R for reproducible geographical research

Edzer Pebesma
(joint work with Daniel Nüst* and Roger Bivand**)

ifgi
Institute for Geoinformatics
University of Münster

edzer.pebesma@uni-muenster.de

AAG, Feb 24, 2012, NY, USA

* 52North GmbH, **NHH Bergen

Outline

This is ongoing work, and most of the core ideas are not original.

- What is reproducible research?
- What is R? Why R?
- How can R be used for reproducible geographical research?
- Challenges
- Outlook

Why is reproducible research a good thing?

- the credibility of science is at stake when research is not reproducible
- we cannot repeat observation, but we can repeat the procedures that led us from observations to research findings and conclusions
- even in cases where data cannot be shared, sharing procedures will increase credibility
- even in case of errors, being able to trace them back to the source (data? script? software?) increases credibility

Why is reproducible research a good thing?

- the credibility of science is at stake when research is not reproducible
- we cannot repeat observation, but we can repeat the procedures that led us from observations to research findings and conclusions
- even in cases where data cannot be shared, sharing procedures will increase credibility
- even in case of errors, being able to trace them back to the source (data? script? software?) increases credibility

then, why don't we do this? Why is reproducibility not compulsory?

Claerbout's Principle

Claerbout's Principle[1]:

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

[1] J. de Leeuw, Reproducible research: the bottom line, statistics Program, Unniversity of California, Los Angeles, CA, USA (2001), quoting: J. Buckheit, D. Donoho, WaveLab and reproducible research, statistics Department, Stanford University, CA, USA (1995).

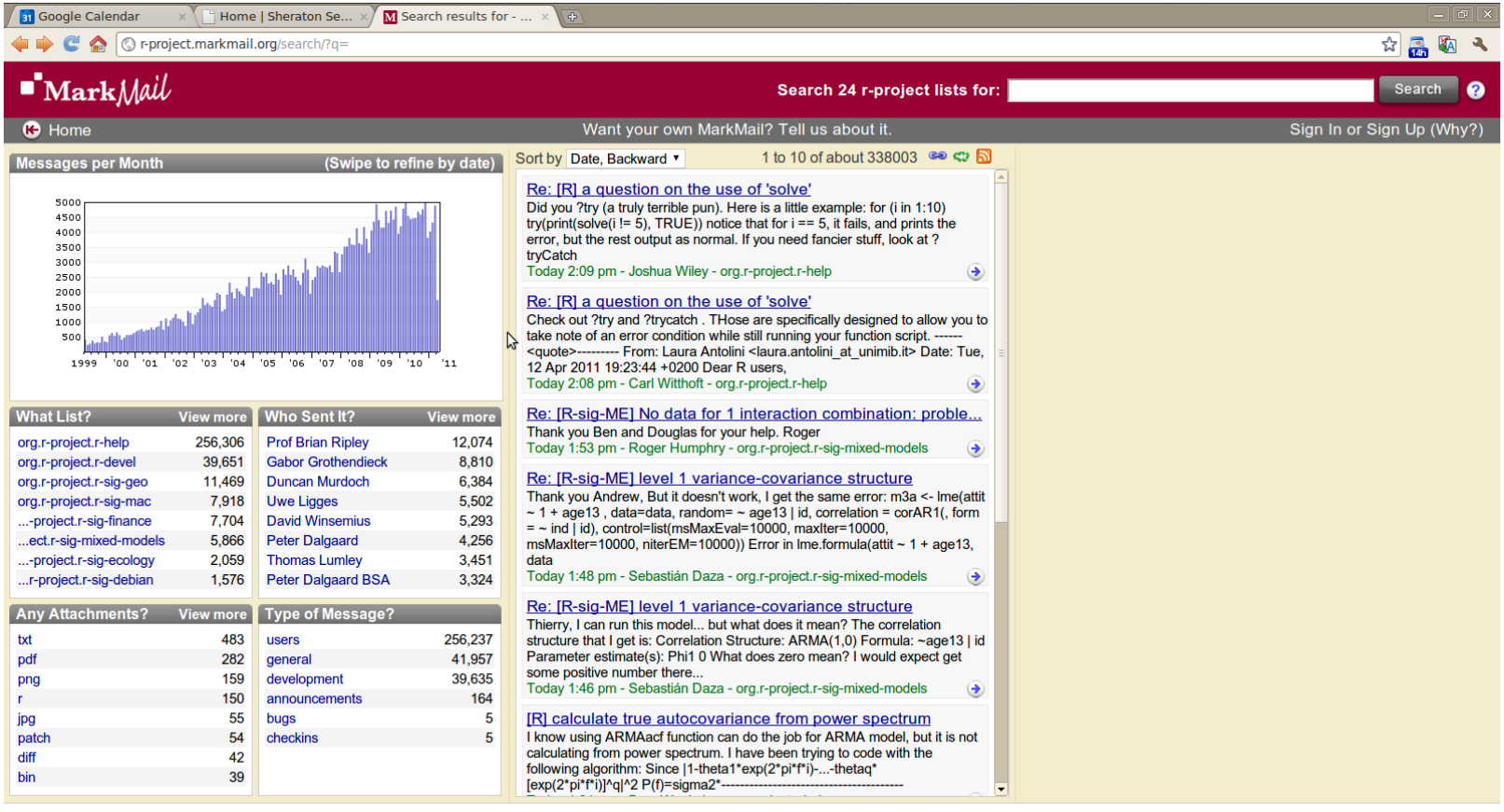
Versioning of research?

