A practical history of R-sig-geo (where things came from)

Roger Bivand

20 August 2018

The presentation files are on github at: rsbivand/geostat18

- Understanding the original uses of S and R (in the 1980s and 1990s), and seeing how these uses affected the development of R lets us appreciate the robustness of R's ecosystem.

- This talk uses readings of the R sources and other information to explore R's history. The topics to be touched on include the "colour" books (brown, blue, white, green), interlinkages to SICP (Scheme) and LispStat

- This should show where and when impulses flowed between R and the development of capabilities in R for spatial data analysis

- The focus will be on the period 1997-2004, which is when a lot of the foundations for R spatial came into being

History of R and its data structures

- Rasmus Bååth has a useful blog piece on R's antecedents in the S language

- Something similar is present in the second chapter of (Chambers, 2016), from the viewpoint of one of those responsible for the development of the S language

- In addition to S, we need to take SICP and Scheme into account (Abelson and Sussman, 1996, second edition), as described by Ihaka and Gentleman (1996) and Wickham (2014)

- Finally, LispStat and its creators have played and continue to play a major role in developing R (Tierney, 1990, 1996, 2005)

R did not always look like an alternative implementation of the S language. It started as a small Scheme-like interpreter (loosely based on work by Sam Kamin [4] and David Betz [2]). This provided a platform for experimentation and extension. The following dialog shows a very early version of the R interpreter at work.

```
> (define square (lambda (x) (* x x)))
square
> (define v 10)
v
> (square v)
100
```

The S-like appearance of R was added incrementally. We initially moved to an S-like syntax for our experiments because, although we were both familiar with Lisp syntax, we didn't feel that it was really suitable for expressing statistical computations. This choice of syntax set us on a path towards compatibility with S. Once initiated, the move towards compatibility with S was irresistible. This was partly because we wanted to see just how far we could push it and partly because it have us access to code resources and developers.

6

Becker and Chambers (1984): S: An Interactive Environment for Data Analysis and Graphics, A.K.A. the Brown Book
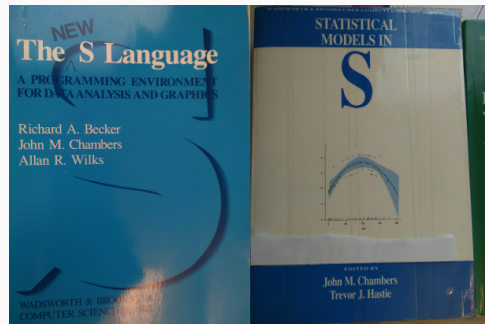
Becker and Chambers (1985): Extending the S System
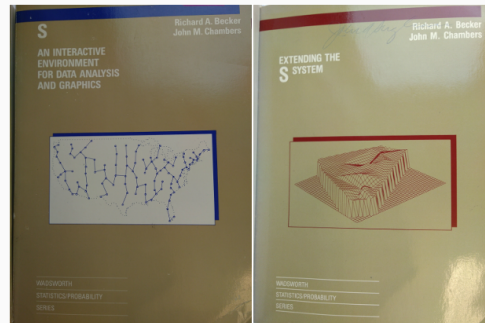
Becker et al. (1988): The New S Language: A Programming Environment for Data Analysis and Graphics, A.K.A. the Blue Book.
Chambers and Hastie (1992): Statistical Models in S, A.K.A. the White Book.

Chambers (1998): Programming with Data: A Guide to the S Language, A.K.A. the Green Book. Venables and Ripley (2000): S Programming

- The S2 system was described in the Brown Book, S3 in the Blue Book and completed in the White Book, finally S4 in the Green Book

- The big advance from S2 to S3 was that users could write functions; that data.frame objects were defined; that formula objects were defined; and that S3 classes and method dispatch appeared

- S4 brought connections and formal S4 classes, the latter seen in R in the **methods** package (still controversial)

- S-PLUS was/is the commercial implementation of S and its releases drove S3 and S4 changes

- S was a Bell Labs innovation, like Unix, C, C++, and many interpreted languages (like AWK); many of these share key understandings

- Now owned by Nokia, previously Alcatel-Lucent, Lucent, and AT&T

- Why would a telecoms major (AT&T) pay for fundamental research in computer science and data analysis (not to sell or market other products better)?

