Quo Vadis?

Bill Venables, CSIRO, Australia

UseR! 2012

Nashville

15 June, 2012

Contents

1 Outline

- Introduction and motivation
- Examples of **R** in action: "elementary **R** from an advanced standpoint".
- What are people doing with **R**, now? Some reasons for concern?
- Strengths, Weaknesses, Opportunities and Threats.

2 Introduction and genesis

- Recent interest by the programing language community, (Cook, 2012; Morandat et al., 2012)

  Do they have a point?

- Alternating emphasis in **S** (or nearly):
  - 1984–5: *interactive*: (Becker and Chambers, 1984, 1985)
  - 1988 *programming*, **S** *3*: (Becker et al., 1988)
  - 1991 *statistics*: (Chambers and Hastie, 1991)
  - 1998 *data*, *programming*, **S** *4*: (Chambers, 1998),
  - 2008 *software*, *data*: (Chambers, 2008).
- Should the two sides be more explicitly separated?

New York, 23 March, 1999...The Association for Computing Machinery (ACM) today named Dr John M. Chambers of Bell Labs as the recipient of the 1998 Software System Award for developing the **S** System, an innovative software program[a] that helps users to manage and extract useful information from data.

The ACM's citation notes that Dr Chambers' work "will forever alter ***the way people analyse, visualize, and manipulate data*** ... **S** is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers."

The System Software Award recognizes those who develop software systems having a lasting influence. It will be presented on 15 May, 1999 during a special ACM awards banquet in New York City, and will be accompanied by a $10,000 prize.

---

[a] environment?

Why is R so popular, really?

Why has **R** attracted such a wide following? John Chambers, (Chambers, 2009), suggested 6 *facets*

1. an interface to computational procedures of many kinds;
2. interactive, hands-on in real time;
3. functional in its model of programming;
4. object-oriented, "everything is an object";
5. modular, built from standardized pieces; and,
6. collaborative, a world-wide, open-source effort.

To this list I would add that specifically the **R** system is:

7. extensible, may be augmented by compiled code in other languages;
8. cross-platform; and
9. international.

Perhaps 7 is just an amplification of 1 (interface)

(Oh, and it's free.)

Just what does "interactive" mean, too?

Allows the user:

- To analyse data quickly, easily and comprehensively,
- To synthesise the information in data about the process they were collected to explore
- To *keep track* of the exercise in a *reproducible* way

i.e. to "program with data".

Two cultures, or two aspects of the same culture?

- *Users* of **R** to do real data analysis and graphics work, now, quickly, effectively (and reproducibly).

  "Statistics work is *detective* work!" – Tukey (1969)

  (Typically users, and lovers, of **S** 3 classes and methods. . .)

- *Developers* using **R** to write "*understandable* and *trustworthy* software", (Chambers, 2008).

  (Typically users, and sometimes lovers, of **S** 4 classes and methods. . .)

These are *not* separate: most users will operate somewhere on a spectrum at any one time, moving often.

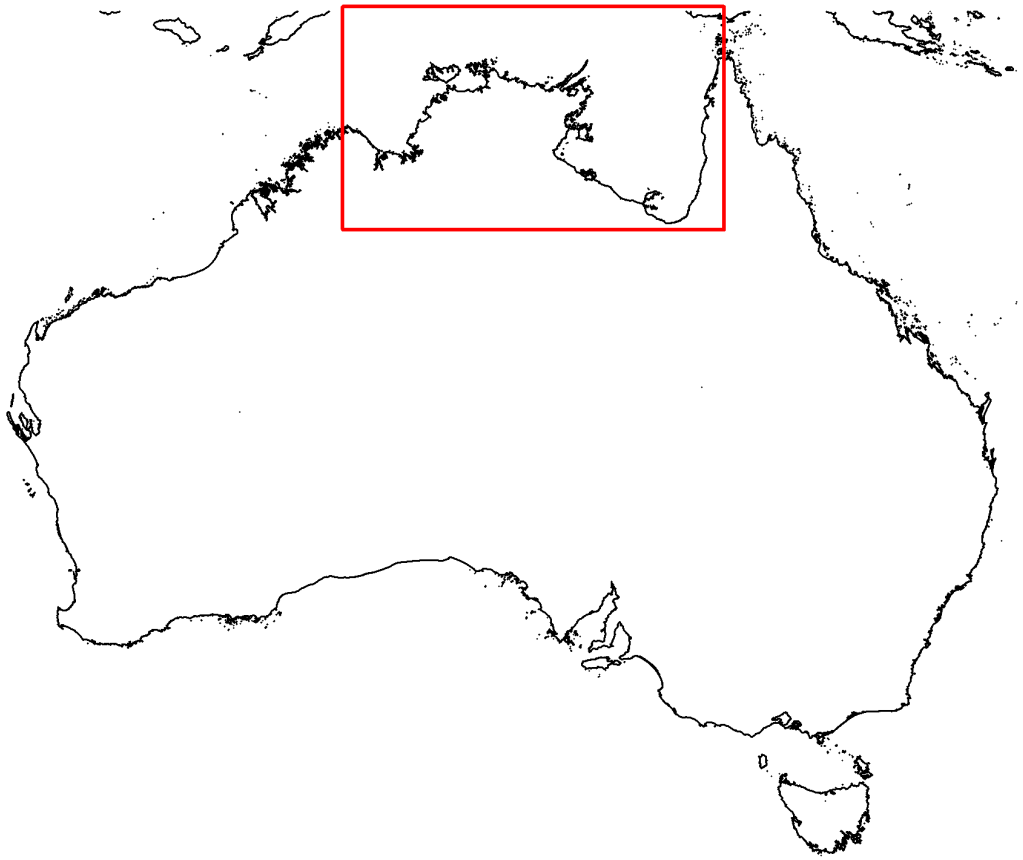
> "Can one be a good data analyst without being a half-good programmer?
> The short answer to that is, 'No'.
> The long answer to that is, '*No!*'."
>
> —*Frank Harrell*
>
> 1999 **S-PLUS** User Conference, New Orleans

- **R** balances the *current*, *partly conflicting* needs of data analysts and software developers in a near-optimal way (for now).
  - interactive flexibility *vs* code efficiency
  - informal *vs* formal programming

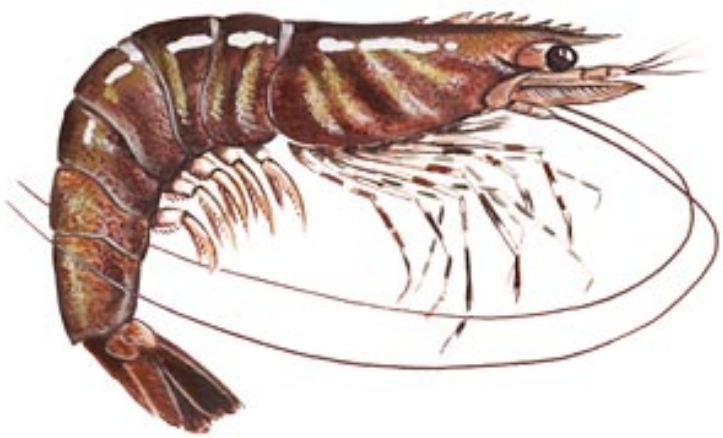
3 The Northern Prawn Fishery

- Mixed species prawn (= shrimp) fishery in Northern Australia
- Four species groups: *Tigers*, *Bananas*, *Endeavours*, *Kings*, each a composite of two *biological* species.
- Tiger prawns (*Penaeus semisulcatus* 'Grooved', and *P. esculentis*, 'Brown') are the most valuable, and currently the only species which have a stock assessment.
- Vessel daily logbook records provide only a single weight of Tiger prawns caught: the two species are not differentiated.
- Problem: build a predictive model to "split" the Tiger catch into the two component species catches, by weight.

Where is it and how big is it?