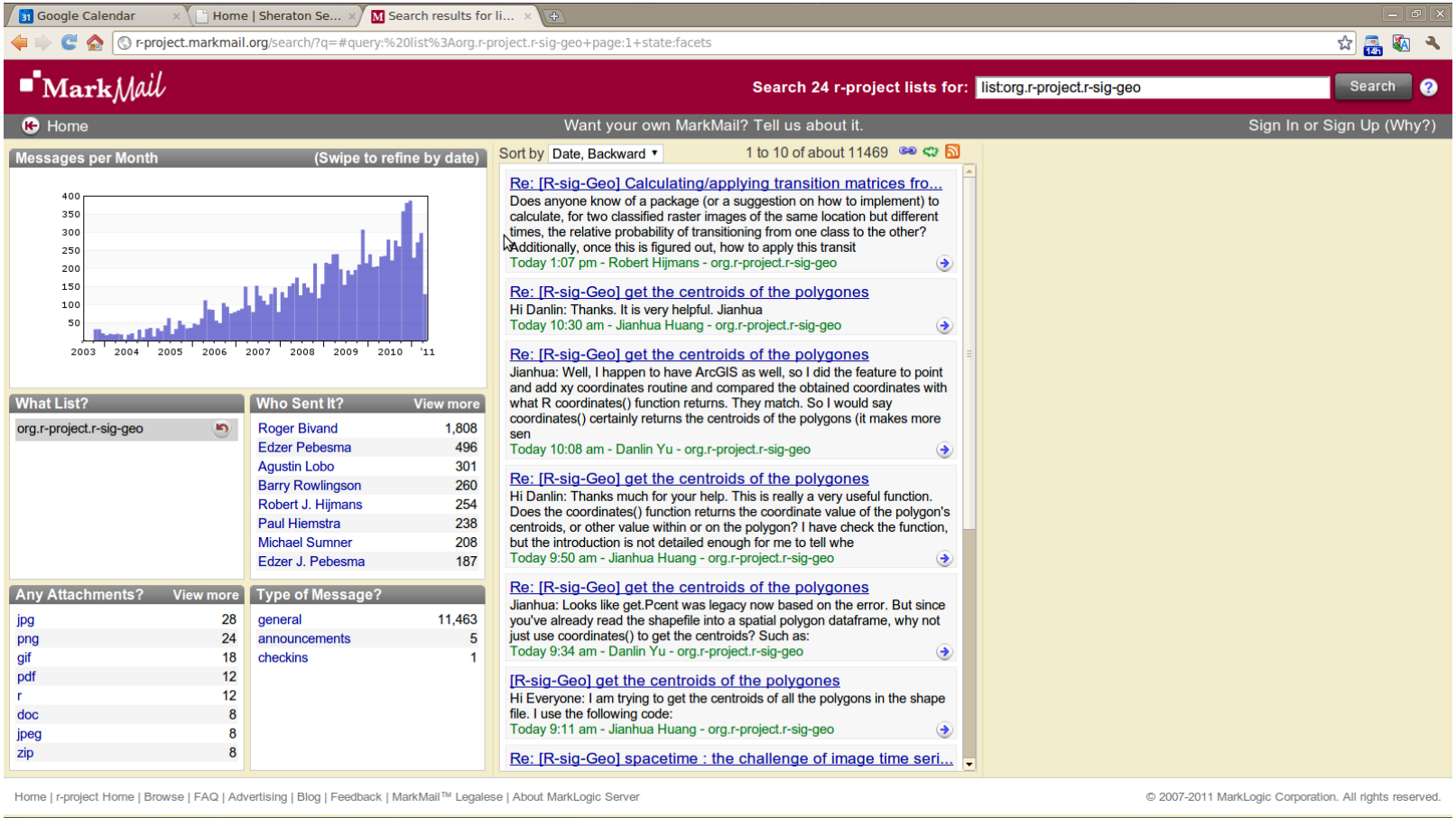
- once accepted, papers are rarely revised
- data and analysis scripts are, over time, typically improved
- to keep the link between submitted paper and the research to reproduce it, data and scripts should be equally frozen, and be part of the submission procedure, and downloadable with the paper
- the author can provide (documented) updates of procedures.