- Some Green Book examples are for quality control of telecoms components

- S-PLUS was quickly adopted for teaching and research, and with S3, provided extensibility in the form of libraries

- Most links have died by now, but see this FAQ for a flavour - there was a lively community of applied statisticians during the 1990s

- S built on a long tradition of documentation through examples, with use cases and data sets taken from the applied statistical literature; this let users compare output with methods descriptions

- ... so we get to R

and what about LispStat?

- Luke Tierney was in R core in 1997, and has continued to exert clear influence over development

- Because R uses a Scheme engine, similar to Lisp, under the hood, his insight into issues like the garbage collector, namespaces, byte-compilation, serialization, parallelization, and now ALTREP has been crucial (see also the proposal by Luke Tierney, Gabe Becker and Tomas Kalibera)

- Many of these issues involve the defensive copy on possible change policy involved in lazy evaluation, which may lead to multiple redundant copies of data being present in memory

- Luke Tierney and Brian Ripley have fought hard to let R load fast, something that is crucial to ease the use of R on multicore systems or inside databases

R spatial and R-sig-geo

- The AI-GEOSTATS listserve/mailing list was started by Gregoire Dubois in 1995; AI was from Arc/Info

- GRASS accepted code contributions, then kept in `src.contrib` and `src.garden` and GRASS mailing lists were active (the archives seem to have been lost in moving from minordomo to mailman in 2001)

- Bao et al. (2000) describe S-PLUS links to GRASS and ArcView; the module for spatial statistics was also from the mid-1990's

- `MASS` included S and compiled code for kriging, trend surfaces and point pattern analysis, `splancs` provided point pattern analysis

- Much of the porting of S code to R for spatial statistics was begun by Albrecht Gebhardt

- The CRAN archives show `tripack` and `akima` - both with ACM licenses - from August 1998; `ash` and `sgeostat` followed in April 1999

- The date for the MASS `spatial` package is unknown because it was bundled in the VR metapackage subsequently split into `spatial`, MASS, `nnet` and `class`

- In the earliest period, CRAN admins helped practically with porting

- Rowlingson and Diggle (1993) describes the S-PLUS version of `splancs` for point pattern analysis

- I'd contacted Barry in 1997 but only moved forward with porting `splancs` as R's ability to load shared objects advanced

- In September 1998, I wrote to him:

```
It wasn't at all difficult to get things running, which I think is
a result of your coding - thank you! An issue I have thought about
a little is whether at some stage Albrecht and I wouldn't integrate
or harmonize the points and pairs objects in Splancs, spatial, and
Sgeostat - they aren't the same, but for users maybe ought to appear
to be so. Maybe when I get the GRASS interface up?
```

- The earliest `splancs` in the CRAN archive is from November 2000

- Albrecht and I were teaching spatial data analysis using R and code that would go into CRAN packages (I still have the package I wrote, and the exams)

- In 1998, the ERSA (European Regional Science Association) meeting was in Vienna, and we gave a talk subsequently published in Bivand and Gebhardt (2000)

- We covered R, point pattern analysis, geostatistics, ESDA and spatial econometrics

tering. It would be of advantage if both Splancs and the spatial package defined point objects in the same way, or at least if conversion functions were written, permitting data to be moved between them without uncertainty about the bounding region being used; both initially date from a period when object-oriented mechanisms were only beginning to enter S.

· Markus Neteler and I were in touch too, as I'd been trying to interface R and GRASS (Bivand, 2000)

```
I just uploaded GRASS-0.1-1.tar.gz - a standard R package, and
paper2.pdf.gz - the latest version of the C&G submission, to
hgeo02.geog.uni-hannover.de. The package was tested on 5.0b5 on the GRASS
side, and 0.99.0 (released 7/2) on the R side. In fact, the package caused
a patch to be issued for R yesterday - I was overstressing a strsplit()
function in R in parsing the output of g.list! The same two files are on
ftp://reclus.nhh.no/pub/candg.
```

· The first GRASS package in the CRAN archive is from November 2000

· Before that, we'd presented at GeoComputation 2000 in Chatham (Bivand and Neteler, 2000)

- Nicholas Lewin-Koh posted on the R list (subsequently split between R-help and R-devel) in September 2000, asking about coding the reading of ESRI shapefiles

- He made a number of key contributions which ended up in `spweights` - currently part of `spdep` - and what became `maptools`

- He wrote the first vector representation based on the shapefile import format, and I should be blamed for subsequent fumbling implementations in `maptools`

- From interaction with Kurt Hornik, Ross Ihaka and Brian Ripley on the R list during late 2000, contacts widened, including Tim Keitt (`rgdal`)

