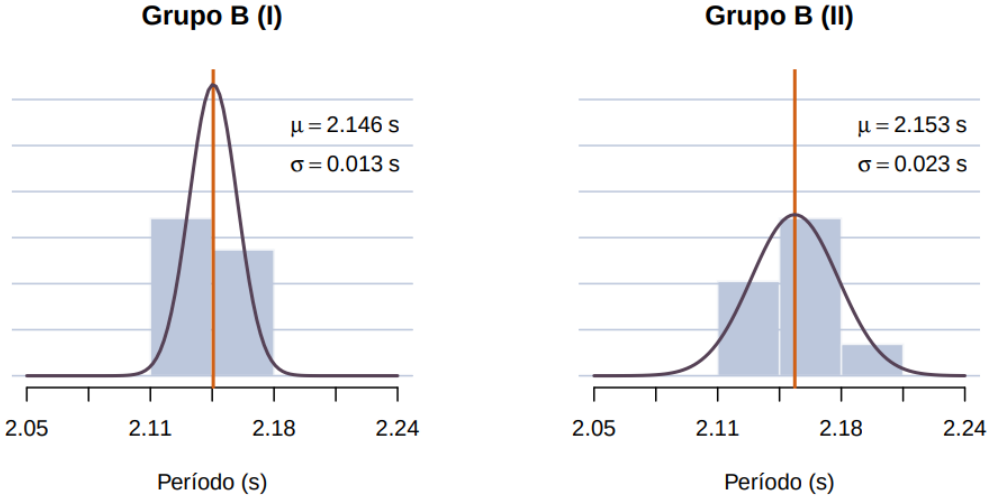


Figura 1. Comparación de dúas mostras diferentes da medida do período dun péndulo.

Axuste a unha recta

É moi habitual facer un axuste dos datos experimentais a unha recta para poder extraer do axuste os valores de certos parámetros físicos (por exemplo, para determinar a constante elástica dun resorte, a potencia dunha lente ou a masa da Terra). Este problema pode afrontarse representando a man os datos en papel milimetrado e facendo un axuste "a ollo". Aínda que este é un método que os alumnos deben coñecer (e, de feito, é o método que se lles esixe nas probas da ABAU), a súa eficacia deixa moito que desexar. Os programas de folla de cálculo proporcionan unha maneira máis exacta de obter este axuste mediante o método dos mínimos cadrados. Como non podía ser menos, R tamen ofrece esta opción, coa vantaxe engadida de que ao lado dos valores da pendente e da ordenada na orixe proporciona tamén as súas incertezas, información que é imprescindible á hora de expresar correctamente o valor da medida.

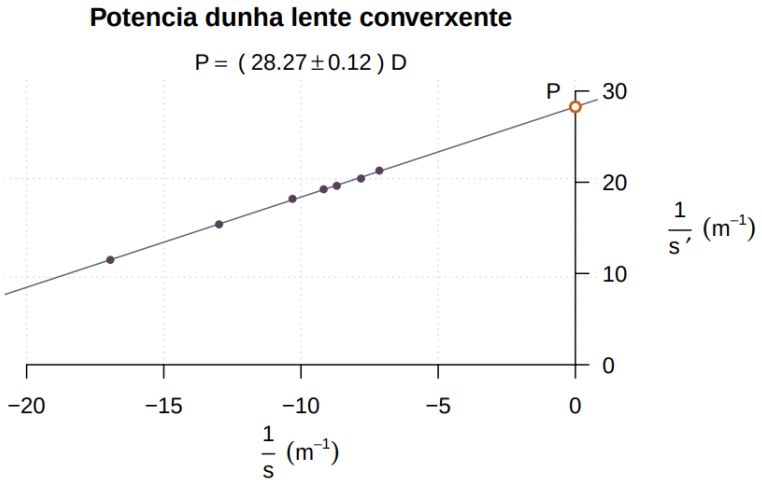


Figura 2. Obtención da potencia dunha lente, xunto coa súa incerteza.

