Rapport de mission useR!2017 - Bruxelles

Frédéric Vanwindekens

04.07.2017 - 07.07-2017

Contents

1 Généralités

1.1 Introduction

Du 4 au 7 juillet 2017, j'ai assisté à la conférence internationale annuelle "useR!" dont l'objet est l'état d'avancement du développement du langage de traitement de données R, les nouvelles voies explorées. Cette année, elle était organisée par des partenaires belges à Bruxelles.

Pour moi, l'intérêt de participer à la conférence était multiple

- de présenter les développements méthodologiques réalisés dans le cadre du projet MIMOSA (CMASOP) et le package R en développement lié à cette approche (RCogMap)

- de suivre les dernières évolutions et tendances dans le domaine de la cartographie

1.2 Points marquants pour tous

1.2.1 Tidy way of thinking (#dplyr #tidyr #ggplot @tom @pat)

According to the 'tidy data' paper (Wickham 2014), data is tidy when

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Programming with tidyverse grammars

- `http://sched.co/Axoq`
- `https://github.com/tidyverse/rlang`

1.2.2 Reproductible science (#rmarkdown)

Apprendre Rmarkdown ! + bonnes pratiques, voir détails ci-dessous.

1.2.3 Cartographie (#meteo #satellite #mapping #openstreetmap #dynamicmap)

1. Cours pour le traitement de données spatiale (ULg, Lisboa, Montpellier) `http://www.openspat.eu/`
2. open data météo Allemagne : (Deutsh Weather Data)
   - a package R for that : rdwd
   - leaflet
   - `https://github.com/brry/prectemp`
3. New direction : sf package Important [à creuser!, see details Spatial Data in R: New Directions]

   Avis sur les outils :

   - "sf is the future" (en remplacement de sp) : plus de caractères à écrire, mais plus flexible et plus précis .

- pour les cartes statiques : "ggplot2 is the future", préférez $\text{geom}_{\text{polygon}}$ à geom_{map}. (ggviz : avis négatif).
- pour les cartes dynamiques : leaflet, avis ++, activement maintenu par RStudio (shiny : avis pas enthousiaste pour les cartes).

4. Package lié à OSM, ... - Maps are data, so why plot data on a map? osmplotr, osmdata (see details)

   Présentation EXTRA !!: `https://schd.ws/hosted_files/user2017/88/padgham.pdf`

   - vidéo : `https://channel9.msdn.com/Events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/Maps-are-data-so-why-plot-data-on-`

   à suivre : premier test ok via le script R `file://FILESERVER/Commun/Documentation/Statistiques/ScriptExemples/essaicartoosmplotr.R` (dossier réseau)

   Mais pas encore beaucoup de fonctionnalités (??), test plus avancé non concluant.) creuser.

5. linking R and other GIS (QGIS/GRASS/...)

   - Package RQGIS : `http://sched.co/AxpV`
   - Package Link2GI : `http://sched.co/Axpb`

6. Raster et satellite : développement de STARS Scalable, Spatiotemporal Tidy Arrays for R (stars)

   `http://sched.co/AxpE`

1.2.4 Autres thèmes importants (mais non creusés, à creuser)

1. Parallel computing
2. Modules in R voir détails (dernier jour), c'est entre un script et un package

1.3 Points d'intérêt plus spécifiques

1.3.1 KEYNOTE: Teaching data science to new useRs

`https://channel9.msdn.com/events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/KEYNOTE-Teaching-data-science-to-new-useRs?term=tidyverse`

1.3.2 A Spreadsheet for R : Jamovi (@Mary, @Patoche, . . .)

Intéressant, à tester : https://www.jamovi.org/

1.3.3 URBAN GREEN SPACES AND THEIR BIOPHONIC SOUNDSCAPE COMPONENT

Analyse des sons des environnements (ex. oiseaux, . . .)
slides : https://schd.ws/hosted_files/user2017/51/useR%212017_Brussel___distrib_PaulDevos.pdf
Packages : soundecology and seewave

1.3.4 Dose-response (@Gilles)

KEYNOTE: Dose-response analysis: considering dose both as qualitative factor and quantitative covariate- using R*
http://sched.co/Ay02

1.3.5 text mining

Package **tidytext** à creuser

1.3.6 Morphological Analysis in R

Peut-être intéressant pour multicritères/LCA ?? Voir les slides : http://schd.ws/hosted_files/user2017/ba/morph-talk-user2017.7z (exemple à la fin sur les t de CO2 en fonction des choix)

1.4 Liens utiles

1.4.1 Programme et liens vers les ressources

https://user2017.brussels/schedule

1.4.2 Vidéos des présentations, slides + son :

- all : https://channel9.msdn.com/Events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference
- spatial : https://gist.github.com/anonymous/3d5b56cb16526db96dcaa0a579980187
- une keynote importante sur les graphiques (Hadley Wickham): https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/Towards-a-grammar-of-interactive-graphics

- R Extracting data from the web APIs and beyond : `https://github.com/ropensci/user2016-tutorial`, @tom @jean-pierre #geekhighxslevel