- I ended up presenting at the DSC meeting in Vienna in March 2001, getting to meet many R core members, developers and CRAN admins

15:15-15:40	Coffee Break	Coffee Break	
	Applications	**R Connectivity**	
	Anna Bartkowiak	Roger Bivand	
	A Set of XLispStat Subroutines for Detecting Outliers	R and geographical information systems, especially GRASS	
15:40-17:20	Balasubramanian Narasimhan	Erich Neuwirth	
	A Java-based test bed for Random Number Generators Chen Xun	Embedding R in standard software, and the other way round	
	A Quasi-Exact Method for the Confidence Intervals for the Difference of Two Independent Binomial Proportions in Small Sample Cases	Erich Neuwirth R Sings or using R to sonify data	
	Martin Theus	Brian Ripley	
	What Dataminers Want	Applications of R clients and servers	

- Further contacts came from a CSISS workshop organised by Luc Anselin in May 2002 in Santa Barbara, including Konstantin Krivoruchko (ESRI) and Gilberto Câmara (INPE)

- My talk eventually appeared as Bivand (2006), but the contacts were more useful

- Visiting the INPE earth observation lab and presenting at a conference in December 2002 provided further ideas and links to TerraLib

- Luc himself contributed code to spdep and continues be very supportive

- I did a further piece for the ERSA 2002 conference in Dortmund (Bivand, 2002)

- The DSC meeting in March 2003 was also being planned for Vienna, and I'd talked about it with Kurt Hornik when he gave a keynote at the GRASS meeting in Trento in September 2002

- I met Virgilio Gómez-Rubio at the GRASS meeting among many others and talked about `DCluster` and `RArcInfo`

10.30-12.30	**Keynote speeches**
	Chairman Battista Benciolini, University of Trento, Italy.

Current status and future development of GRASS
Markus Neteler, ITC-irst, Italy.

GRASS as a multidimensional dynamic GIS: past vision and where we are now
Helena Mitasova, Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, USA.

The dynamics/development of a free software project in statistical computing: R
Kurt Hornik, Vienna University of Economics and Business Administration, Austria.

- In mid November I started emailing contacts to invite sumbissions for a spatial statistics session at DSC, and a workshop at which we could discuss working together

- Of the replies, far more would have liked to attend than were able to

Friday 21 10:30-12:30	**Spatial Statistics**
	C. Furlanello, M. Neteler, S. Merler, S. Menegon, S. Fontanari, A. Donini, A. Rizzoli and C. Chemini *GIS and the Random Forest Predictor: Integration in R for Tick-Borne Disease Risk Assessment*
	V. Gómez Rubio, J. Ferrándiz and A. López *Detecting Disease Clusters with R*
	Barry Rowlingson, Adrian Baddeley, Rolf Turner and Peter Diggle *Rasp: A Package for Spatial Statistics*
	Virgilio Gómez Rubio *RArcInfo: Using G.I.S. Data with R*
15:30-17:30	Edzer J. Pebesma *Gstat: Multivariable Geostatistics for S*
	Paulo J. Ribeiro, Ole F. Christensen and Peter J. Diggle *Geostatistical Software -- geoR and geoRglm*
	Albrecht Gebhardt *PVM-Kriging with R*
	E. Glatzer and R. Müller *An Interactive Tool for Residual Diagnostics in Variogram Fitting (with Implementation in R)*
	Links: Open pre-DSC'03 spatial data workshop Back to Overview

Open pre-DSC'03 spatial data workshop: R, map(), GIS, ...

Wednesday 19 March 2003, Department of Economic Geography and Geoinformatics, Vienna University of Economics and Business Administration

Location: Room RL 1.16, Department of Economic Geography and Geoinformatics, Vienna University of Economics and Business Administration, Rossauer L�nde 23/1, A-1090 Vienna; access by crossing road from U4 Metro Station Rossauer L�nde. The Department is at the red ring in the centre of this map.

Times: we expect to run to the following schedule: 0930-1030, coffee, 1100-1230, lunch, 1400-1530, coffee, 1600-end, where "end" dovetails with the DSC icebreaker.

Who: this open workshop is "open" - anyone with an interest in what we will be discussing is welcome to attend, whether you are an active contributor or not.