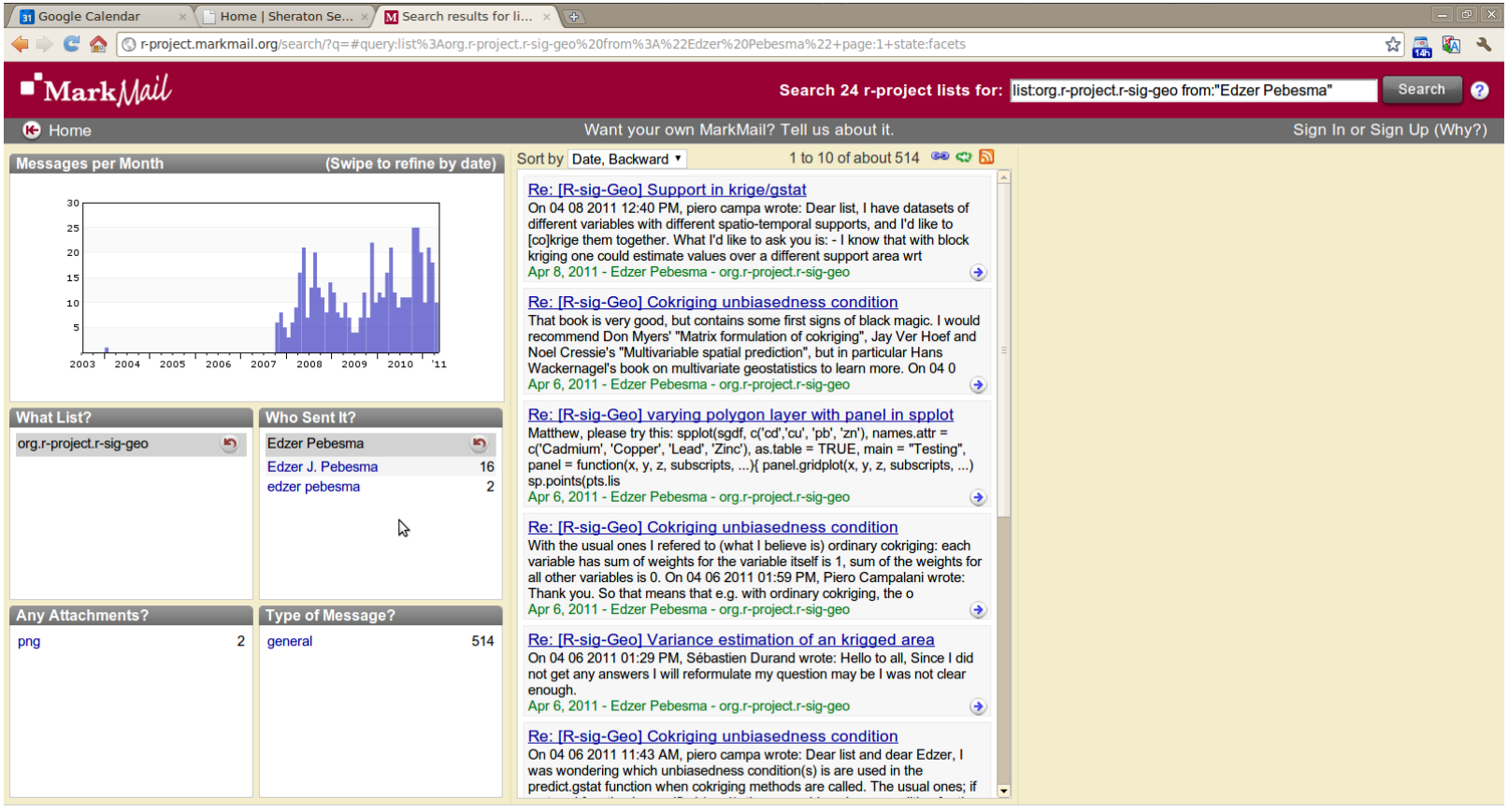
What is R? Why R?

- R is a free, open source environment for statistical computation and graphics, running on all operating systems
- R is developed and maintained by about 20 PhD/professors in academia
- R has an unknown number of users[2]
- R can be extended by add-on packages
- around 4000 of such packages are part of R, and are developed and maintained by a similar number of developers
- last year, R entered the top-20 most used programming languages
- increasingly, R is the platform of choice for teaching and research, both in academia and industry