1.4.3 Autres liens

- Herbergement de projets intéressants : `https://github.com/ropensci`
- Nasa Datanauts (open data) : `https://open.nasa.gov/explore/datanauts/`

1.5 Contacts réalisés

- Paul Devos - Ghent University - Department INTEC – WAVES
- Emily Burchfield - Vanderbilt University, Nashville USA (Energy and Environment)
- Benjamin Ortiz Ulloa - Data Viz Engineer

2 Détails de la mission, des présentations

2.1 4 juillet : Tutoriels

2.1.1 Geospatial Data Visualization Using R

à creuser : codes disponibles, etc `https://github.com/bhaskarvk/user2017.geodataviz`

2.1.2 Spatial Data in R: New Directions

`https://edzer.github.io/UseR2017/`

Centered on the package sf (simple features)

1. In comparison to sp, package sf
   - implements all types and classes of simple features
   - has no support for gridded (raster) data
   - uses data.frames for features, and list columns for feature geometry
   - uses S3 instead of S4
   - is also built on top of GDAL, GEOS and Proj.4
   - tries to provide a simpler and cleaner API (or user experience) conversions to/from sp makes it still easy to work "backwards"

2. What's really new, and better in sf, compared to sp?

   - simple features is a widely adopted standard
   - tidyverse compatibility
   - ggplot2 support ($\text{install}_{\text{github}}$, under development)
   - support for measurement units
   - partial support for computations using geographic coordinates
   - support by mapview, tmap, mapedit; 15 revdeps on CRAN
   - binary geom ops: fast (indexed), low memory footprint; flexible st_{join}
   - (c)lean Rcpp interface to external dependencies GDAL/GEOS/Proj.4
   - fast WKB (de)serialization, in C++

2.1.3 Data Carpentry – Open and Reproducible Science with R

`http://rpubs.com/minebocek/user2017-ors` `https://github.com/fmichonneau/2017-useR-reproducibility`

1. Four facets of reproducibility

   Documentation explanation and commenting of why and how an analysis is carried out in human readable language

   Organization tools to organize your projects so that you don't have a single folder with hundreds of files

   Automation the power of scripting to create automated (and self documenting) data analyses

   Dissemination publishing is not the end of your analysis, rather it is a way station towards your future research and the future research of others

2. Three key principles for (file) names

   (a) Machine readable

   - easy to search for files later
   - easy to narrow file lists based on names
   - easy to extract info from file names (regexp-friendly)

   (b) Human readable

- name contains information on content, or
- name contains semantics (e.g., place in workflow)

(c) Plays well with default ordering

- use numeric prefix to induce logic order
- left pad numbers with zeros
- use ISO 8601 standard (YYYY-mm-dd) for dates

3. Organisation of the R folder

Some organizing principles for files

data-raw the original data, do not edit or directly alter any of the files in this folder.

data-output intermediate datasets that will be generated by the analysis. We write them to CSV or other portable formats, particularly useful when it take a long time (or expensive resources) to generate.

fig where we can store the figures used in the manuscript.

R our R code (the functions)

- For large projects it often becomes easier to keep the prose separated from the code.
- If you have a lot of code (and/or manuscript is long), it's easier to navigate.
- Or just create an R package.

tests the code to test that our functions are behaving properly and that all our data is included in the analysis.

```
|
+-- data-raw/
|   |
|   +-- gapminder-5060.csv
|   +-- gapminder-7080.csv.csv
|   +-- ....
|
+-- data-output/
|
+-- fig/
|
```

```
+-- R/
|   |
|   +-- figures.R
|   +-- data.R
|   +-- utils.R
|   +-- dependencies.R
|
+-- tests/
|
+-- manuscript.Rmd
+-- make.R
```

4. Licences `http://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1002598&type=printable`

2.1.4 Efficient R programming

Tips
- Timing your code : with system.time() or micorbenchmark()
- Compile function : ByteCompile: true in the DESCRIPTION file !

Rules
1. never grow a vector !
2. vectorise if possible !

More complex
- parallel computing
- Rcpp

2.2 5 juillet

2.2.1 Sponsor talk : Microsoft

1. History :
   - We never heard about R (5th language)
   - R does not do what I need (number of packages !)
   - We can't use R because it is opensource (Microsoft !)
   - We can't use R in a regulated environment (FDA !)
   - R is too slow (Microsoft server, SQL)
   - R is not deeling with big data (spark, ...)

- You can't use R in production : trains, planes, planes, chemicals, ...
- We can't get support for R, Who stands behind R? (microsoft, ...)

now R ecosystem

2.2.2 Keynote : Structural Equation Modeling

Yves Rosseel

multivariate statistical modeling technique. hypothesis/model about the data

special :

- latent variables
- allows indirect effects

comparaison univariate and multivariate regression in multivariate : strict distinction between dependent and independant.

Sem : path analysis all variables are observed (manifest) we allow indirect effect we allow for cycles (ex. reading motivation, reading frequencey, reading ability)

SEM is used in social science, but also in medecine, neuroimaging, biology/ecology(climate change!), operation research...