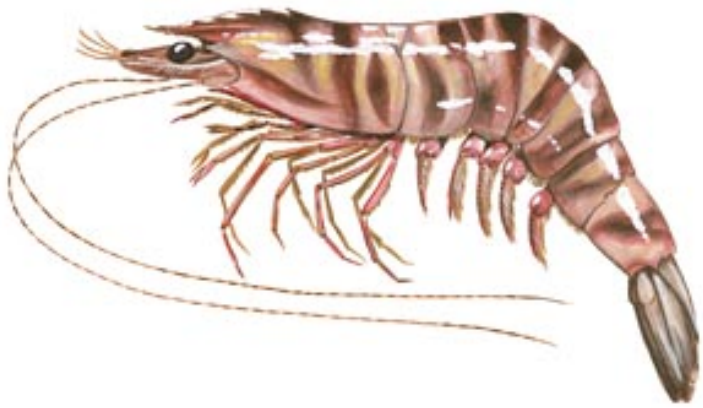


What do the prawns look like?



Penaeus semisulcatus

Grooved tiger prawn



Penaeus esculentus

Brown tiger prawn

The fishing map, 1970-2011 effort

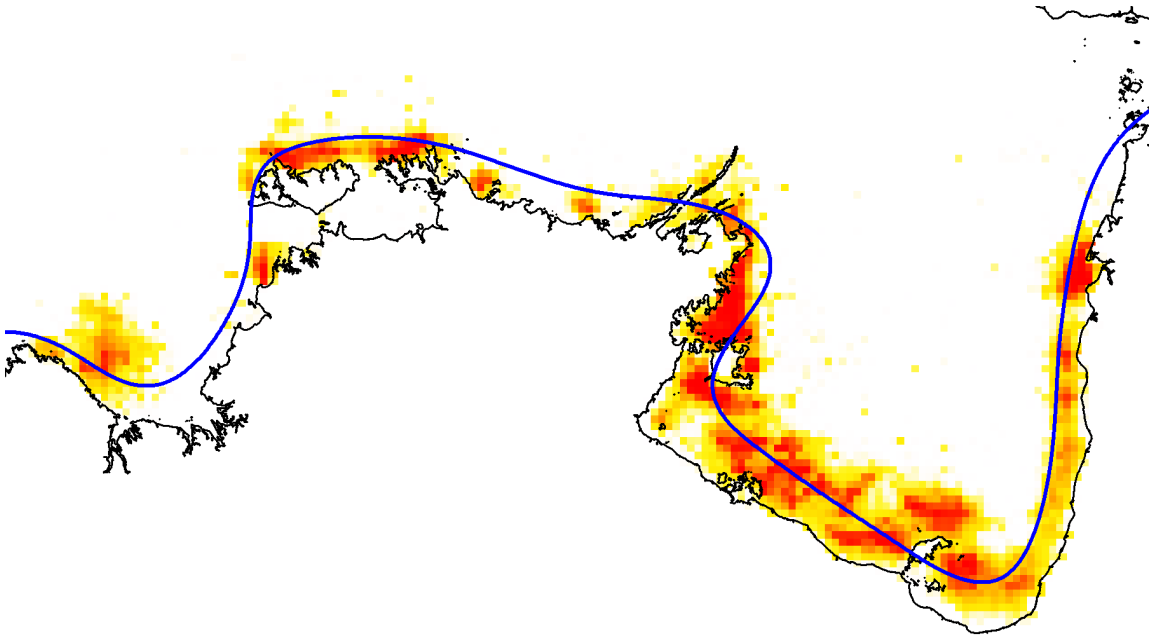


Figure 1: A spatial representation of NPF effort since 1970, in boat days (on a log scale). The deeper the orange colour, the higher the recorded effort.

Available data

- Response: 12 scientific surveys where the catch is separated by weight for every "shot" (12 surveys, $\approx$ 12000 shots)
- Predictors:
  - Spatial location, `Latitude, Longitude, Rdist, Rland`
  - Depth, `Depth`
  - Benthic substrate: `Mud` (not initially)
  - Time: time of year, `PDay`, Time since 1970-01-01 `Day`

Models

Three phases:

- Species split, 1990: Static model based on empirical data.
  (Before my time.)
- Species Split, 2000: Generalized linear model, spatio-temporal,
  (My **S-PLUS** days, verging towards **R**)
- Species Split, 2004: Generalized additive model, spatio-temporal.
  (**R** conversion complete!)

Species split, 2000

(An account of this work is given in Venables and Dichmont (2004).)

10 surveys, ≈9000 records.

The response was $Y = S/T$ where

- S = Weight of *P. semisulcatus* (Grooved Tiger) and
- T = Total weight of Tiger prawns in the catch.

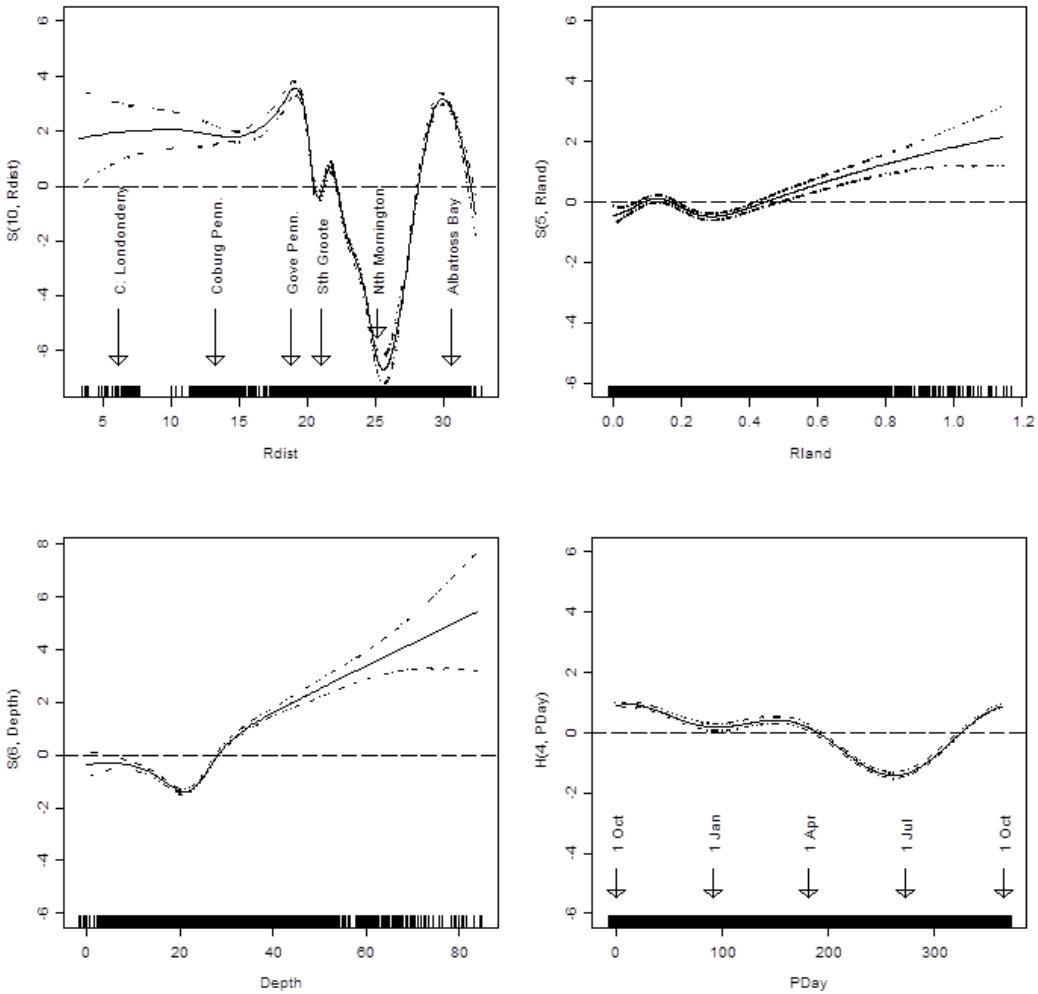
with all weights in grams.