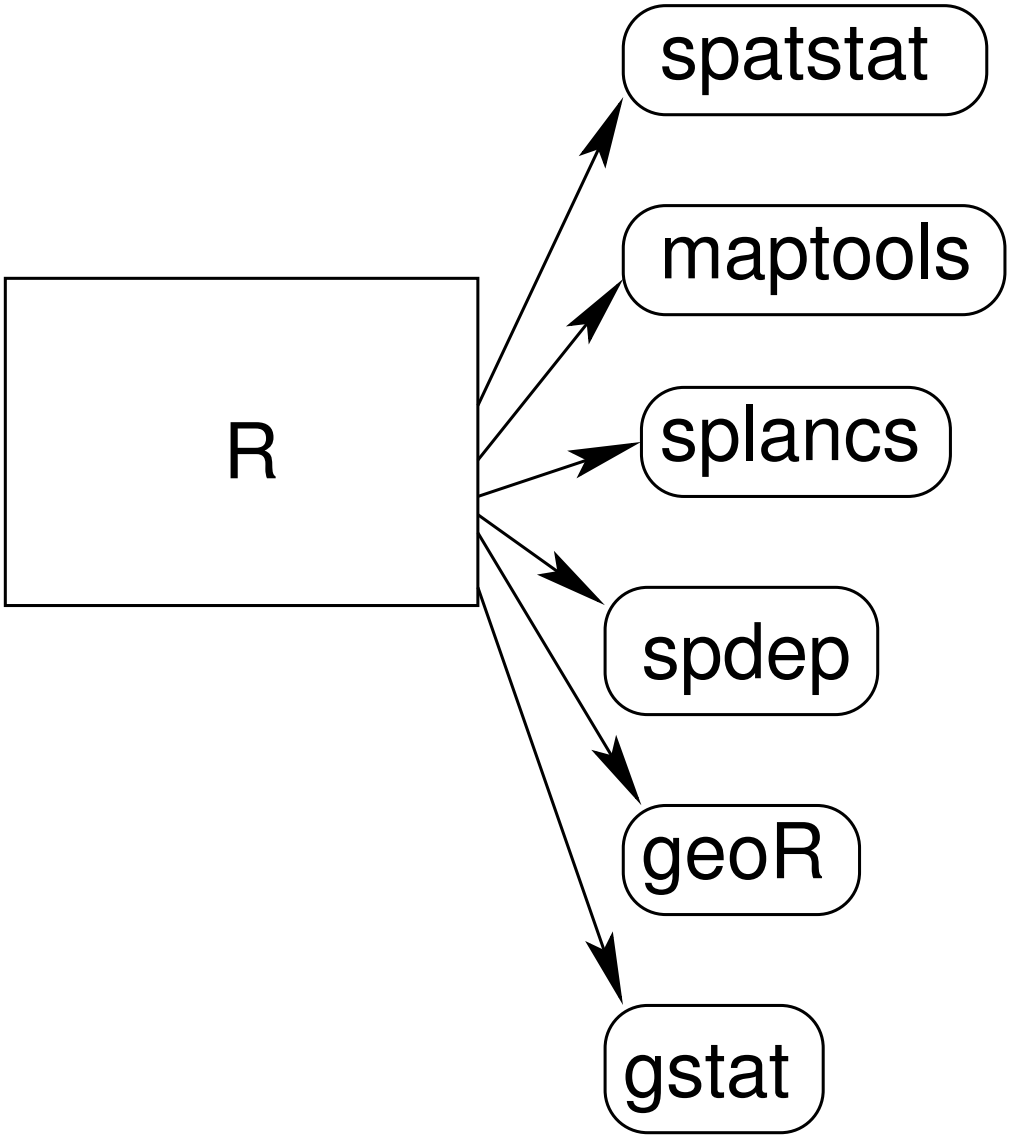
[2] Forbes (2011) estimated it to be 2 Million