Aims: the open workshop is intended to permit as many as possible of DSC participants interested in the practical handling of spatial/geographical data to exchange views. During the day, we should be able to hear about various forms of R/GIS integration, both loose and tight coupled. Topics here include packages already on CRAN (RArcInfo, GRASS, and the spatial statistics packages), and initiatives under development, whether active or not (especially Rmap, the R/GDAL package, functions to read and write shapefiles, PostGIS, GRASS 5.1, TerraLib, using StatConnector in ArcGIS, others?).

It would be good if we can exchange and dig up enough information to try to avoid too much duplication of code and work, and to think through some of the issues involved in working with spatial data in particular in the R setting. In some cases, R is a compute engine embedded in the application, in others the spatial data is ingested into R - glancing at Omegahat might also be relevant as a source of ideas.

Facilities: there will be plenty of DHCP network points, power cables (local sockets!) and a beamer all day. If you would like to join lunch at Serviten St�berl, please let me know by 5 March so that we can book a room there. We will make arrangements if you have particular requirements.

Programme: Please let me know if you would like to present or demonstrate your software for linking R and GIS, and/or for accessing and handling spatial data in R. Since some participants are present for only part of the workshop, please indicate your time constraints. This paragraph will be updated to reflect status.

Contact: Roger Bivand, Economic Geography Section, Department of Economics, Norwegian School of Economics and Business Administration, Breiviksveien 40, N-5045 Bergen, Norway.

From Edzer's reply 13 November 2002

I wonder whether I should start writing S classes. I'm afraid I should.

Correspondence developed quickly (21 November)

```
> > > I wonder whether I should start writing S classes. I'm afraid I should.
> >
> > Definitely, and probably new-style - have you got Venables & Ripley S
> > programming? Plenty is happening on classes, and they are among the things
> > it would be great to reach a consensus on at the little pre-meeting in
> > Vienna (or by correspondence, but brainstorming many-to-many may be better
> > than multiple one-on-one?
> >
> I more or less worked through the green book, by now.

Tim Keitt's interface to GDAL uses new-style classes, so does pixmap now.
I think it will be hard but worthwhile to get a shared set of spatial data
classes built, making it easier for everything on top of them, although I
don't see this as being fun!
```

Contribution even though he couldn't get to Vienna:

```
I was looking over all the DSC material, especially the spatial stuff. I
did notice, after looking through peoples packages that there is a lot
of duplication of effort (I am sure you are aware of this). My
suggestion is that we set up a repository for spatial packages similar
to the bio-conductor model. Where we have a base spatial package that
has S-4 based methods and classes that are efficient and general, Say
Spatial_base. Then a Maps, PointProc, LatticeProc, and GeoProc and
whatever else packages could be built on top of the base class. i would
```

```
Cc: statsgrass@grass.itc.it
Subject: Re: [STATSGRASS] proj4

On Thu, 8 May 2003, Agustin Lobo wrote:

>
> Hi!
>
> Does anyone know if there is a package to
> call proj4 from R (or python)?
> (proj4: library to make coordinate conversions
> betwee different projections, http://remotesensing.org/proj/)
>

I've been working on one, based on Barry Rowlingson's code from his draft
Rmap package. How soon do you need it, and on what platform? I'm trying to
write class objects for coordinates, projection arguments, etc., so any
early release will be subject to change.
```

```
Date: Thu, 22 May 2003 17:55:56 +0200
From: mailman-admin@stat.math.ethz.ch
To: maechler@stat.math.ethz.ch
Subject: Your new mailing list: r-sig-geo

The mailing list `r-sig-geo' has just been created for you.  The
following is some basic information about your mailing list.
```

- Albrecht Gebhardt was a local organiser for a meeting at which many of those at the DSC workshop met in October 2003

○ **Geostatistics: Theory and New Methods**

(Methods of statistical learning theory, Bayes methods, spatial prediction and risk analysis, spatial sampling and monitoring network design etc)

Conveners: W. G. Müller, J. Pilz

○ **Combining Statistics and GIS**

(Interfaces between (geo-)statistics and GIS, integration of geostatistics and GIS, mutual benefits, statistical problems of error propagation and uncertainty in GIS, etc)

Conveners: R. Bivand, G. Dubois

○ **Geostatistical Applications**

(Application of geostatistical methods in the Earth and Environmental Sciences, in Agriculture and Forestry, Epidemiology and Health Sciences, Econometrics and Telecommunications)

Conveners: H. Glass, M. Kanevsky

○ **Spatial Data Bases and Mapping**

(Interfaces between (geo-)statistical software systems, spatial database management systems and visualization and mapping software systems)