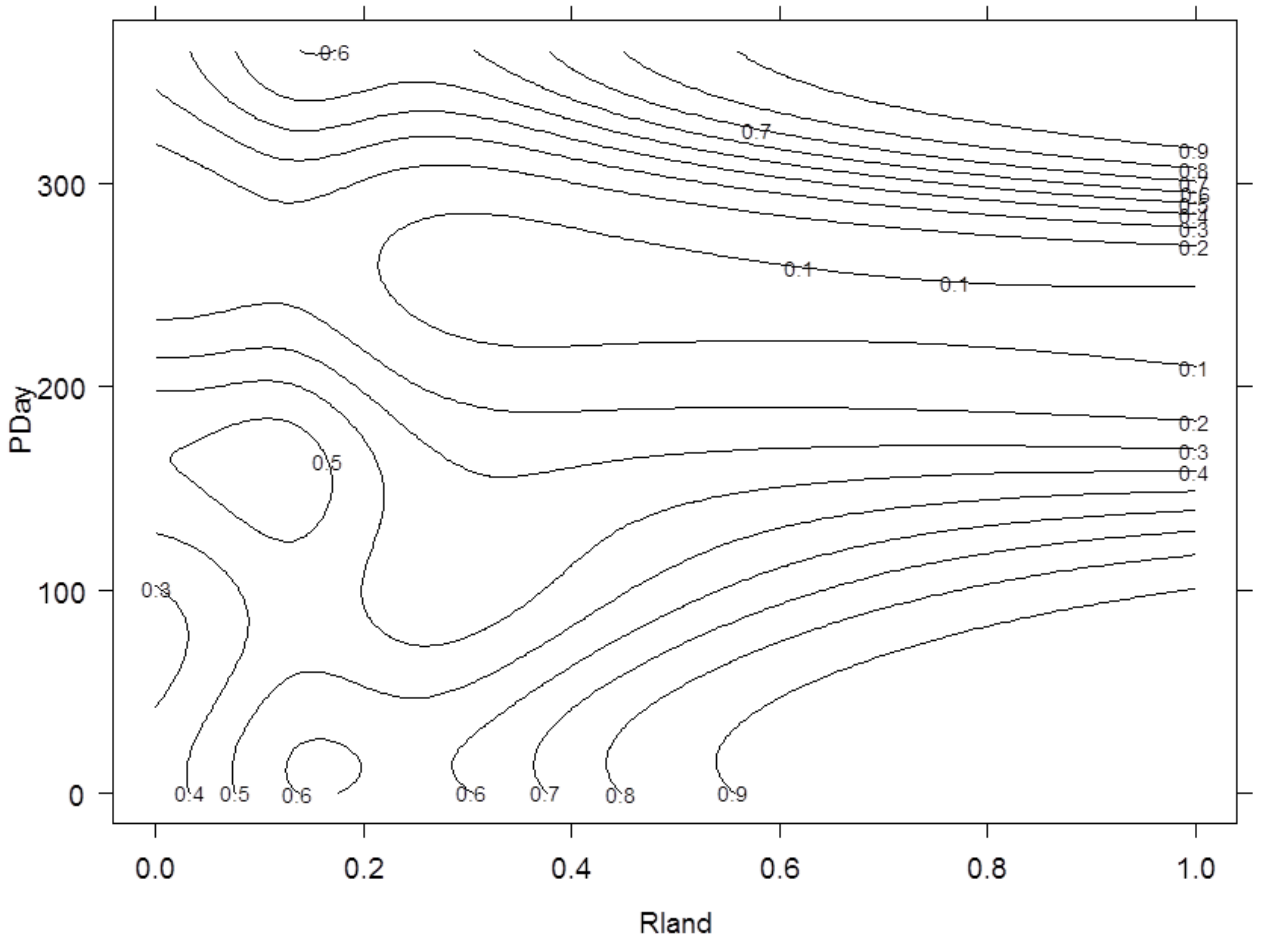
If $\mathrm{E}\, Y = \mu$ then we used a `quasibinomial` GLM with

$$\text{logit}\,\mu = \beta_0 + S_{10}(\text{Rdist}) + S_5(\text{Rland}) + S_6(\text{Depth}) + H_4(\text{PDay}) + LH_2(\text{Rland}, \text{PDay})$$

where the $S_k()$-terms are natural splines with k degrees of freedom, $H_k()$ is an harmonic (Fourier polynomial) with k degrees of freedom and $LH_2(,)$ is a linear $\times$ harmonic interaction with 2 d.f.

Components:


S(10, Rdist)
C. Londonderry
Coburg Penn.
Gove Penn.
Sth Groote
Nth Mornington
Albatross Bay
Rdist
S(5, Rland)
Rland
S(6, Depth)
Depth
H(4, PDay)
1 Oct
1 Jan
1 Apr
1 Jul
1 Oct
PDay


PDay
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
Rland
0.6
0.7
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.1
0.2
0.3
0.4
0.5
0.3
0.4
0.5
0.6
0.6
0.7
0.8
0.9

A long-term trend?

To investigate if the model really were temporally stable, we included an extra term $S_5(\text{Day})$, i.e. a natural spline with 5 d.f. in the elapsed time, in days, since 1970-01-01.

The result was ambiguous:

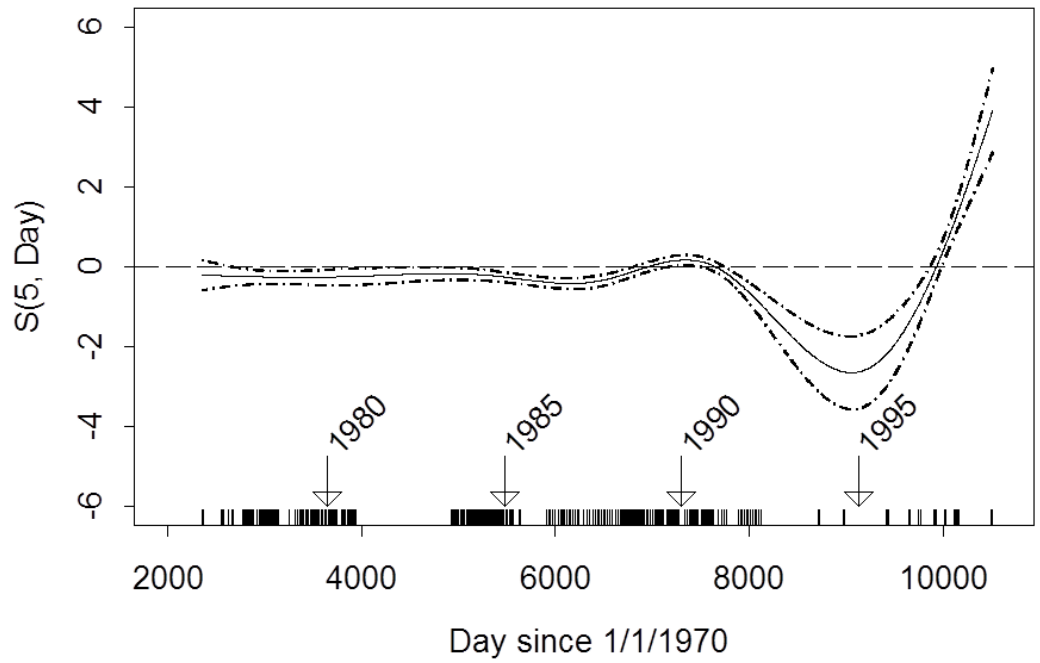


Figure 2: Long-term trend component, additional to the temporally stable model of 2000.

Species split, 2004

- A new project with field data collection: led by the statistician
- A new technology had become available via the `mgcv` package, (Wood, 2003, 2004, 2011).