Representación gráfica

No eido das representacións gráficas R ofrece non só unha gran variedade de tipos de gráficas, senón que ademais permite a súa personalización e adaptación ata o mínimo detalle. Dada a súa versatilidade, con R pódense mostrar graficamente relacións entre diferentes magnitudes dun xeito que outros programas non permiten. Por exemplo, na práctica "Satélites terrestres e as súas órbitas" os alumnos deben utilizar datos reais de diferentes satélites para analizar o seu movemento. Representando gráficamente estes valores, é doado sacar conclusións que son difíciles de alcanzar unicamente cos valores numéricos, como son a relación entre a altura da órbita e o período orbital do satélite, a excentricidade da órbita ou o tipo de satélite.

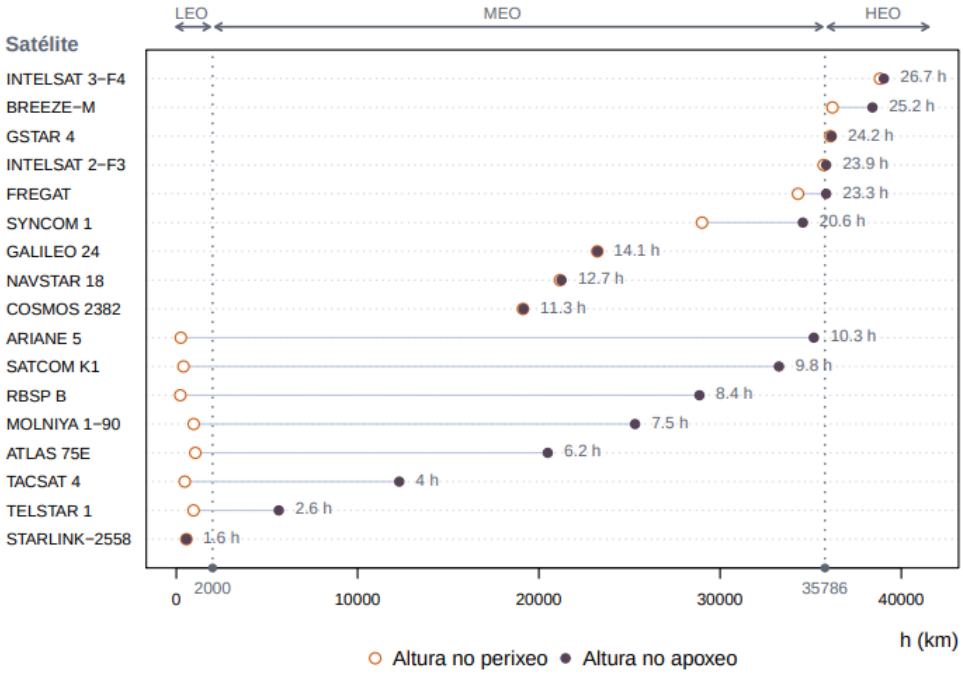


Figura 3. Altura no apoxeo e no perixeo, período orbital e tipo de satélite según a altura da órbita (LEO: órbita baixa, MEO: órbita media, HEO: órbita alta).

Linguaxe de programación

O que fai de R unha ferramenta transformadora no laboratorio é o feito de que, ademais de todo o sinalado, é tamén unha linguaxe de programación. Isto posibilita que, nunha mesma contorna, os alumnos dispoñan de recursos para estudar

estatísticamente os datos, representalos graficamente, facer unha análise deles e utilizar os resultados para facer calquera tipo de cálculo que precisen. No estudo do efecto fotoeléctrico, por exemplo, despois de facer o axuste dos datos experimentais a unha recta pódense calcular, a partir dos coeficientes obtidos, outros moitos valores sen máis que escribir un sinxelo programa.

```
[1] "------ PREGUNTA 1 ------"
[1] "Constante de Planck experimental: h = 6.61e-34 J·s"
[1] "Incerteza: u(h) = 9.4e-36 J·s"
[1] "Discrepancia entre o valor real e o valor medido: 0.24 %"
[1] "------ PREGUNTA 2 ------"
[1] "Traballo de extracción: W0 = 3.593e-19 J = 2.243 eV"
[1] "Incerteza: u(W0) = 9.7e-21 J = 0.06 eV"
[1] "------ PREGUNTA 3 ------"
[1] "Frecuencia umbral: f0 = 5.436e+14 Hz"
[1] "Lonxitude de onda umbral: l0 = 5.514e-07 m = 551.4 nm"
[1] "------ PREGUNTA 4 ------"
[1] "Lonxitude de onda: l = 215 nm"
[1] "Velocidade dos fotoelectróns: v = 1111148.03 m/s"
```

Figura 4. Efecto fotoeléctrico. Os alumnos escriben un programa co que, a partir do axuste dos datos experimentais a unha recta, calculan magnitudes como a constante de Planck, a frecuencia umbral de metal ou a velocidade dos fotoelectróns.