Google Calendar
Home | Sheraton Se...
Search results for - ...
r-project.markmail.org/search/?q=
MarkMail
Search 24 r-project lists for:
Search
Home
Want your own MarkMail? Tell us about it.
Sign In or Sign Up (Why?)
Messages per Month
(Swipe to refine by date)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
1999 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11
What List?
View more
org.r-project.r-help 256,306
org.r-project.r-devel 39,651
org.r-project.r-sig-geo 11,469
org.r-project.r-sig-mac 7,918
...-project.r-sig-finance 7,704
...ect.r-sig-mixed-models 5,866
...-project.r-sig-ecology 2,059
...r-project.r-sig-debian 1,576
Who Sent It?
View more
Prof Brian Ripley 12,074
Gabor Grothendieck 8,810
Duncan Murdoch 6,384
Uwe Ligges 5,502
David Winsemius 5,293
Peter Dalgaard 4,256
Thomas Lumley 3,451
Peter Dalgaard BSA 3,324
Any Attachments?
View more
txt 483
pdf 282
png 159
r 150
jpg 55
patch 54
diff 42
bin 39
Type of Message?
users 256,237
general 41,957
development 39,635
announcements 164
bugs 5
checkins 5
Sort by Date, Backward
1 to 10 of about 338003
Re: [R] a question on the use of 'solve'
Did you ?try (a truly terrible pun). Here is a little example: for (i in 1:10) try(print(solve(i != 5), TRUE)) notice that for i == 5, it fails, and prints the error, but the rest output as normal. If you need fancier stuff, look at ? tryCatch
Today 2:09 pm - Joshua Wiley - org.r-project.r-help
Re: [R] a question on the use of 'solve'
Check out ?try and ?trycatch . THose are specifically designed to allow you to take note of an error condition while still running your function script. ------ <quote>--------- From: Laura Antolini <laura.antolini_at_unimib.it> Date: Tue, 12 Apr 2011 19:23:44 +0200 Dear R users,
Today 2:08 pm - Carl Witthoft - org.r-project.r-help
Re: [R-sig-ME] No data for 1 interaction combination: proble...
Thank you Ben and Douglas for your help. Roger
Today 1:53 pm - Roger Humphry - org.r-project.r-sig-mixed-models
Re: [R-sig-ME] level 1 variance-covariance structure
Thank you Andrew, But it doesn't work, I get the same error: m3a <- lme(attit ~ 1 + age13 , data=data, random= ~ age13 | id, correlation = corAR1(, form = ~ ind | id), control=list(msMaxEval=10000, maxIter=10000, msMaxIter=10000, niterEM=10000)) Error in lme.formula(attit ~ 1 + age13, data
Today 1:48 pm - Sebastián Daza - org.r-project.r-sig-mixed-models
Re: [R-sig-ME] level 1 variance-covariance structure
Thierry, I can run this model... but what does it mean? The correlation structure that I get is: Correlation Structure: ARMA(1,0) Formula: ~age13 | id Parameter estimate(s): Phi1 0 What does zero mean? I would expect get some positive number there...
Today 1:46 pm - Sebastián Daza - org.r-project.r-sig-mixed-models
[R] calculate true autocovariance from power spectrum
I know using ARMAacf function can do the job for ARMA model, but it is not calculating from power spectrum. I have been trying to code with the following algorithm: Since |1-theta1*exp(2*pi*f*i)-...-thetaq* [exp(2*pi*f*i)]^q|^2 P(f)=sigma2*----------
Home | r-project Home | Browse | FAQ | Advertising | Blog | Feedback | MarkMail™ Legalese | About MarkLogic Server
© 2007-2011 MarkLogic Corporation. All rights reserved.