In the new model, the response was the same as in the previous one. It was a quasibinomial GAM, with terms as follows:

- An *isotropic* thin-plate smooth spline term in `Longitude` and `Latitude`, intended to capture spatial aspects not otherwise captured by functions of location,

- A bivariate smooth tensor spline in `Day` (of year) and (distance out to) `Sea`, with the spline basis for `Day` cyclic, with period one year,
- A similar bivariate smooth tensor spline in `Day` and `Depth`,
- A bivariate smooth tensor spline in `Sea` and `Depth` to capture more subtle spatial features,
- A smooth spline term in (percent) `Mud` in the sediment.

The non-stable variant of the model contained in addition:

- A smooth spline term in the `ElapsedDays` since 1970-01-01.

The fitted components are shown in Figures 3 and 4.

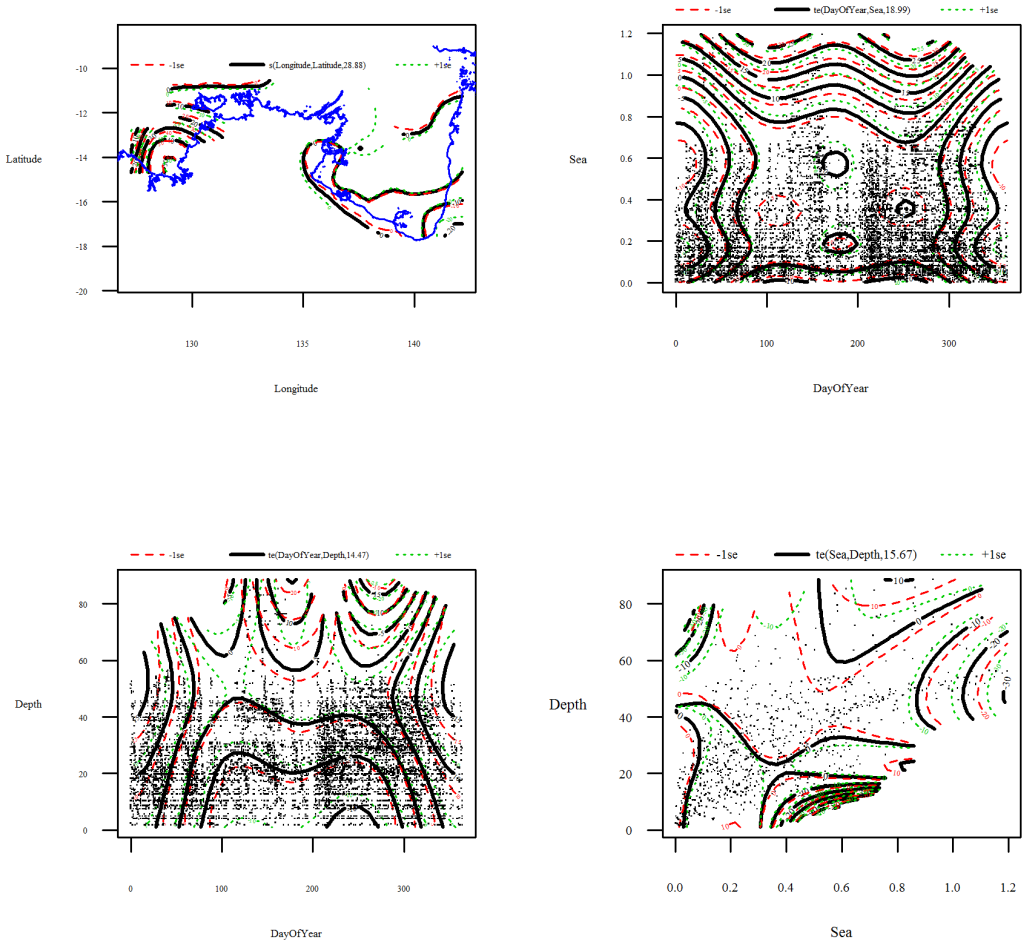


Figure 3: Four smooth bivariate components from the stable model, 2004.

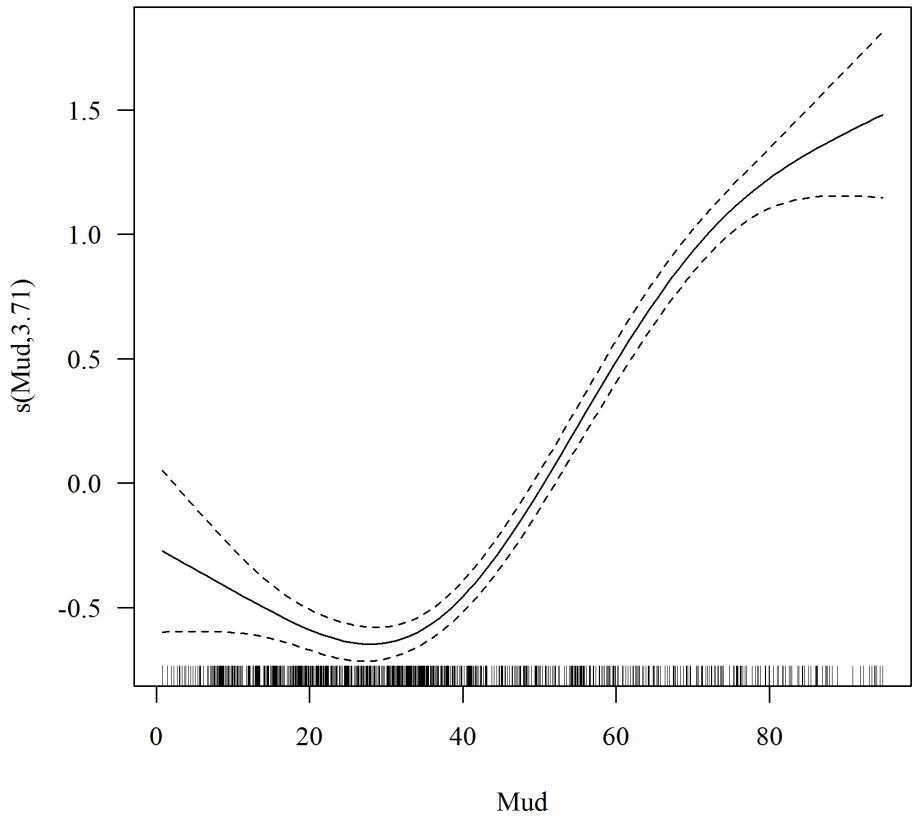


Figure 4: The sediment component of the stable model, 2004.

The annual migration effect

To appreciate the annual migration effect, we can predict the change in proportion over the year in four key places, shown in Figure 5.

