R-SCool

Escola de Análise de Dados com R

ESCOLA DE INVERNO 2019

R: Linguagem e Pacote

Marcelo Camerlo (ICS-ULisboa)

RESUMO

O curso oferece uma introdução prática à compreensão da linguagem R e à gestão do pacote R.Studio. O R tem muitas formas de ser 'falado' e as suas aplicações são desenvolvidas constantemente. Mas o seu uso pode ser tão 'intuitivo' como o dos pacotes Excel, SPSS ou STATA. Para tal fim, no curso se discutem e propõem estratégias para um uso eficiente e partilhado. Adicionalmente, são apresentadas as ferramentas e dados que logos serão explicados e utilizados nos outros cursos da R-SCool.

SESSÕES

- **Sessão I**: A Linguagem R
- **Sessão II**: Gestão do 'espaço de trabalho', com R.Studio
- **Sessão III**: Técnicas de análise
-
- **Sessão IV**: Desenhando com R

CONTEÚDOS DETALHADOS

1. **A Linguagem R**

   1.1. Elementos básicos da sintaxe e semântica R (operadores, funções e condicionais).

   1.2 Criação, exploração e manipulação de variáveis, bases de dados e outros objectos R.

2. **Gestão do 'espaço de trabalho', com R.Studio**

   2.1. Administração dos diferentes componentes de R(*working directory, console, workspace, scripts, e historial, quartz*).

   2.2. Instalação de pacotes.

   2.3. Exportação e importação de dados (STATA, SPSS, Excel).

   2.4. Exploração das bases que serão utilizadas nos outros cursos da R-SCool.

3. **Técnicas de Analise**

   3.1. Execução de técnicas de análise básicas (estatística descritiva, tabelas de contingência, correlações e t-test, ANOVA, MANOVA e regressão múltipla).

   3.2. Simulações.

4. **Desenhando com R**

   Elaboração de:

   - Histogramas
   - Gráficos de barras
   - Gráficos de linhas
   - *Boxplots*
   - *G*ráficos bivariados

RESEARCH DESIGN, The Positivist Approach

Ignacio Lago (Universitat Pompeu Fabra)

OVERVIEW

This course focuses on research design in social sciences, that is, how to design our empirical research in order to make valid causal and descriptive inferences in political and social life. The course is thought for the use of observational data from a positivist approach. In the last session(s), the students will present their research projects.

SESSIONS

- **Session I**: Major components of research design & Causality
- **Session II**: Explanations in Social Sciences
- **Session III**: Estimating causal effects
- **Session IV**: Indeterminate research designs
- **Session V**: Selection of observations
- **Session VI**: Understanding what to avoid (a)
- **Session VII**: Understanding what to avoid (b)
- **Session VIII**: Case-studies

DETAILED CONTENTS

1. **Introduction**
   - 1.1. Key concepts
   - 1.2. Defining scientific research in Social Sciences
   - 1.3. Major components of research design

2. **Causality**
   - 2.1. Description versus explanation
   - 2.2. Causal effects: causal mechanisms and correlation
   - 2.3. Causality versus spuriousness

3. **Explanations in Social Sciences**
   - 3.1. Statistics
   - 3.2. Covering laws
   - 3.3. Causal mechanisms

4. **Estimating causal effects**
   - 4.1. Assumptions required for estimating causal effects.
   - 4.2. Rules for judging causal inferences

5. **Indeterminate research designs**
   - 5.1. More inferences than observations – degrees of freedom
   - 5.2. Multicollinearity

6. **Selection of observations**
   - 6.1. Selection bias
   - 6.2. Selecting observations on the explanatory variable

7. **Understanding what to avoid (a)**
   - 7.1. Specification bias:
     - a) Excluding relevant variables
     - b) Including irrelevant variables: Inefficiency

8. **Understanding what to avoid (b)**
   - 8.1. Endogeneity
   - 8.2. Measurement errors

9. **Case-studies**

READING

FIREBAUGH, GLENN. 2008. *Seven Rules for Social Research.* Princeton: Princeton University Press.

GERRING, JOHN. 2012. *Social Science Methodology. An Unified Framework.* Cambridge: Cambridge University Press.

KING, Gary, Robert KEOHANE y Sidney Verba. 1994. *Designing Social Inquiry. Scientific Inference in Qualitative Research.* Princeton: Princeton University Press.

LAGO, Ignacio. 2017. *La lógica de la explicación en ciencias sociales. Una introducción metodológica*. Madrid: Alianza.