3. CONCLUSIÓNS

Os beneficios que ofrece o uso de R no laboratorio de Bacharelato son innegables… pero a realidade é ben diferente. Os temarios oficiais son de por si moi extensos como para introducir novos contidos, e menos aínda unha linguaxe de programación. A grande maioría dos profesores non teñen formación neste campo. Como ferramenta, R non é intuitiva. Daquela, R si ou R non?

No currículo de Física de segundo de Bacharelato indícase que esta disciplina "debe dotar o/a alumno/a de novas aptitudes que o capaciten para a súa seguinte etapa de formación, con independencia da relación que esta poida ter coa física". E que mellor aptitude que ter coñecementos de programación científica, que é unha habilidade que se lles vai esixir nos seus futuros estudos universitarios. Non só iso; a American Association of Physics Teachers vai máis alá e no documento *Recommendations for the Undergraduate Physics Laboratory Curriculum* indica o seguinte: "Contemporary research in physics and related sciences almost always involves the use of computers. They are used for data collection and analysis, numerical analysis, simulations, and symbolic manipulation. Computational physics has become a third way of doing physics and complements traditional modes of theoretical and experimental physics". E, a pesar disto, a programación científica non está recollida no currículo das asignaturas de ciencias en Bacharelato.

En conclusión: si, estamos tolos se non formamos aos nosos alumnos no uso de ferramentas computacionais como R.

Referencias

[1] Guía da LOMCE. Bacharelato. http://www.edu.xunta.gal/portal/guiadalomce/bacharelato

[2] ABAU Física. https://ciug.gal/PDF/Grupos_Traballo_2021/23_fisica_practicas_2021.pdf

[3] AAPT Recommendations for Computational Physics in the Undergraduate Physics Curriculum. https://www.aapt.org/resources/upload/aapt_uctf_compphysreport_final_b.pdf

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

Mesa redonda: R y COVID

Miguel Ángel Rodríguez Muíños[1] (moderador), Mª Jesús Hernández Vega[2] , Javier Kniffki[3] , Manuel Antonio Novo Pérez[2] , Manuel Vaamonde Rivas[2] , Javier Álvarez Liébana[4]

[1]Consellería de Sanidade. [2]Universidade da Coruña, [3]Kstats, [4]Universidad Complutense de Madrid

RESUMO

En los tiempos que corren, uno de los temas candentes a nivel mundial es la Pandemia de COVID. Creo que no necesita ningún tipo de introducción o presentación. Además de los aspectos clínicos de la misma uno de los papeles más importantes lo han jugado (lo continúan haciendo) los especialistas en análisis de datos. R está jugando un papel fundamental en este campo. Dentro de esta 8ª Xornada de Usuarios de R en Galicia contamos con la presencia de relevantes figuras del análisis y divulgación de datos, sobre COVID, con R.

El orden de intervención de los invitados se ha establecido alfabéticamente por el primer apellido, moviendo a Javier Álvarez al último lugar por problemas de agenda. Cada uno de ellos realizará una breve charla (5 a 7 minutos) y, una vez finalizadas las intervenciones, procederemos a debatir sobre las mismas o sobre cualquier tema relacionado que tengáis a bien preguntar.

INVITADOS PARTICIPANTES EN LA MESA:

- Miguel Ángel Rodríguez Muíños (moderador): Técnico informático del Servicio de Epidemioloxía de la Dirección Xeral de Saúde Pública (Consellería de Sanidade - Xunta de Galicia).