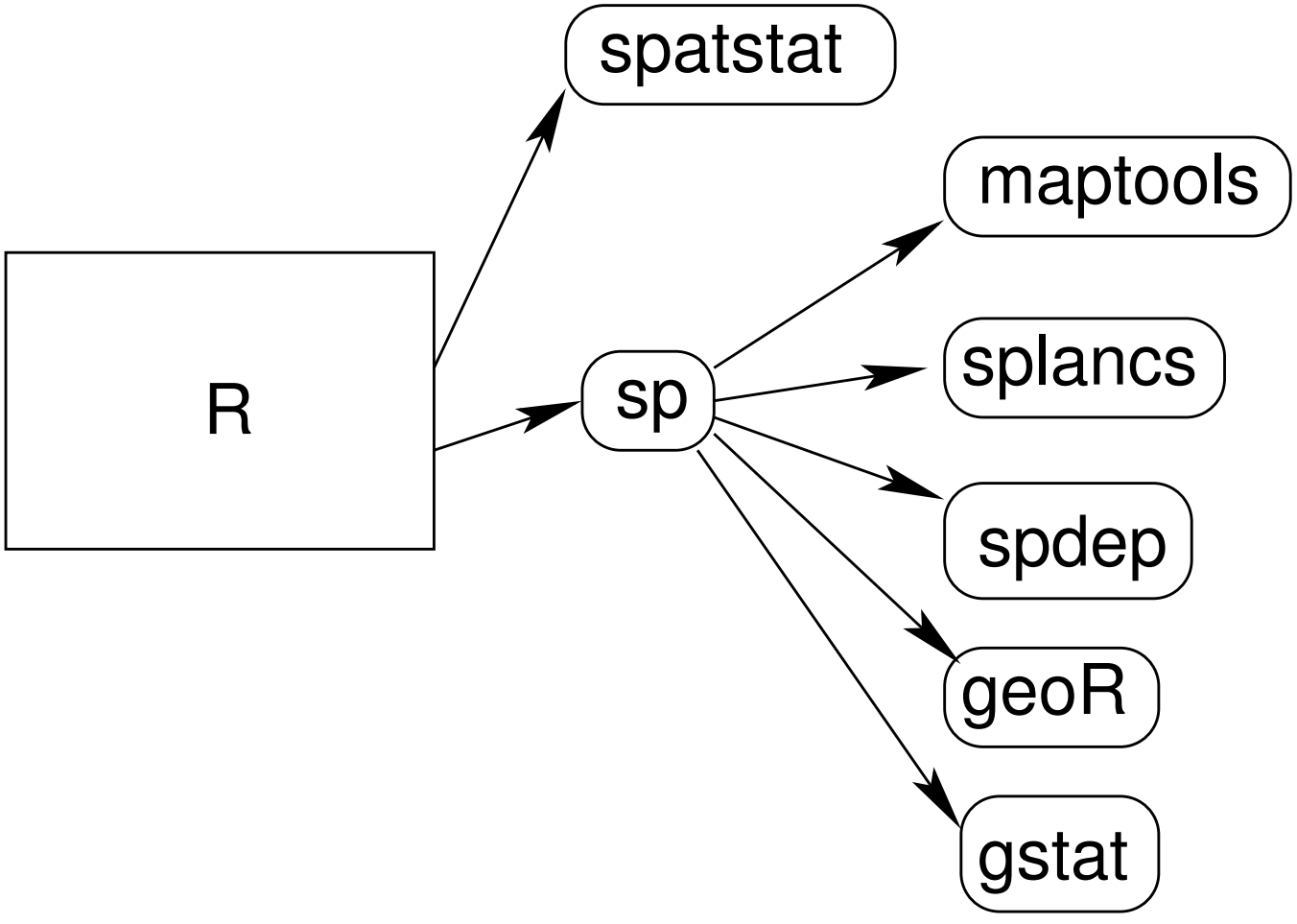

Google Calendar
Home | Sheraton Se...
Search results for li...
r-project.markmail.org/search/?q=#query:%20list%3Aorg.r-project.r-sig-geo+page:1+state:facets
MarkMail
Search 24 r-project lists for:
list:org.r-project.r-sig-geo
Search
Home
Want your own MarkMail? Tell us about it.
Sign In or Sign Up (Why?)
Messages per Month
(Swipe to refine by date)
400
350
300
250
200
150
100
50
2003
2004
2005
2006
2007
2008
2009
2010
'11
What List?
org.r-project.r-sig-geo
Who Sent It?
View more
Roger Bivand 1,808
Edzer Pebesma 496
Agustin Lobo 301
Barry Rowlingson 260
Robert J. Hijmans 254
Paul Hiemstra 238
Michael Sumner 208
Edzer J. Pebesma 187
Any Attachments?
View more
jpg 28
png 24
gif 18
pdf 12
r 12
doc 8
jpeg 8
zip 8
Type of Message?
general 11,463
announcements 5
checkins 1
Sort by Date, Backward
1 to 10 of about 11469
Re: [R-sig-Geo] Calculating/applying transition matrices fro...
Does anyone know of a package (or a suggestion on how to implement) to calculate, for two classified raster images of the same location but different times, the relative probability of transitioning from one class to the other? Additionally, once this is figured out, how to apply this transit
Today 1:07 pm - Robert Hijmans - org.r-project.r-sig-geo
Re: [R-sig-Geo] get the centroids of the polygones
Hi Danlin: Thanks. It is very helpful. Jianhua
Today 10:30 am - Jianhua Huang - org.r-project.r-sig-geo
Re: [R-sig-Geo] get the centroids of the polygones
Jianhua: Well, I happen to have ArcGIS as well, so I did the feature to point and add xy coordinates routine and compared the obtained coordinates with what R coordinates() function returns. They match. So I would say coordinates() certainly returns the centroids of the polygons (it makes more sen
Today 10:08 am - Danlin Yu - org.r-project.r-sig-geo
Re: [R-sig-Geo] get the centroids of the polygones
Hi Danlin: Thanks much for your help. This is really a very useful function. Does the coordinates() function returns the coordinate value of the polygon's centroids, or other value within or on the polygon? I have check the function, but the introduction is not detailed enough for me to tell whe
Today 9:50 am - Jianhua Huang - org.r-project.r-sig-geo
Re: [R-sig-Geo] get the centroids of the polygones
Jianhua: Looks like get.Pcent was legacy now based on the error. But since you've already read the shapefile into a spatial polygon dataframe, why not just use coordinates() to get the centroids? Such as:
Today 9:34 am - Danlin Yu - org.r-project.r-sig-geo
[R-sig-Geo] get the centroids of the polygones
Hi Everyone: I am trying to get the centroids of all the polygons in the shape file. I use the following code:
Today 9:11 am - Jianhua Huang - org.r-project.r-sig-geo
Re: [R-sig-Geo] spacetime : the challenge of image time seri...
Home | r-project Home | Browse | FAQ | Advertising | Blog | Feedback | MarkMail™ Legalese | About MarkLogic Server
© 2007-2011 MarkLogic Corporation. All rights reserved.