ESTATÍSTICA I: Fundamentos e Ferramentas Básicas

Marcelo Camerlo (ICS-Ulisboa)
Helena Carvalho (ISCTE-IUL)

RESUMO

O curso oferece uma introdução aos conceitos e instrumentos básicos da análise estatística, a partir de uma abordagem atenta os desafios metodológicos da investigação social e em diálogo com a análise qualitativa. Está principalmente orientado para dois destinatários: estudantes interessados em iniciar uma formação quantitativa; e investigadores qualitativos à procura de uma utilização sistemática da estatística básica.

SESSÕES

- **Sessão I**: Fundamentos da Análise Quantitativa
- **Sessão II**: Análise exploratória Univariada
- **Sessão III**: Análise exploratória Bivariada (a)
- **Sessão IV**: Análise exploratória Bivariada (b)
- **Sessão VI**: Inferência Estatística (a)
- **Sessão VI**: Inferência Estatística (b)
- **Sessão VII**: Validação e Generalização de Hipóteses (a)
- **Sessão VIII**: Validação e Generalização de Hipóteses (b)

CONTEÚDOS DETALHADOS

1. **Fundamentos da análise quantitativa**
   - 1.1. O raciocínio estatístico
   - 1.2. Descrição vs. inferência; descrição vs. explicação
   - 1.3. Tipos de variáveis e tipos de relações entre variáveis
   - 1.4. Bases de dados
   - 1.5. Estudos observacionais vs. experimentais
   - 1.6. Associação vs. causalidade

2. **Análise exploratória univariada (Estatística descritiva I)**
   - 2.1. Frequências e proporções
   - 2.2. Medidas de tendência central e de dispersão
   - 2.3. Valores extremos. Valores omissos
   - 2.4. Visualização: tabelas, gráfico de barras, histograma, gráfico de linhas e diagrama de extremos e quartis

3. **Análise exploratória bivariada (Estatística descritiva II)**
   - 3.1. Tipos de relações entre variáveis independentes e variáveis dependentes
   - 3.2. Variáveis latentes
   - 3.3. Medidas de associação (comparação de proporções; coeficientes de correlação, e regressão linear)
   - 3.4. tabela de dupla entrada, diagrama de extremos e quartis para comparação de grupos, gráfico de dispersão

4. **Inferência Estatística**
   - 4.1. Distribuição e valor-Z
   - 4.2. Teorema do limite central
   - 4.3. Nível de confiança e intervalos de confiança
   - 4.4. Teste de hipóteses
   - 4.5. Nível de significância e valor de *p*

5. **Validação e Generalização de Hipóteses**
   - 5.1. Hipóteses estatísticas: nula e alternativa
   - 5.2. Teste de hipóteses bivariadas (*teste do qui-quadrado,* ANOVA, testes *t*)
   - 5.3. Significância estatística e prática

BIBLIOGRAFIA

Diez, David, Christopher Barr e Mine Çetinkaya-Rundel (2012). OpenItro Statistics. 3rd edition. Free textbook (http://www.openintro.org).

ESTATÍSTICA II: Análise Factorial e Regressão Linear Múltipla

Helena Carvalho (ISCTE-IUL)
Marcelo Camerlo (ICS-Ulisboa)

RESUMO

O curso está organizada em dois eixos analíticos:

1. O primeiro eixo centra-se na operacionalização de dimensões de análise através de novas variáveis compósitas. Para a identificação das novas variáveis compósitas é usada a Análise em Componentes Principais (ACP) para variáveis quantitativas ou tratadas como tal.
2. O segundo eixo tem por objetivo construir modelações com as novas variáveis compósitas (quantitativas) e outras variáveis (quantitativas ou dummy), a fim de explorar relações de dependência entre elas. Para o efeito usa-se a Regressão Linear Múltipla.

O desenvolvimento dos dois eixos analíticos é introduzido por uma abordagem teórico-prática dos métodos, seguido de uma abordagem mais aplicacional, assente na interpretação de casos práticos.

SESSÕES

- **Sessão I**: Análise em Componentes Principais – conceptualização
- **Sessão II**: Análise em Componentes Principais – interpretação
- **Sessão III**: Análise em Componentes Principais - reportar resultados em artigo, tese
- **Sessão IV**: Validade do constructo
- **Sessão VI**: Regressão Linear Múltipla – conceptualização
- **Sessão VI**: Regressão Linear Múltipla – interpretação
- **Sessão VII**: Regressão Linear Múltipla com blocos hierárquicos
- **Sessão VIII**: Reportar resultados em artigo, tese