- Mª Jesús Hernández Vega: Científica de Datos en CaixaBank. Matemática, Máster en Big Data y participante en el proyecto CEDCOVID "Evaluación, predicción poblacional y personalizada de la evolución de la enfermedad COVID-19" del CITIC (UDC). Nos hablará sobre el análisis de la concentración de CO2 en espacios interiores para la prevención del contagio de la COVID-19

- Javier Kniffki: Máster en Técnicas Estadísticas. CEO y fundador de Kstats (empresa de estudios estadísticos y análisis de datos – Galicia y Mexico-). Nos hablará sobre la recopilación y análisis de datos COVID en colegios de Galicia.

- Manuel Antonio Novo Pèrez: Investigador de la UDC. Matemático. Máster en Estadística. Nos hablará del proyecto ForeCoop y la monitorización de la evolución de la COVID en España.

- Manuel Vaamonde Rivas: Científico de datos en la UDC. Matemático. Máster en Técnicas Estadísticas. Nos hablará sobre el análisis y modelado de datos de carga viral de COVID-19 en aguas residuales en el marco del proyecto COVIDBENS.

- Javier Álvarez Liébana: Matemático (doctor), investigador y divulgador científico. Actualmente investigador y docente en la Universidad Complutense de Madrid. Miembro de "Acción matemática contra el coronavirus"y conocido en Twitter como @DadosdeLaplace, en Instagram y Twitch como @javieralvarezliebana y en Twitter, Discord, Telegram y Youtube como "Dados de Laplace". Nos hablará sobre su labor divulgadora y de la relación con R y el COVID.

Palabras e frases chave: Mesa redonda, R, Shiny, visualización y análisis de datos

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

HDiR: AN R PACKAGE FOR NONPARAMETRIC PLUG-IN ESTIMATION OF DIRECTIONAL HIGHEST DENSITY REGIONS

Paula Saavedra-Nieves[1] and Rosa M. Crujeiras[1]

[1]Department of Statistics, Mathematical Analysis and Optimization
Universidade de Santiago de Compostela

ABSTRACT

A deeper understanding of a distribution support, being able to determine regions of a certain (possibly high) probability content is an important task in several research fields. These regions are known as Highest Density Regions (HDRs) and such a task can be accomplished from a set estimation perspective. Set estimation deals with the problem of reconstructing a set (or estimating any of its features such as its boundary or its volume) from a random sample of points. The reconstruction of this particular type of sets has been mainly considered for densities supported on an Euclidean space. There are only very few contributions where this theory has been extended to the directional domain (see [1] for a recent revision on this topic).

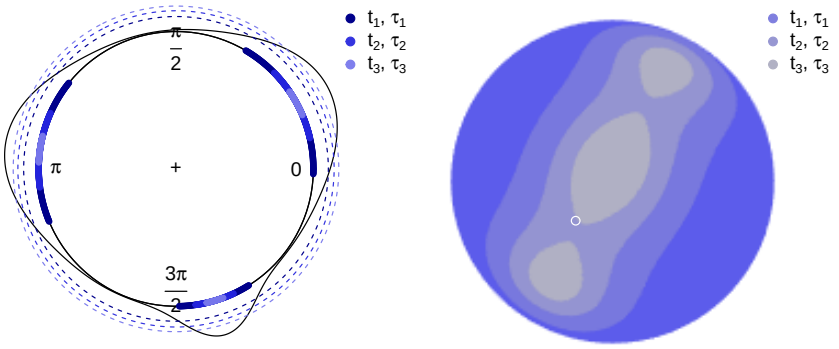


Figura 1: For a circular density (left) and a spherical density (right), HDR $L(f_\tau)$ for $\tau = \tau_1 = 0{,}2$, $\tau = \tau_2 = 0{,}5$ and $\tau = \tau_3 = 0{,}8$.

Formally, given $\tau \in (0,1)$, the $100(1-\tau)\,\%$ HDR is the subset

$$L(f_\tau) = \{x \in S^{d-1} : f(x) \geq f_\tau\} \tag{1}$$

where f denotes the density of a random vector X taking values on a $d-$dimensional unit sphere S^{d-1} and f_τ is the largest constant such that

$$P(X \in L(f_\tau)) \geq 1 - \tau$$

with respect to the distribution induced by f. According to Figure 1, if large values of τ are considered, $L(f_\tau)$ is equal to the greatest modes and the most distinct clusters can be easily identified. However, for small values of τ, $L(f_\tau)$ is almost equal to the support of the distribution.