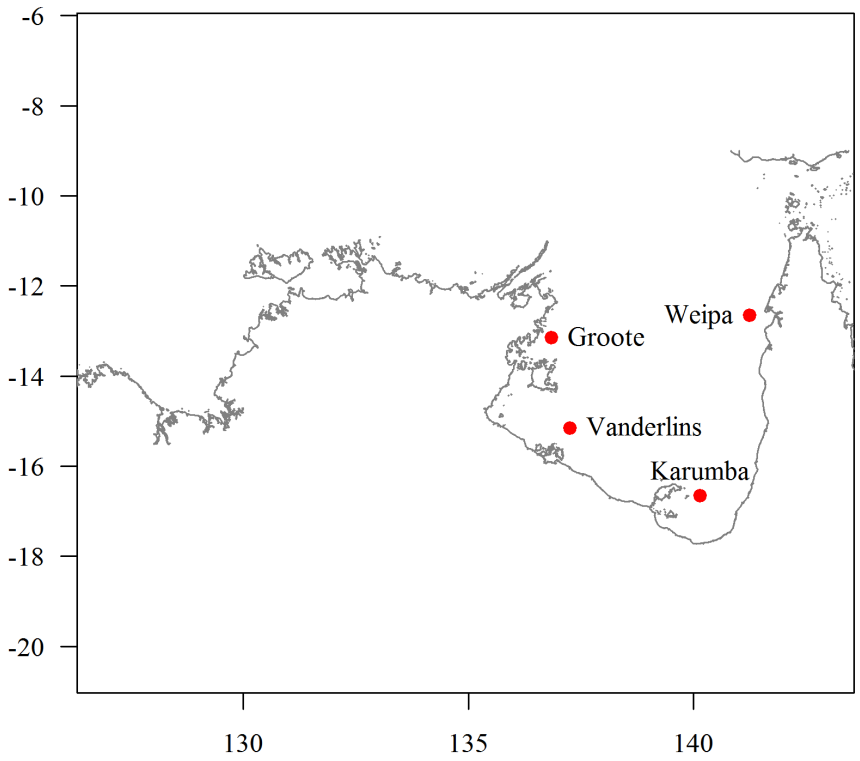


Figure 5: Four key places in the NPF

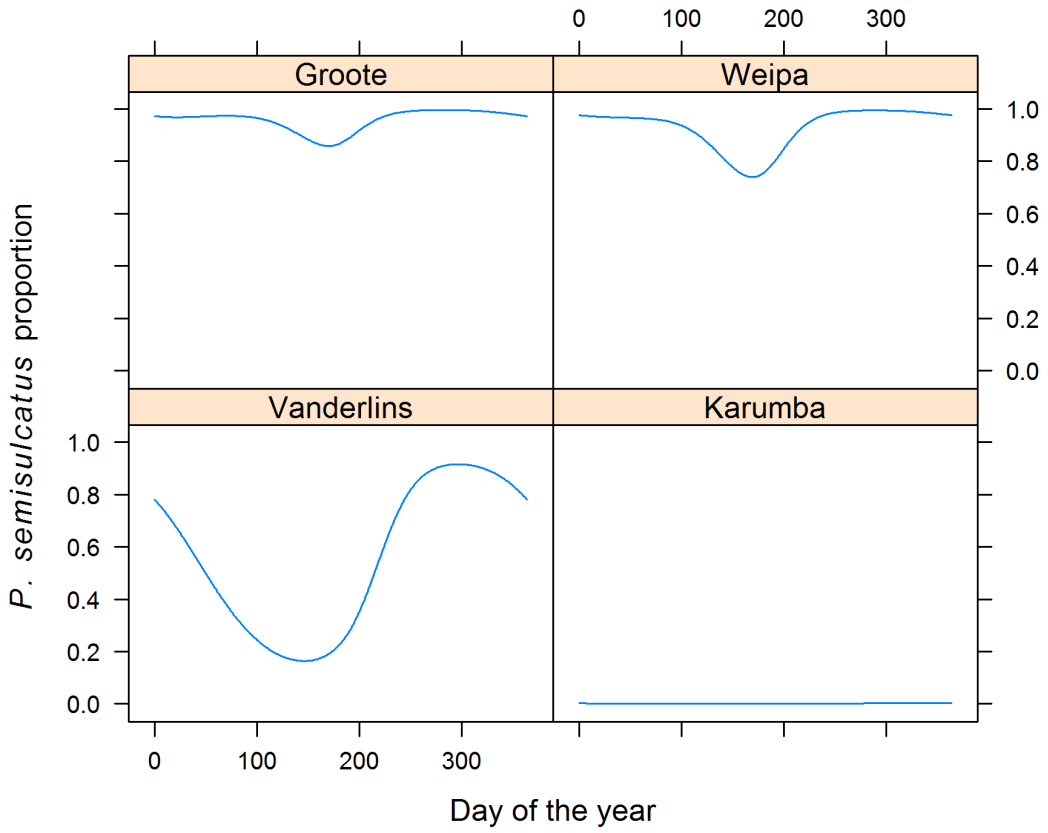
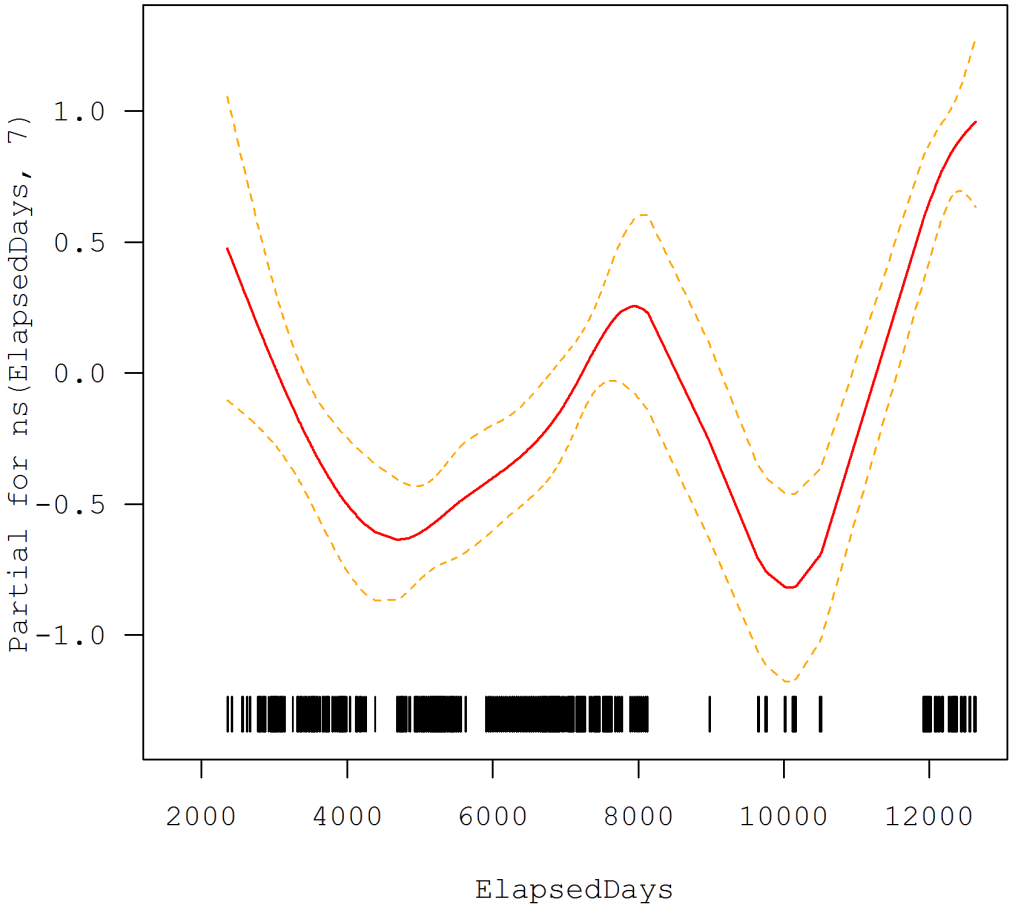


Figure 6: Variation in *P. semisulcatus* proportions over the year

The long-term trend component



A technical note on complex in R

All Euclidean distances between points in two dimensions:

```
> dist2d <- function(x, y=NULL) {
    z <- with(xy.coords(x, y), complex(real = x, imaginary = y))
    as.dist(outer(z, z, function(x, y) Mod(x-y)))
  }
```

Drawing thin blue line, "by hand". With the coastline displayed:

```
> curvCoord <- with(locator(type = "o"), {
    z0 <- complex(real = x, imaginary = y)
    d0 <- c(0, cumsum(Mod(diff(z0))))
    z <- complex(real      = spline(d0, Re(z0), n=2000)$y,
                 imaginary = spline(d0, Im(z0), n=2000)$y)
    data.frame(Longitude = Re(z), Latitude = Im(z),
               Coast = c(0, cumsum(Mod(diff(z)))))
  })
> lines(Latitude ~ Longitude, curvCoord, col="blue")
```

Some packages

- `splines`, (Doug Bates - now part of standard **R**)
- `gam`, (Hastie, 2011),
- `glm2`, (Marschner, 2011)
- `SearchTrees`, (Becker, 2012)
- `mgcv`, (Wood, 2006)

4 The morphometric problem

Problem: Calibrate the relationship between weight and carapace length of an animal.