Conveners: V. Gomez, B. Rowlingson

○ **Geostatistical Software Developments**

(New implementations of geostatistical methodology in R, S-Plus, SAS and other statistical software systems)

Conveners: E. Pebesma, P. Ribeiro

| Friday 21
16:45-18:15 | **Spatial Statistics**

Roger Bivand, Edzer Pebesma and Barry Rowlingson
Generic functions for spatial data

Albrecht Gebhardt and Claudia Gebhardt
Bayesian Methods in Geostatistics: Using prior knowledge about trend

Nadine Henkenjohann
spatialreg - A package to fit spatial regression models with isotropic and anisotropic covariance functions

Juergen Pilz and Philipp Pluch and Gunter Spoeck
Taking into account uncertainty in spatial covariance estimation for Bayesian prediction

Back to Overview | Session Format |

CRAN arrivals

- `fields` and `geoR` in mid-2001, `spatstat` in early 2002

- `gstat` in February 2003, `maptools` in August 2003, `rgdal` in November 2003

- `sp` in April 2005 (after a workshop in Lancaster November 2004 and before one in Valencia May 2005)

- `classInt` in March 2006, `spgrass6` in August 2006

Tutorial: Analysing Spatial Data in R

presented by Roger Bivand

Goals:

The workshop will introduce the foundation classes for spatial data provided in the sp package, and show how they make work with spatial data easier. The classes have shared interfaces to common file formats for spatial data, and a range of subsetting, plot, overlay, and sampling methods.

The workshop will let participants needing to import spatial data into R, analyse it in one or more packages, display it in graphical form, and perhaps export it from R, learn how these steps can be made conveniently and robustly. The final section will show how the new class structure makes writing custom interfaces less challenging, making R analysis functions available for a wide range of spatial data formats.

Programme:

1. Representing spatial data in R (30 minutes)
2. Vizualising spatial data in R (30 minutes)
3. Accessing spatial data in R (30 minutes)
4. break
5. Worked examples: geostatistics (30 minutes)
6. Worked examples: disease mapping (30 minutes)
7. Worked examples: customising spatial data classes (30 minutes)

Workshop materials:

pre-workshop notes, scripts, and data sets on Rgeo website
download and install at least the core packages from the Spatial task view.

Post-workshop follow-up on the R-sig-geo mailing list.

Back to Tutorials

useR-2006@R-project.org

Thursday 15 15:00-18:30	**Ecology & Spatial Statistics** (Spotlights: HS 0.5, Forum: Atrium)

Philippe Grosjean, Richard Hillary, Ernesto Jardim, Laurie Kell, Iago Mosqueira, Jan Jaap Poos, Robert Scott and Hunter S. Thompson
Fisheries modelling in R: The FLR (Fisheries Library in R) project

Thomas Petzoldt, Karsten Rinke and Louis Kates
Population ecology modelling with R: A comparison of object oriented approaches

Daniel Doktor
Spatial and statistical modelling of phenological data

T. Laurent, A. Ruiz-Gazen, and C. Thomas-Agnan
GEOXP: An R package for interactive exploratory spatial data analysis

Vojtech Janousek, Vojtech Erban and Colin Farrow
Using the R language for graphical presentation and interpretation of compositional data in mineralogy: Introducing the package GCDkit-Mineral

Rudolf Dutter
Data Analysis System with Graphical Interface

Norbert Solymosi, Andrea Harnos, Jenő Reiczigel and Ferenc Speiser
RpostGIS an R-library for using PostGIS spatial structures and functions

A. Pedro Duarte Silva, Jorge Cadima, Manuel Minhoto and Jorge Orestes Cerdeira
Subselect0.99: Selecting variable subsets in multivariate linear models

Lisbeth Riis and Mikkel Grum
Using R to Reduce Pesticide Usage in the Horticultural Industry

Keiji Osaki
Spatial characteristics of vegetation index map in urban area derived by variogram analysis

Back to Overview | Session Format

35

CRAN packages

- Once S3 permitted extension by writing functions, and packaging functions in libraries, S and R ceased to be monolithic

- In R, a library is where packages are kept, distinguishing between base and recommended packages distributed with R, and contributed packages

- Contributed packages can be installed from CRAN (infrastructure built on CPAN and CTAN for Perl and Tex), Bioconductor, other package repositories, and other sources such as github

- With over 12000 contributed packages, CRAN is central to the R community, but is stressed by dependency issues (CRAN is not run by R core)

- Andrie de Vries Finding clusters of CRAN packages using igraph looked at CRAN package clusters from a page rank graph