Although there are other alternative routes for estimating Euclidean HDRs, the plug-in approach has received considerable attention in the literature. This is with no doubt a natural methodology, which can be easily generalized to the directional setting. Given a random sample $\mathcal{X}_n \in S^{d-1}$ of the unknown directional density f, plug-in methods reconstruct the $100(1-\tau)\,\%$ HDR namely $L(f_\tau)$ in (1) as

$$\hat{L}(\hat{f}_\tau) = \{x \in S^{d-1} : f_n(x) \geq \hat{f}_\tau\}$$

where $\hat{f}_\tau$ is an estimator of the threshold f_τ and f_n denotes a nonparametric directional density estimator.
`HDiR` package provides tools for directional (circular and spherical) HDRs exact computation also including their plug-in estimation. This library also implements the first specific bandwidth selector devised for directional HDRs proposed in [1], but it also allows to use the existing directional bandwidth selection methods devised for kernel density estimation. Moreover, confidence regions for circular HDR are also available and can be depicted for illustration. Two exploratory tools are also implemented. The first one is a scatterplot computed from HDRs plug-in reconstructions. Sample points are coloured according to the directional HDRs in which they fall. Finally, several distances between sets can be also computed. Their roles are crucial to measure the distances between directional clusters or, for instance, to quantify the estimation error between the theoretical HDRs and the corresponding plug-in estimators. All mentioned capabilities of the `HDiR` package will be shown in this talk.

Keywords: Bandwidth selection, directional estimation, Highest Density Regions, `R`.

ACKNOWLEDGEMENTS

Authors acknowledge the financial support of Ministerio de Economía y Competitividad and Ministerio de Ciencia e Innovación of the Spanish Government under Grants MTM2016-76969-P, MTM2017-089422-P, PID2020-118101GB-I00 and PID2020-116587GB-I00 and ERDF.

Referencias

[1] Saavedra-Nieves P., Crujeiras R. M. (To appear). Nonparametric estimation of directional highest density regions. *Advances in Data Analysis and Classification.*

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

CONTROL ESTATÍSTICO DE PROCESOS MEDIANTE O PAQUETE `qcr`

Salvador Naya[1], Javier Tarrío-Saavedra[1], Miguel Flores[2], e Rubén Fernández-Casal[3]

[1]Grupo MODES, CITIC, Departamento de Matemáticas, Escola Politécnica Superior, Universidade da Coruña. Ferrol, Spain.
[2]Grupo MODES, SIGTI, Departamento de Matemática, Facultad de Ciencias, Escuela Politécnica Nacional. Quito, Ecuador.
[3]Grupo MODES, CITIC, Departamento de Matemáticas, Falcultade de Informática, Universidade da Coruña. A Coruña, Spain.

RESUMO

O paquete R `qcr` proporciona un completo conxunto de ferramentas para o control estatístico de procesos (SPC), xa estean definidos por unha variable, por varias (caso multivariante) ou por unha variable funcional. Deste xeito, inclúense gráficos de control univariantes estándar para medidas individuais, $\bar{x}$, desviacións típicas S, rangos R, proporción de elementos defectuosos p, número de unidades non conformes np, número de defectos c, número medio de defectos u, ademais de gráficos con memoria coma os EWMA ou CUSUM. Por outro lado, inclúense os principais gráficos para o control de procesos definidos por varias variables coma o T^2 de Hotelling, o EWMA multivariante ou MEWMA e o MCUSUM. É tamén importante destacar que este paquete tamén proporciona alternativas non paramétricas para o control de procesos, como son os gráficos de control de rangos r, o Q e o S, baseados no concepto de profundidade de datos. Os autores tamén desenvolveron e implementaron no paquete novos gráficos de control para a Fase I e a Fase II do control de procesos a partires de datos funcionais. Aparte das técnicas SPC arriba indicadas, o paquete `qcr` tamén inclúe un completo conxunto de ferramentas para a avaliación da capacidade dun proceso para cumplir especificacións marcadas polas normas, pola xerencia ou polos clientes. Especificamente, móstrase un amplo abano de índices de capacidade de proceso, tanto paramétricos como non paramétricos.