Standard model form: $W = \alpha L^{\beta}$ (usually $2 < \beta < 3$)

Statistical model:

$$W \sim \mathrm{N}\left(\mu = \alpha L^{\beta}, \sigma^2 \mu^2\right)$$

Alternative log-linear version:

$$\log W = \alpha^{\star} + \beta \log L + \varepsilon, \qquad \alpha^{\star} = \log \alpha, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

Package: `robust`, (the Insightful Robust Library), (Wang et al., 2012).

Aggregated data:

Issue: Individual animals from a shot are classified by species and sex but have their carapace lengths measured. Only the total weight of the Tiger catch is measured.

Can we infer anything about the morphometric relationships from this kind of data?

If W_i is the total weight of a group of g_i animals with carapace lengths l_{ij}, and sexes s_{ij}, $j = 1, \ldots, g_i$, $i = 1, \ldots, n$, an approximate model would be

$$W_i \sim \mathrm{N}\left(\sum_{j=1}^{g_i} \mu_{ij}, \sum_{j=1}^{g_i} \sigma^2_{s_{ij}} \mu^2_{ij}\right), \qquad \text{where} \quad \mu_{ij} = \exp\left(\alpha_{s_{ij}} + \beta_{s_{ij}} \log l_{ij}\right)$$

Robust version:

Let the $t-$distribution with low degrees of freedom replace the normal. The likelihood, $\mathcal{L}$, to maximize is specified as

$$-2\log\mathcal{L} = \sum_{i=1}^{n}\left\{(\nu+2)\log\left(1+\frac{\left(w_i-\sum_{j=1}^{g_i}\mu_{ij}\right)^2}{\nu\sum_{j=1}^{g_i}\sigma^2_{s_{ij}}\mu^2_{ij}}\right)+\log\left(\sum_{j=1}^{g_i}\sigma^2_{s_{ij}}\mu^2_{ij}\right)\right\}$$

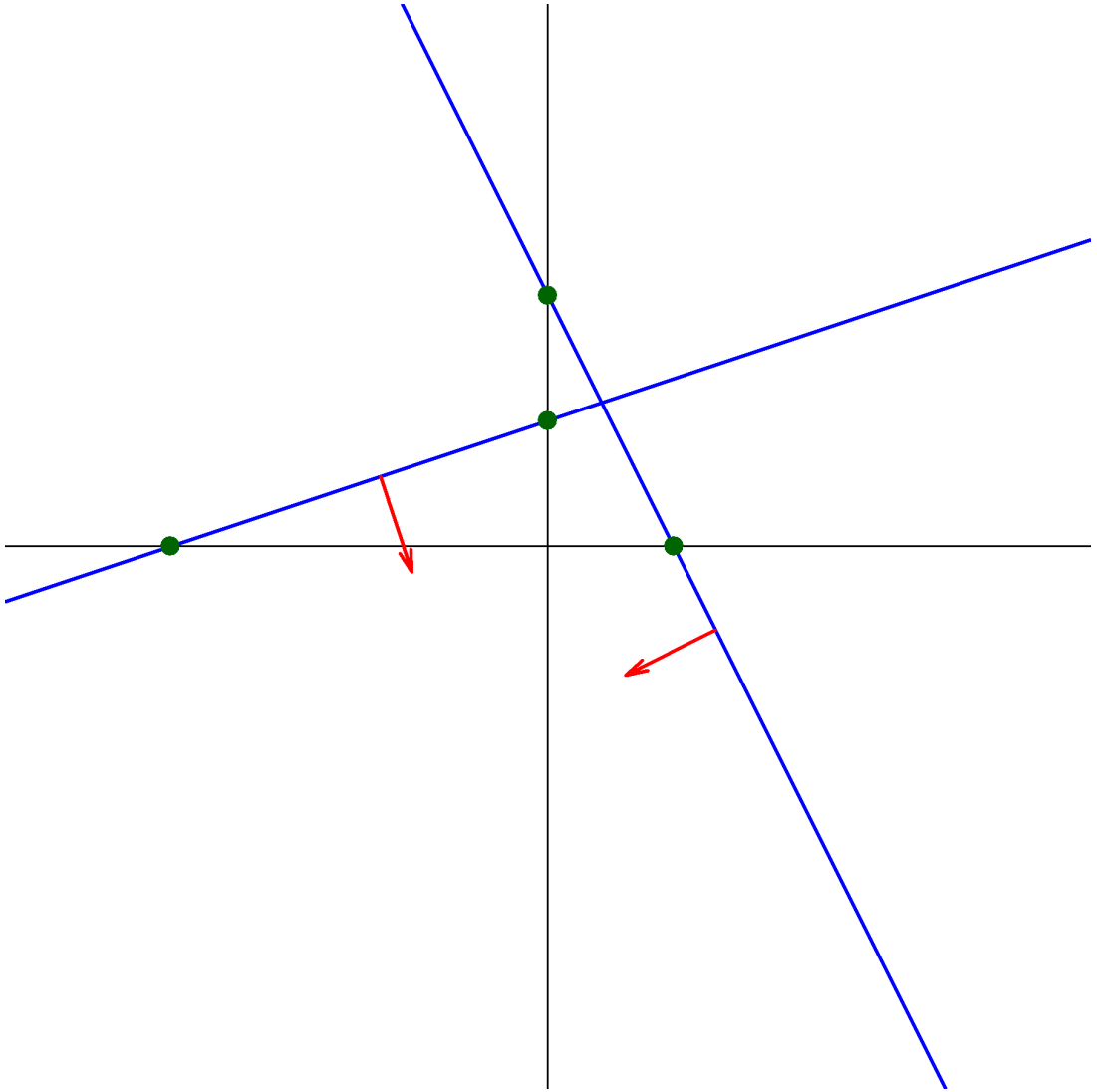
where the degrees of freedom, ν, is normally set at 3 or 5.

- Outliers are difficult to detect since information comes from both location and scale. An automated procedure is useful!
- Unless the group is relatively homogeneous in size, the total weight gives little information on morphometric relationships.
- A pure **R** solution is not difficult, but has to take full advantage of vectorization for the computations to be efficient.