Google Calendar
Home | Sheraton Se...
Search results for li...
r-project.markmail.org/search/?q=#query:list%3Aorg.r-project.r-sig-geo%20from%3A%22Edzer%20Pebesma%22+page:1+state:facets
MarkMail
Search 24 r-project lists for:
list:org.r-project.r-sig-geo from:"Edzer Pebesma"
Search
Home
Want your own MarkMail? Tell us about it.
Sign In or Sign Up (Why?)
Messages per Month
(Swipe to refine by date)
30
25
20
15
10
5
2003
2004
2005
2006
2007
2008
2009
2010
'11
What List?
org.r-project.r-sig-geo
Who Sent It?
Edzer Pebesma
Edzer J. Pebesma 16
edzer pebesma 2
Any Attachments?
png 2
Type of Message?
general 514
Sort by Date, Backward
1 to 10 of about 514
Re: [R-sig-Geo] Support in krige/gstat
On 04 08 2011 12:40 PM, piero campa wrote: Dear list, I have datasets of different variables with different spatio-temporal supports, and I'd like to [co]krige them together. What I'd like to ask you is: - I know that with block kriging one could estimate values over a different support area wrt
Apr 8, 2011 - Edzer Pebesma - org.r-project.r-sig-geo
Re: [R-sig-Geo] Cokriging unbiasedness condition
That book is very good, but contains some first signs of black magic. I would recommend Don Myers' "Matrix formulation of cokriging", Jay Ver Hoef and Noel Cressie's "Multivariable spatial prediction", but in particular Hans Wackernagel's book on multivariate geostatistics to learn more. On 04 0
Apr 6, 2011 - Edzer Pebesma - org.r-project.r-sig-geo
Re: [R-sig-Geo] varying polygon layer with panel in spplot
Matthew, please try this: spplot(sgdf, c('cd','cu', 'pb', 'zn'), names.attr = c('Cadmium', 'Copper', 'Lead', 'Zinc'), as.table = TRUE, main = "Testing", panel = function(x, y, z, subscripts, ...){ panel.gridplot(x, y, z, subscripts, ...) sp.points(pts.lis
Apr 6, 2011 - Edzer Pebesma - org.r-project.r-sig-geo
Re: [R-sig-Geo] Cokriging unbiasedness condition
With the usual ones I refered to (what I believe is) ordinary cokriging: each variable has sum of weights for the variable itself is 1, sum of the weights for all other variables is 0. On 04 06 2011 01:59 PM, Piero Campalani wrote: Thank you. So that means that e.g. with ordinary cokriging, the o
Apr 6, 2011 - Edzer Pebesma - org.r-project.r-sig-geo
Re: [R-sig-Geo] Variance estimation of an krigged area
On 04 06 2011 01:29 PM, Sébastien Durand wrote: Hello to all, Since I did not get any answers I will reformulate my question may be I was not clear enough.
Apr 6, 2011 - Edzer Pebesma - org.r-project.r-sig-geo
Re: [R-sig-Geo] Cokriging unbiasedness condition
On 04 06 2011 11:43 AM, piero campa wrote: Dear list and dear Edzer, I was wondering which unbiasedness condition(s) is are used in the predict.gstat function when cokriging methods are called. The usual ones; if
Home | r-project Home | Browse | FAQ | Advertising | Blog | Feedback | MarkMail™ Legalese | About MarkLogic Server
© 2007-2011 MarkLogic Corporation. All rights reserved.