Palabras e frases chave: SPC. SQC. Gráficos de control. Análise de capacidade. R.

1. DESCRIPCIÓN DO PAQUETE

Seguidamente se describen as funcións implementadas no paquete qcr que permiten a construcción de gráficos de control univariantes, multivariantes e funcionais. Específicamente, no caso no que o proceso que se quere controlar estea definido por só unha variable, sexa esta cuantitativa ou cualitativa, o paquete qcr dispón das seguintes funcións.

Gráficos de control para variables:

- `qcs.xbar`, que permite o cálculo do gráfico $\bar{X}$, desenvolvido a partir da representación dos valores medios dunha variable cuantitativa que representa un proceso que se pode definir mediante lotes ou grupos de medicións homoxéneas.
- `qcs.R`, proporciona utilidades para contruir un gráfico R ou de rangos, que é o gráfico de control máis común para controlar a dispersión. Nun gráfico R, se representan en dúas dimensións os rangos de diferentes medidas tomadas sobre unidades agrupadas en mostras racionais.

- `qcs.S` proporciona o gráfico de desviacións típicas, S, que serve para controlar a dispersión dun proceso cando este se pode medir a partir de lotes ou mostras racionais.
- `qcs.one` proporciona un xeito de construir os gráficos para medidas individuais ou I. Son gráficos que permiten controlar a posición dun proceso cando este non se pode dividir en submostras de máis dunha observación. É a alternativa aos gráficos $\bar{X}$ cando isto acontece.

Gráficos de control para atributos:

- `qcs.p` é unha función que permite o cálculo do gráfico de control para a proporción de unidades defectuosas en cada un dos lotes estudados, etiquetado como gráfico p. Asúmese distribución binomial e a súa aproximación á distribución normal.
- `qcs.np` proporciona a estimación do gráfico de control np, no que se representa o número de unidades defectuosas por lote ou mostra racional. Asúmese distribución binomial e a súa aproximación á distribución normal.
- `qcs.c` é a función que permite o cálculo do gráfico c, é dicir, do número de non conformidades ou defectos por unidade. Asúmese que o número de defectos segue unha distribución de Poisson.
- `qcs.u` permite a estimación e representación do gráfico de control u o de medias do número de defectos por unidade.
- `qcs.g` é unha función que estima e representa o gráfico g, que é o gráfico bidimensional do número de eventos entre unidades defectuosas.

Gráficos de control con memoria (aplicables tanto a variables coma a atributos):

- `qcs.cusum` permite a estimación do gráfico con memoria das sumas acumuladas de diferencias con respecto a media do proceso (CUSUM). Este gráfico permite a detección de cambios pequenos no proceso (menores en magnitude que dúas desviacións típicas con respecto á media do proceso). Neste paquete está dispoñible a versión para medidas individuais.
- `qcs.ewma` permite a estimación do gráfico con memoria das medias móviles ponderadas exponencialmente (EWMA). Este é un gráfico ideal para detectar cambios pequenos no proceso, ao igual que o gráfico CUSUM. Ademais ten a vantaxe de ser máis robusto que outras alternativas cando os datos non se distribuen normalmente ou cando, incluso, as sucesivas observacións están lixeiramente autocorreladas. Neste paquete está dispoñible a versión para medidas individuais e para mostras racionais.

Gráficos de control para datos multivariantes, é dicir, cando o proceso está definido por máis dunha variable, que é o caso máis frecuente, sobre todo no marco da Industria 4.0:

- `mqcs.t2` é unha función que estima o gráfico T^2 de Hotelling, permitindo o control de procesos definidos por máis de unha variable.
- `mqcs.mcusum` permite o cálculo do gráfico CUSUM multivariante ou MCUSUM para vectores de medidas individuais. É unha alternativa para o caso no que os cambios a detectar no proceso sexan polo común relativamente pequenos.
- `mqcs.ewma` é unha función definida para construir os gráficos de control EWMA multivariantes ou MEWMA, definidos para medidas individuais. Ao igual que os gráficos MSUCUM, están deseñados para a detección de cambios pequenos con respecto á media do proceso.