SEM used in standard analysis (regression) when :

- missing data
- robust standard errors, diagnostics
- (in)equality constraints
- ...

Examples of applications :

- ...
- !!!! growth curve model (random slopes, random intercept)
- potlitical democracy

UNIX philosophy : do one thing, do it well
lavaan : latent variables analysis
Avec le diagramme, on apprend la syntaxe, c'est assez simpe

- =~ latent variables
- ~ regression
- – co-factor (?)

2.2.3 Shiny app for 700 end users

not so interesting

2.2.4 sonification of data

for disabled people, e.g.
not so interesting data music on Shiny

2.2.5 sports data

not interesting

2.2.6 soundecology

interesting !!! Analyse des sons des environnements (ex. oiseaux, ...) slides : `https://schd.ws/hosted_files/user2017/51/useR%212017_Brussel___distrib_PaulDevos.pdf`
Packages : soundecology and seewave

2.2.7 Maps are data, so why plot data on a map?

Mark padgham (Salzburg, Austria)
Interesting !!!
==> Maps are data ! osmdata

- osmplotr

Ex : Anju (north korea) (compare MS, google, et OSM !!)
Full slides (un must) : `https://schd.ws/hosted_files/user2017/88/padgham.pdf`

2.2.8 Text mining session

tidytext (package) count frequency (songs - usa states) sentiment analysis (-5 to +5)

term frequency - inverse document frequency idf(term)=ln(n doc / n doc with the term)

TF-IDF (!! try with survey !!)

2.2.9 R and GIS session

1. Jannes Muenchow

   qgis and R : interaction dans r : ce qui manque : des geoalgorithme

   RSAGA RQGIS RGRASS ...

   Qgis : python API lots of geoalgorithm and python console then also R via reticulate ($\text{apen}_{\text{app}}()$)

   reticulate = interface of python on R

   use functions : $\text{find}_{\text{algorithms}}()$, $\text{get}_{\text{usage}}()$

   library(RGIS) Jannes Muenchow (2013) : plant diversity prediction

2. Ege Rubak : package S2 Interfacing Google's spherical geometry library (S2) for spatial data in R

   It both facilitates geometric operations directly on the sphere such as polygonal unions, intersections, differences etc. without the hassle of projecting data in the common latitude and longitude format, and provides an efficient quadtree type hierarchical geospatial index.

3. Scalable, Spatiotemporal Tidy Arrays for R (stars) Edzer Pebesma anticipate the queston : raster include simple features ask collaboration for developping the package

2.3 6 juillet

2.3.1 Sponsors Talk : OpenAnalytics

Antwerp, consulting. Ex : parasite worms, mexican soup (Dutch women)

- optimal design algorithms for PK/PD modeling
- software
- remote health monitoring for off-shore activities

- risk management platform for decision making European Commission
- Architect : an IDE for data science

2.3.2 Plenary : dose-reponse

Ludwif A. Hoth
Très spécifique - high level !

2.3.3 Networks analysis (NL)

Finlay Scott (JRC)
FLR ecosystem (various packages)

2.3.4 JAMOVI

Jonathon Love spreadsheet for R
different from other : live Demonstration : intéressant. +

2.3.5 data journalism

Timo Grossenbacher very interesting !++

2.3.6 Lightning talks - ecology (really small talk !)

1. deutch weather data a package R for that : rdwd leaflet github.com/brry/prectemp
2. Earth movement package ESEIS toolbox to weld geomorphic, seismologic, spatial and time series analysis

   www.micha-dietze.de
3. WEF : water-energy-food connection, real important for the future swat model, best management practices
4. subsurface hydrology critic : non reproducable, than is it science ?
5. PM2.5 in India
6. Andy South (@southmapr) rnaturalearth sustainability, (soft. sust. institute) plot($ne_{countries}$(country="belgium",scale="medium")

7. open spat Ulg, lisboa, Montpellier supagro très intéressant !!! +

   Learning modules to get skills on free tools

   analyze and interpret spatial data

   prequel online course

   Rmarkdown

   - and scenari (free !)(check !!)

   `http://openspat.eu`

8. Tobias Gauster - package smires European project on Science and Management of Intermittent Rivers and Ephemeral Streams

2.4 7 juillet

2.4.1 KEYNOTE: R tools for the analysis of complex heterogeneous data

bof bof

- phone call migrants / lampedusa disaster
- Sampson's Monks Network (crisis in cloister)
- proteins of yeast (1500)

2.4.2 tidyverse grammar Tidy evaluation

Interesting, but no slides ?? `http://schd.ws/hosted_files/user2017/43/tidyeval-user.pdf`

2.4.3 modules

S. Warnholz à creuser

Un module est qqchose entre un script et un package `https://channel9.msdn.com/events/useR-international-R-User-conferences/useR-International-R-User-2017-C Modules-in-R?term=user%20module`

2.4.4 R and Haskell: Combining the best of two worlds

Haskell is statically typed, purely functional, lazy, fast, and with that, you know, cool and mathy touch ... ;-) `http://sched.co/Axqq`