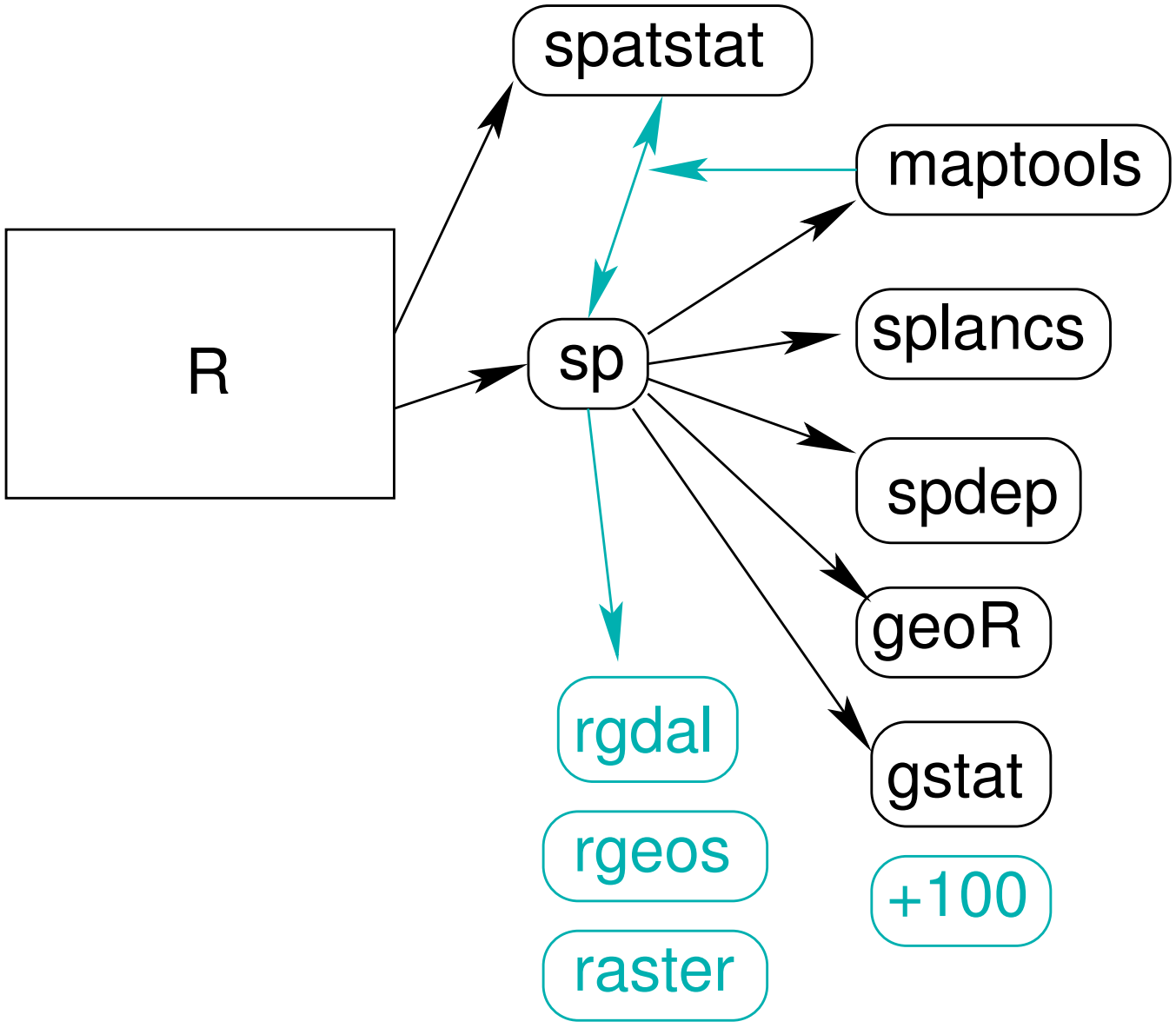
Before 2005



After 2005



2011: over 100 spatial packages on CRAN



How can R be used for reproducible geographical research?

Stage I: share data and scripts

- share and discuss research ideas, as a basic attitude, by scripts that need three mouse clicks maximally to reproduce (this is basically what `r-help` and `r-sig-geo` require)
- provide, in your paper, reference to the software used
- write, in your paper, *that* you are willing to share (data and) scripts needed to reproduce the research
- share these scripts through a web site (*)
- provide the URL in your paper (*)
- submit data and procedures (scripts) as supplementary material

(*) this puts responsibility on the author's side

How can R be used for reproducible geographical research?

Stage II: the executable paper

- inspired by Donald Knuth's *literate programming*, R has since long had `Sweave`, for literate analysis.
- `Sweave` documents mix text (the journal paper), and R scripts.
- `Sweave` runs R, and automatically merges ("weaves") the output (text, figures) needed to generate the final document (e.g. pdf, html)
- R code is *replaced* or *augmented* with the output of running R
- many R books are being written in this system, including "Applied Spatial Data Analysis with R" (Bivand, Pebesma, Gomez-Rubio)
- Sweave papers are *executable* papers

Challenges

- when submitting an executable paper, how will the journal publisher react when we submit 15 years of Landsat imagery for the Amazon basin?
- if historic data is improved over time, can we still acess the version of the data on which a particular outcome was based?
- how does versioning of data carry through in the linked data cloud?
- scripts work with particular versions of software (R, add-on packages). Where do we find working instances of the software as it was used to execute the paper in 2002?
- What are the responsibilities of journal publisher, and what are those of the author?

Outlook

- we need to convince journals that reproducible papers are (i) useful, (ii) reproducible, and (iii) better than non-reproducible ones.
- we ought to start rejecting papers without accompanying readable and reproducible procedures
- we probably need more disasters, climate-gates, etc., before this will fly
- we need to adopt open science, and reproducibility, as the default case, not as an exception.

Further reading

- Friedrich Leisch, Manuel Eugster, Torsten Hothorn, 2011. Executable Papers for the R Community: The R2 Platform for Reproducible Research. International Conference on Computational Science, ICCS 2011; *Procedia Computer Science* 4 (2011) 618–626.
- Pebesma, Nüst, Bivand, 2012. R in reproducible geoscientific research. *EOS*, accepted for publication.