Moitas veces, os vectores de datos non se distribuen segundo unha distribución normal multivariante, polo que, para a detección das anomalías do proceso, do xeito máis fiable posible, é preciso a utilización de versións non paramétricas (sen asumir unha distribución paramétrica de partida) para os gráficos de control multivariantes. Unha das familias de gráficos non paramétricos máis utilizada é a que se basea no concepto de profuncidade de datos. A continuación amósanse as funcións implementadas no paquete `qcr` que permiten a construcción de gráficos de control non paramétricos:

- `npqcs.r` é unha función que estima o gráfico de rangos a partir de vectores multivariantes. Primeiramente se calculan as profundidades multivariantes de cada vector mediante métricas coma a de Mahalanobis, Tukey, profundidade simplicial ou de proxeccións aleatorias qcr. Seguidamente calcúlase o estatístico

$$r_G(y) = P\{D_G(Y) \leq D_G(y) \mid Y \sim G\},$$

onde $Y \sim G$ amosa que os valores de Y se distribúen seguindo a distribución G. Unha práctica habitual cando G é descoñecida é a de utilizar a distribución empírica definida coma G_m a partir da mostra $\{Y_1, \ldots, Y_m\}$, redefinindo o estatístico de rango coma

$$r_{G_m}(y) = \frac{\#\{D_{G_m}(Y_j) \leq D_{G_m}(y), j = 1, \ldots, m\}}{m}.$$

É importante destacar que o rango dunha nova observación calcúlase a partir da súa comparación cunha mostra de calibrado $\{Y_1, \ldots, Y_m\}$, que se asume coma homoxénea e representativa do estado actual do proceso que se está a controlar.

- `npqcs.Q` permite o cálculo do gráfico Q, que é a versión non paramétrica, baseada na profundidade de datos multivariantes, do gráfico de control de medias $\bar{X}$.

- `npqcs.S` proporciona unha versión multivariante dos gráficos de control con memoria, ideais para detectar cambios pequenos no proceso, equivalentes aos gráficos MCUSUM.

Para obter máis información acerca da expresión e uso dos gráficos de control non paramétricos, consúltese o traballo de Flores et al. [1].
En termos xerais, cando a variable que define un proceso é unha curva ou volume definido con respecto ao tempo ou a frecuencia, o control do mencionado proceso poderá levarse a cabo utilizando os denominados gráficos de control para datos funcionais, tamén chamados gráficos de perfís. Neste eido, o paquete `qcr` implementa unha serie de funcións que permiten a estimación de gráficos de control para a Fase I e a Fase II propostos por Flores et al. [2].

- `fdqcs.depth` é unha función que proporciona un gráfico para a Fase I do procedemento para o control dun proceso cando a variable que o define é funcional. A Fase I consiste na estimación da distribución da variable crítica para a calidade do proceso mediante a detección e eliminación dos valores atípicos. Neste caso, os valores atípicos detéctanse mediante métodos baseados no cálculo da profundidade de datos funcionais, é dicir, na profundidade das curvas que definen a calidade do proceso a medir.

- `fdqcs.rank` é unha función que proporciona un gráfico de control do tipo Fase II, cando os datos de partida que definen ao proceso son curvas con respecto ao tempo ou frecuencia, é dicir, datos funcionais. Especificamente, o gráfico Fase II proposto é un gráfico non paramétrico de rangos.

- `plot. fdqcs` é a función que proporciona a envolvente de curvas máis profundas, definindo previamente unha proporción específica para as mesmas. Na saída gráfica correspondente, indícanse aquelas curvas fóra da envolvente, detectadas como anomalías ou estados fóra de control no gráfico de rangos correspondente.