CONTEÚDOS DETALHADOS

1. Análise de Componentes Principais (ACP)

1. Adequabilidade da matriz de *input* à aplicação da ACP
2. Definição das componentes principais
3. Definição de valor próprio e de comunalidade
4. Critérios de extração das componentes principais
5. Métodos de rotação das componentes principais
6. Aplicações com o *software* R
7. Interpretar e reportar resultados em tese, artigo

2. Validade do constructo

2.1. Validar estrutura unidimensional
2.2. Consistência (via Alpha de Cronbach)
2.3. Construção de novas variáveis compósitas
2.4. Aplicações com o *software* R
2.5. Interpretar e reportar resultados em tese, artigo

3. Regressão Linear Múltipla

3.1. Correlação linear vs regressão linear
3.2. Modelo de regressão linear múltipla
3.3. Qualidade e adequabilidade do modelo
3.4. Variáveis preditoras (coeficientes B e Beta, teste t)
3.5. Operacionalização com *software* R
3.6. Apresentação de resultados em tese, artigo

4. Regressão Linear Múltipla com blocos hierárquicos

4.1. Definição dos blocos
4.2. R^2 *change* e F *change*
4.3. Operacionalização com *software* R
4.4. Apresentação de resultados em tese, artigo

BIBLIOGRAFIA

Hair, J., Anderson R., Tatham, R. e Black, W. (2010). Multivariate Data Analysis: A Global Perspective, Upper Saddle River, Pearson International Edition (7ª ed).

Maroco, J. (2010). Análise Estatística com o PASW Statistics (ex-SPSS), Pero Pinheiro, ReportNumber.

Tabachnick, B. & Linda F. (2000). Computer-assisted research design and analysis, Boston: Ally and Bacon.

MAPAS E GEOGRAFIA ELEITORAL

Rodrigo Rodrigues-Silveira
(*Instituto de Iberoamérica* -Universidad de Salamanca)

RESUMO

O curso tem como objetivo oferecer um marco metodológico básico para a análise das questões sociais a partir de uma perspectiva espacial. Nesse sentido, consiste em um curso de formação predominantemente introdutória e metodológica (ainda que inclua um panorama dos principais conceitos da geografia política), com vistas tanto a familiarizar os alunos com a análise da dimensão territorial dos fenômenos sociais como oferecer-lhes os instrumentos necessários tanto para incorporar o território em seus trabalhos científicos como para interpretar as publicações especializadas na área. Espera-se que, no final do curso, o aluno seja capaz de incorporar informações de diferentes bases de dados a um instrumental básico de análise espacial e que apresente condições de continuar seu desenvolvimento na área de modo autônomo.

SESSÕES

- **Sessão I:** Conceitos básicos da geografia política
- **Sessão II:** Fontes e georreferenciamento dos dados
- **Sessão III:** Visualização Espacial de Dados I – Procedimentos básicos, classificação e uso da cor
- **Sessão IV:**Visualização Espacial de Dados II – Mapas Temáticos
- **Sessão V:**Análise Exploratória Espacial de Dados (ESDA)
- **Sessão VI:** Autocorrelação Espacial I – Vizinhança e Autocorrelação Global
- **Sessão VII:** Autocorrelação Espacial II - Análise de Clusters Espaciais (LISA)

CONTEÚDOS DETALHADOS

1. Conceitos básicos da geografia política

1.1. Apresentação do curso e elementos básicos da análise espacial
1.2. O nacionalismo metodológico e seus limites
1.3. Território e territorialidade
1.4. Fronteira
1.5. Escalas
1.6. Local/lugar/paisagem
1.7. A centralidade da ideia de contexto

2. Fontes e georreferenciamento dos dados

Este sessão está dedicado à preparação de dados para a análise espacial. Descrevem-se aqui as principais diferenças entre dados espaciais e não espaciais, suas propriedades básicas e as diferentes estratégias de agregação de dados para seu ajuste a distintas unidades espaciais de referência. Finalmente, o processo de georreferenciamento de dados é explicado passo a passo.

3. Visualização Espacial de Dados I – Procedimentos básicos, classificação e uso da cor

Neste tópico, são explorados os tipos básicos de mapa vetorial (pontos, linhas e polígonos) e os procedimentos básicos necessários para a visualização de dados sociais na forma de mapas. Em particular, três aspectos pontuais são examinados. O primeiro corresponde à composição (*layout*) dos mapas e a localização adequada dos diferentes elementos gráficos empregados (título, escala, rosa dos ventos, legenda). O segundo refere-se aos diferentes modos de classificação de dados e suas consequências para a interpretação das informações que se deseja comunicar. Finalmente, o terceiro aspecto consiste em uma breve introdução às técnicas básicas do uso da cor nos mapas e aos cuidados essenciais exigidos na confecção de uma paleta de cores.