- Here we will be choosing the packages listed in the Spatial and SpatioTemporal task views as contained in the **ctv** package, and their pagerank relationships

CRAN package page rank scores

```
##            Rcpp              sp            MASS        maptools        spatstat
##        0.027231        0.021910        0.012782        0.010123        0.009503
##           rgdal           spdep         lattice           knitr          raster
##        0.009236        0.009207        0.009009        0.008963        0.008007
##           rgeos              sf        testthat         plotKML          Matrix
##        0.007885        0.007794        0.007756        0.007095        0.006835
##         ggplot2    RandomFields       rmarkdown     RColorBrewer           gstat
##        0.006824        0.005949        0.005694        0.005418        0.005364
##           spaMM         leaflet       spacetime          stplanr            ade4
##        0.005197        0.005193        0.005174        0.004784        0.004779
##        magrittr           dplyr           Hmisc robCompositions            maps
##        0.004778        0.004754        0.004707        0.004634        0.004550
```

First package cluster

```
##          sp     maptools       rgdal        knitr       raster        rgeos
## 0.021909980  0.010122714  0.009236221  0.008963208  0.008007089  0.007884672
##          sf     testthat      plotKML      ggplot2    rmarkdown RColorBrewer
## 0.007794233  0.007756111  0.007095066  0.006824349  0.005693899  0.005417506
##       gstat      leaflet    spacetime        Hmisc      mapview       stringr
## 0.005364255  0.005193224  0.005173799  0.004707367  0.004458524  0.004377673
##        tmap         plyr
## 0.004241166  0.004045955
```

Second package cluster

```
##          MASS       spatstat         spdep        lattice    RandomFields
##    0.012782095    0.009503133    0.009206800    0.009008741    0.005949009
##          maps            FRK         fields       CARBayes          georob
##    0.004549607    0.004021172    0.003748315    0.003717594    0.003312704
##        marmap            car           geoR         geospt           abind
##    0.003279676    0.002944835    0.002861347    0.002854347    0.002775868
##           VIM           mgcv           spam         PReMiuM   RcppArmadillo
##    0.002623556    0.002465480    0.002398066    0.002364441    0.002206238
```

Third package cluster

```
##         Rcpp      stplanr       dplyr      tibble       rlang          BH
## 0.027231166 0.004783552 0.004754131 0.003553933 0.003351115 0.002891366
##     roxygen2    tmaptools       tidyr   RcppEigen       readr       broom
## 0.002771766 0.002551049 0.002401763 0.001968913 0.001967407 0.001899899
##          rio       haven      lwgeom       purrr       dodgr       later
## 0.001700361 0.001680877 0.001676834 0.001663077 0.001661423 0.001649899
##     promises   tidyselect
## 0.001632014 0.001456890
```

```
##     magrittr     jsonlite         httr   geojsonio     osmdata     geojson
## 0.0047784772 0.0043322548 0.0034418393 0.0031150461 0.0021470149 0.0021417830
##         lawn         xml2           V8       gistr  rmapshaper        curl
## 0.0017566903 0.0017564029 0.0017510706 0.0014674384 0.0014575060 0.0014398722
##      lazyeval  geojsonlint          oai        rvest     leafletR         jqr
## 0.0013994314 0.0012896173 0.0011876625 0.0010895131 0.0010592619 0.0008586266
##     protolite      selectr
## 0.0008464338 0.0007460838
```

Cluster 1: Hmisc, maptools, ggplot2, spacetime, RColorBrewer, sp, stringr, gstat, plotKML, tmap, leaflet, mapview, rgdal, sf, rgeos, testthat, raster, knitr, plyr, rmarkdown

Cluster 2: georob, spdep, mgcv, marmap, spam, fields, MASS, RcppArmadillo, VIM, FRK, car, geoR, RandomFields, maps, geospt, CARBayes, abind, PReMiuM, spatstat, lattice

Rcpp

lwgeom
broom
promises stplanr
rlang
purrr RcppEigen
dplyr read tidyselect
rio later
BH tibble
roxygen2 dodgr
tmaptools
tidyr

magrittr
httr
geojsonio
osmdata
jsonlite
geojson

- R itself is very varied in the way that approaches have established niches through time

- sp classes date back to the same time as Bioconductor, which has a fairly definite S4 preference

- Communities develop and change over time, and continue to do so, with simultaneous differences in preferences

- so what comes after **tidyverse** and "data science"? Should we be preparing already?