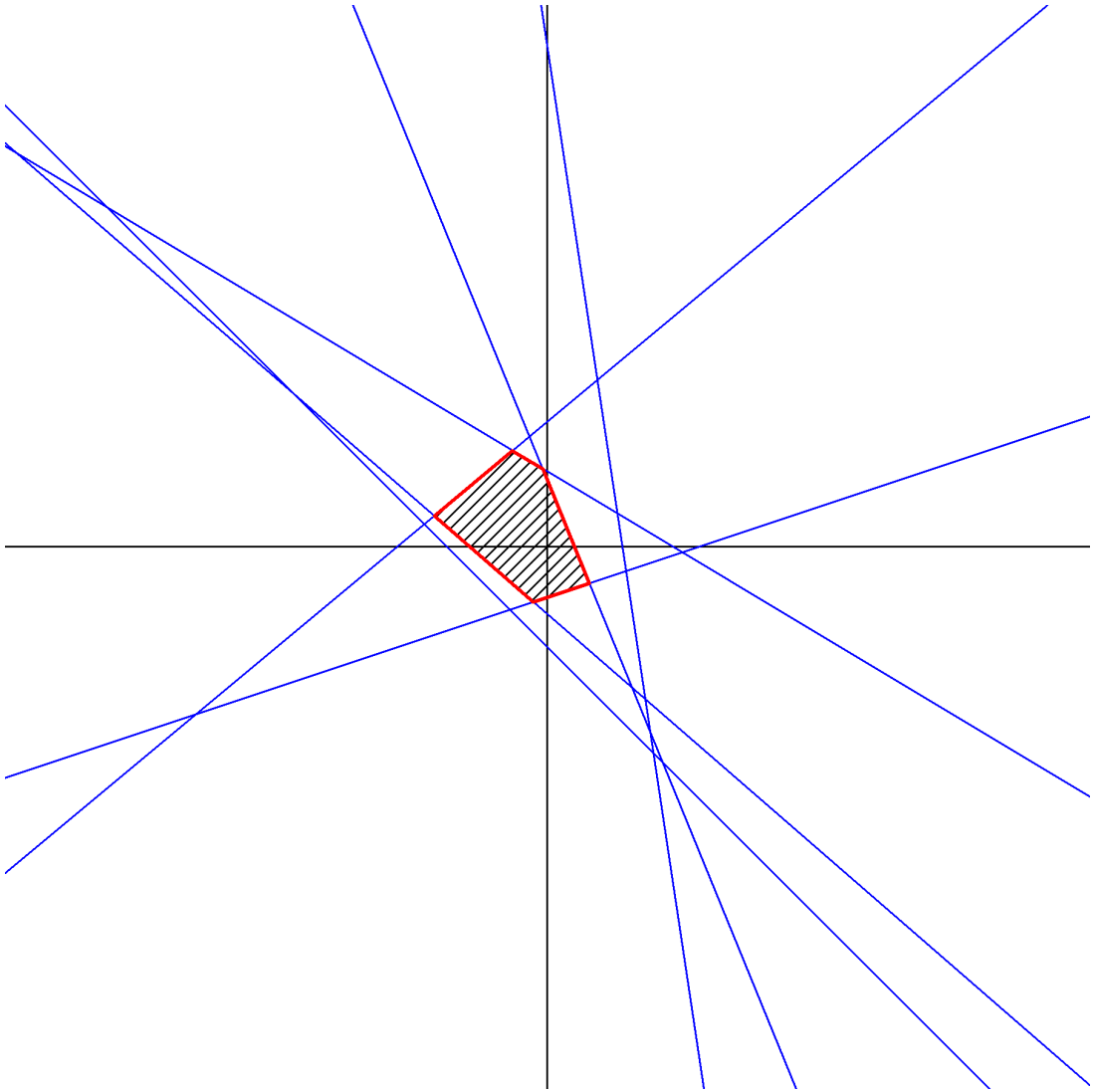
5 Dan's problem

A real problem in Time Series, but only an abstract version given here.

- The data supply a series of *intercepts* on the $x-$ and $y-$axes, typically thousands.
- Lines passing through the intercept pairs are guaranteed
  - *Not* to pass through the origin;
  - *Eventually* to enclose the origin.
- Dan's first problem: devise an efficient algorithm to *identify* and *define* the *smallest* polygon enclosing the origin.

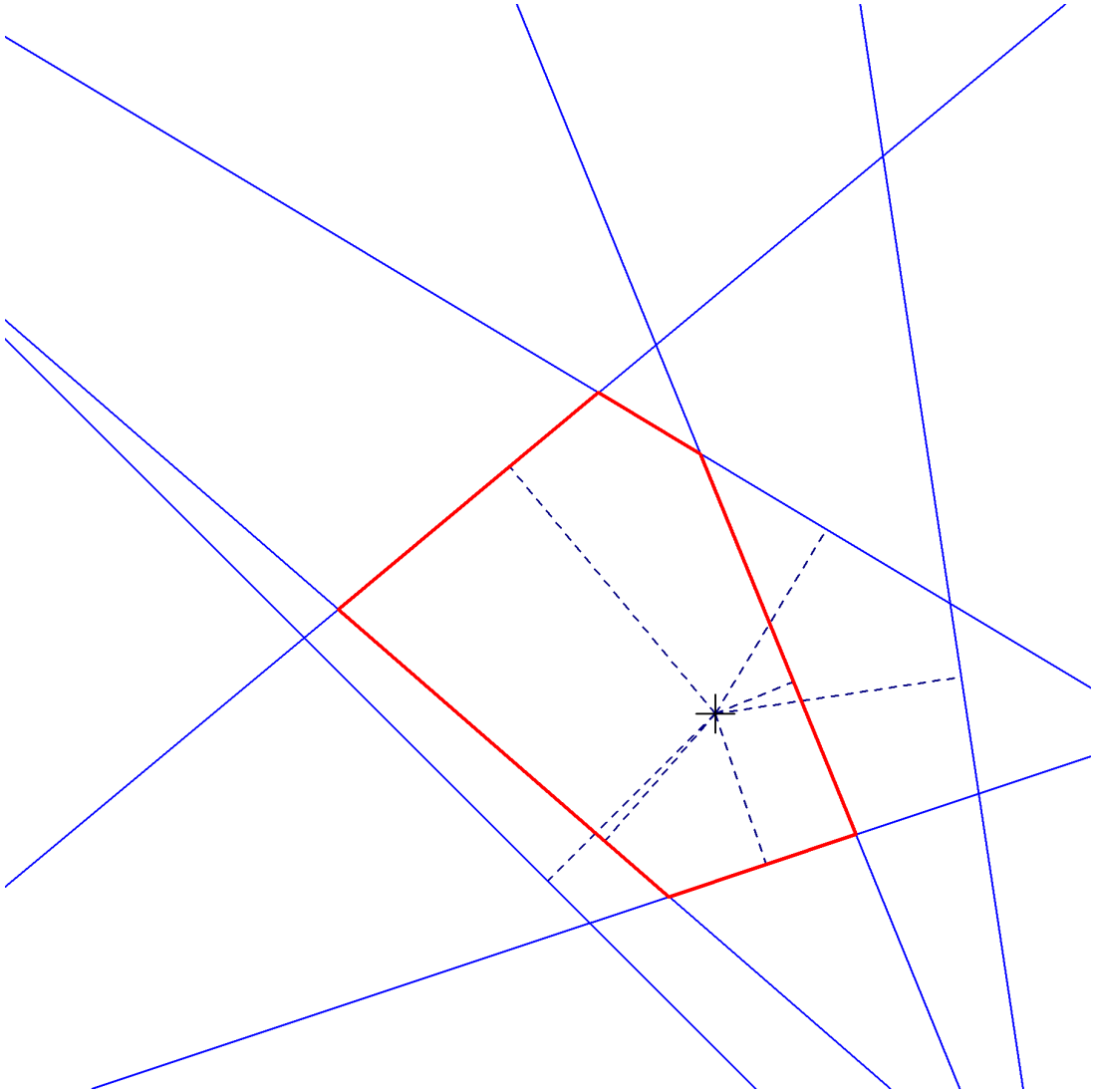


An artificial example:



An algorithm

- Find the *normals* to the lines, and hence the line *closest* to the origin.
- Rotate the plane so that this normal points South, i.e. the line is parallel to the $x-$axis.
- Find the points of intersection of this line with all the others.
- The two intersection points closest to the $y-$axis define the first (positive side) and last (negative side) corners of the minimum enclosing polygon.
- Rotate the plane again so that the *next* side is parallel to the $x-$axis. The next corner is then the intersection point with this line whose $x-$component is the smallest one larger than that of the current corner.
- Continue until the (known) last corner is reached.