4. Visualização Espacial de Dados I – Procedimentos básicos, classificação e uso da cor

A representação visual das informações sociais no espaço muda segundo a natureza e finalidade do mapa que se deseja gerar. O percentual de votos num partido, por exemplo, costuma ser representado por um mapa de área com graus progressivos de cor que vão dos tons mais claros aos mais escuros (de coropletas). A localização das escolas nos bairros de uma metrópole é representada em geral por meio de um mapa de pontos (densidade de pontos). A concentração ou dispersão destas mesmas escolas é representada por um mapa que utiliza cores mais quentes para as áreas em que se observa maior concentração e mais frios para a dispersão (isolinhas). Outra aplicação pode ser a identificação dos fluxos migratórios (entre países, por exemplo) por meio de flechas (mapa de fluxos).

Estes diversos tipos de mapa enquadram-se sob a categoria de mapas temáticos e servem para representar quantidades, localização e fluxos. O objetivo do quinto tema é precisamente ensinar: (a) a discriminar qual tipo de mapa temático é o mais apropriado para representar o tipo de informação utilizada; e (b) como realizar os principais tipos de mapas temáticos utilizando exemplos das ciências sociais.

5. Análise Exploratória Espacial de Dados (ESDA)

A análise exploratória Espacial de dados (ESDA em sua sigla em inglês) corresponde ao conjunto de técnicas de análise do formato e dos padrões de distribuições dos dados no

espaço. Este tipo de análise é bastante conhecido nas ciências sociais nas que ferramentas ou técnicas como a correlação, o histograma, o boxplot, o scatterplot são de amplo uso. A inovação introduzida pelo ESDA consiste em combinar tais elementos com mapas, considerando estes últimos como um instrumento complementário aos anteriores na identificação de padrões nos dados e na formulação de hipóteses de pesquisa. Seu objetivo é identificar padrões espaciais nos dados analisados.

6. Autocorrelação Espacial I – Vizinhança e Autocorrelação Global

Esta sessão dedica-se inteiramente à introdução aos conceitos fundamentais da análise estatística espacial: o de vizinhança (que é a operacionalização do conceito mais geral de contexto) e autocorrelação espacial. Será descrito o que se entende por matriz de vizinhança e autocorrelação espacial. Também são examinados os diversos tipos de matrizes de vizinhanças e como gera-las no R. Uma vez obtidas tais matrizes, será apresentado o conceito de autocorrelação espacial, a diferença entre a autocorrelação global e local, assim como quais as consequências para a avaliação de relações de associação espacial nos fenômenos sociais. Em particular, será ensinado como ponderar um indicador utilizando o contexto espacial como critério base (variável *spatial lagged*) e gerar um *scatterplot* de Moran, que consiste no primeiro passo na medida da associação espacial das unidades espaciais. Finalmente, são introduzidas diferentes medidas de autocorrelação espacial global como o I de Moran ou o C de Geary e sua aplicação na análise exploratória espacial de dados.

7. Autocorrelação Espacial II - Análise de *Clusters* Espaciais (LISA)

Esta parte do curso trata da autocorrelação espacial local, que permite identificar aglomerados (*clusters*) espaciais ou *hot-spots* na distribuição de um ou mais indicadores. A ausência de padrões espaciais globais, isto é, de uma lógica espacial que explique a distribuição da maior parte dos casos, não exclui a existência de uma organização espacialmente significativa em certas partes ou regiões do território. Por esta razão, o emprego de distintas técnicas de análise de *cluster* espacial pode ser uma ferramenta bastante útil para entender porque alguns indicadores possuem um comportamento particular em certos contextos geográficos específicos.

BIBLIOGRAFIA

Anselin, Luc. 1999. “Interactive techniques and exploratory spatial data analysis”. In P. Longley, M. Goodchild, D. Maguire, D. Rhind (Eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications*, pp. 251-264. New York.

Bivand, Roger S., Edzer J. Pebesma, Virgilio Gómez-Rubio. 2008. *Applied spatial data analysis with R*. New York: Springer. Cap. 9 – “Areal Data and Spatial Autocorrelation”.

Fazal, Sahab. 2008. “Spatial Data Structures and Models”. In S. Fazal (Ed.), *GIS Basics*. New Dehli: New Age International.

Hair, J., Anderson R., Tatham, R. e Black, W. (2010). Multivariate Data Analysis: A Global Perspective, Upper Saddle River, Pearson International Edition (7ª ed).

Slocum, Terry A. 1999. *Thematic Cartography and Visualization*. Upper Saddle River, NJ: Prentice Hall.