Aparte das ferramentas SPC, o paquete `qcr` inclúe unha serie de utilidades para o cálculo de índices de capacidade paramétricos e non paramétricos, mediante a funcións `qcs.cp`, ademais de proporcionar unha saída gráfica que permite discernir se un proceso é capaz ou non, e, en caso de non se capaz de cumplir especificacións, indicar se é debido a súa variabilidade ou debido á existencia dun nesgo con respecto o obxectivo ou target. A Figura 1 amosa a saída gráfica da análise de capacidade do proceso consistente no paso de buques portacontenedores a través do novo Canal Ampliado de Panamá, específicamente pola esclusa Agua Clara (no océano Atlántico), en dirección sur [**?**]. Calculouse o índice de capacidade de cuarta xeración C_{pm}. O eixo Y indica desplazamentos con respecto ás especificacións debidos á variabilidade do proceso, namentres que

o eixo X indica desplazamentos con respecto os límites de especificación debidos á posición do proceso. Amósase tamén unha rexión para a que o proceso, segundo o indicador C_{pm}, é capaz de cumplir especificacións. Neste caso, o proceso de paso a través da esclusa Agua Clara non é capaz de cumplir as especificacións de paso no intervalo [50 min, 180 min], marcados pola Autoridade do Cana de Panamá, debido a un problema de sesgo ou posición (tárdase máis en media en pasar a través da esclusa).

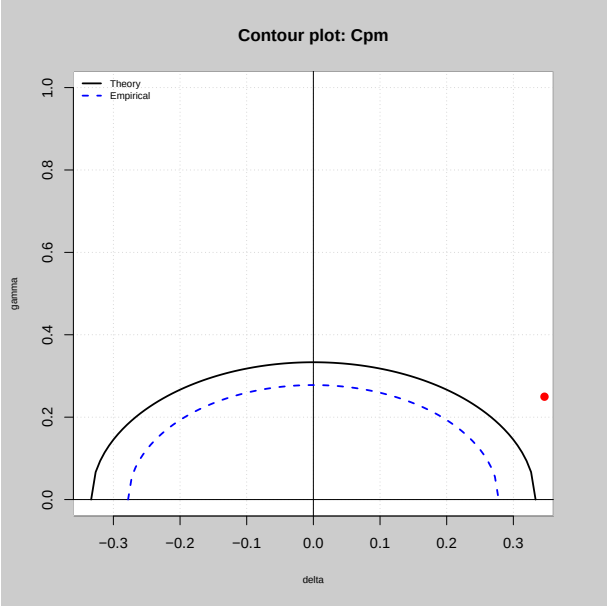


Figura 1: Saída gráfica da análise de capacidade do proceso de paso a través da esclusa Agua Clara, dirección sur, do Canal Ampliado de Panamá.

AGRADECEMENTOS

O traballo foi financiado polos proxectos MINECO MTM2017-82724-R e PID2020-113578RB-100, ademais de pola Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 e Centro de Investigación do Sistema universitario de Galicia ED431G2019/01), todos eles a través do FEDER.

Referencias

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

[2] Flores M., Naya S., Fernández-Casal R., Zaragoza S., Raña P., Tarrío-Saavedra, J. (2020). Constructing a control chart using functional data. `Mathematics`, 8(1), 58.

VIII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 14 de outubro do 2021

EXEMPLOS DE USO DE R NA EMPRESA

Antonio Vidal Vidal[1]

[1]Senior Data Scientist Nextail

RESUMO

Esta presentación resume máis de 10 años de uso de R para o desenrolo de proxectos de análise de datos nas empresas, sempre con unha perspectiva práctica e de aplicación aos problemas que se atopaban.

Aínda que R naceu como unha linguaxe orientada ao análise estatístico, o desenrolo ao longo dos años de multitude de paquetes con funcionalidades moi diversas que complementan a linguaxe base, fai que na actualidade sexa posible usalo para o desenrolo de aplicacións onde a xestión dos datos é igual de importante que o desenrolo do propio programa. Precisamente neste ámbito, o da exploración e tratamento de datos, é onde o seu uso cobra moito máis sentido.

Nesta presentación farase un resumo das aplicacións desenroladas en R para dar solución a unha ou máis funcionalidades das que se describen a continuación:

- Captura e recollida de datos
- Automatización de accións
- Tratamento de datos e movemento entre sistemas
- Análise e modelado de ML
- Xeración de informes e representación de datos

Por último, tamén se fará un resumo de boas prácticas aplicadas ao longo destes anos.

Palabras e frases chave: casos de uso, R na empresa

AUTORES

VIII 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 346622