Some code

```
> normals

function(x, y = NULL) { ## normals from intercepts
    z <- with(xy.coords(x, y),
              complex(real = x, imaginary = y)) ## Lazy evaluation
    x <- Re(z)
    y <- Im(z)
    m <- Mod(z)
    (x/m)*(y/m)*complex(, y, x)  ## numerical caution!
  }

> intersections

function(n1, n2) { ## two lines in 'normal' form
    n12 <- n1*Conj(n2)
    lambda <- ifelse(Im(n12) == 0, as.complex(Inf),
                     (Re(n12) - Mod(n2)^2)/Im(n12))
    n1*(1 + 1i*lambda)
  }
```

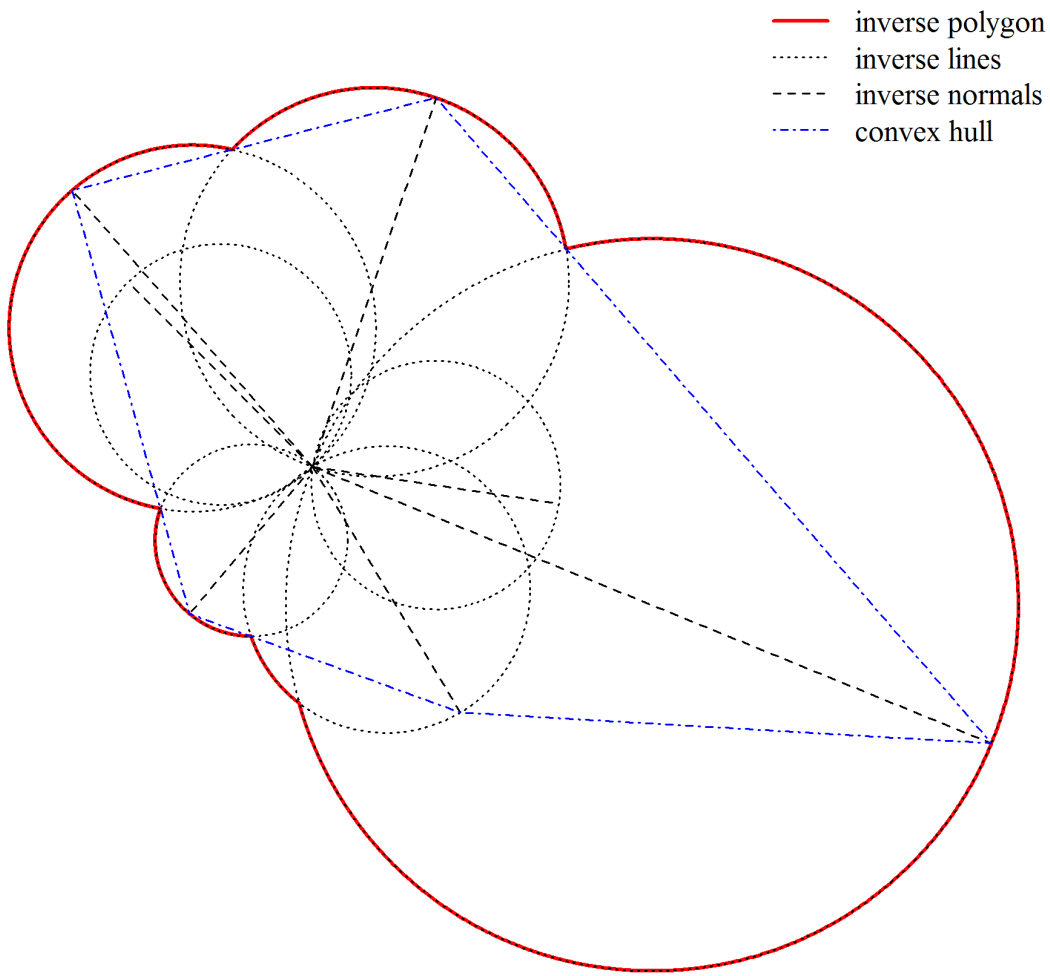
```
> convexCentre

function (x, y = NULL) {
    z <- with(xy.coords(x, y),
              complex(real = x, imaginary = y))
    z <- nz <- normals(z)
... <omitted code: check that the origin is enclosed> ...
    j0 <- which.min(Mod(nz))
    nz <- nz * complex(argument = South - Arg(nz[j0]))
    r0 <- Re(intersections(nz[j0], nz))
    i0 <- which(r0 == max(r0[r0 < 0])); I <- i0;  J <- j0
    while (!any(duplicated(J))) {
      I <- c(I, j0); i0 <- j0
      j0 <- which(r0 == min(r0[r0 > 0])); J <- c(J, j0)
      nz <- nz * complex(argument = South - Arg(nz[j0]))
      r0 <- Re(intersections(nz[j0], nz)); r0 <- r0 - r0[i0]
    }
    structure(intersections(z[I], z[J]), indices = cbind(I, J),
              class = "convexCentre")
  }
<environment: 0x03584224>
```

The breakthrough: inversion and convex hull

- The transformation $z \mapsto 1/z$ in the complex plane maps:
  - lines *not passing through* the origin into circles *passing through* the origin,
  - the (external) normal to the line to the diameter of the circle starting at the origin.
- The points defining the convex hull of the reciprocals of the normals give the lines, in order, defining the polygon enclosing the origin.

```
> convexCentre2

function(x, y=NULL) { ## using convex hull assumption
  z <- with(xy.coords(x, y),
            complex(real = x, imaginary = y))
  h <- chull(1/normals(z))
  z0 <- convexCentre(z[h])
  attr(z0, "indices")[] <- h[attr(z0, "indices")]
  z0
}
```

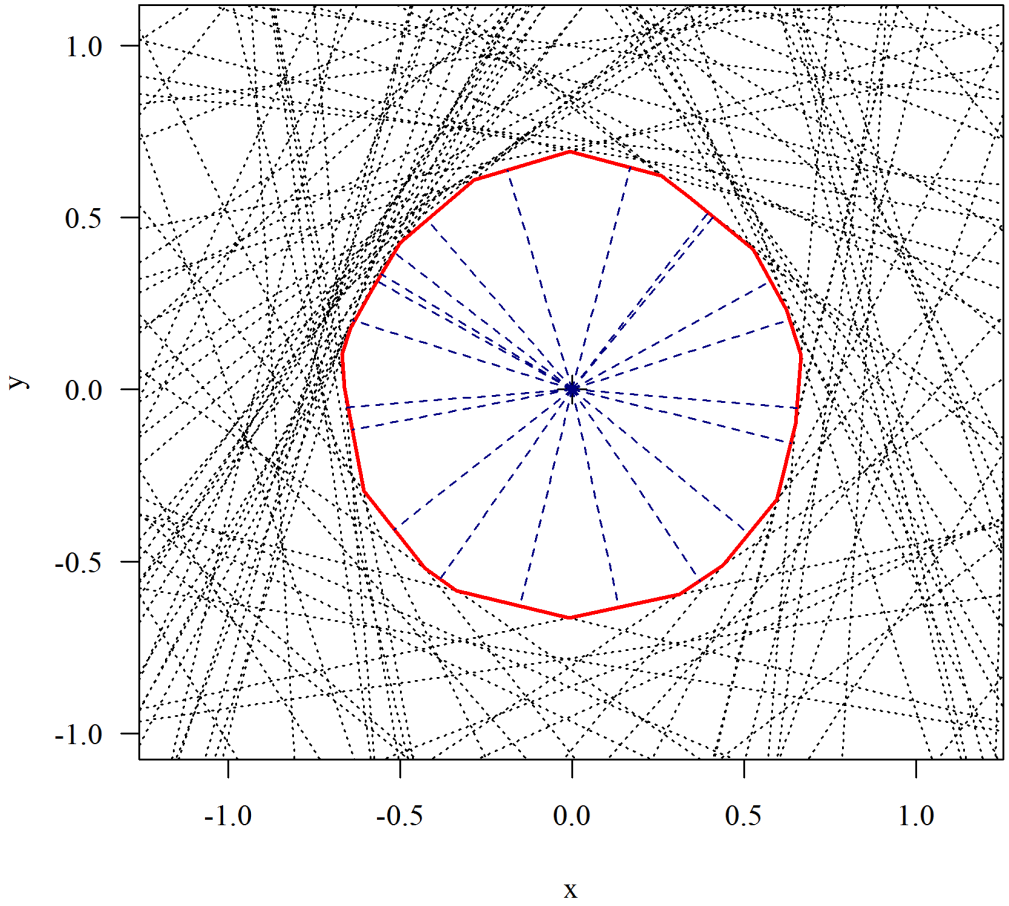
The artificial example: the inverse view



Some timings

```
> w <- complex(,rt(500000, 5), rt(500000, 5))
> rbind(orig = system.time(cc1 <- convexCentre(w))[1:3],
        hull = system.time(cc2 <- convexCentre2(w))[1:3])

     user.self sys.self elapsed
orig      4.41     0.43    5.62
hull      0.61     0.03    0.65

> identical(cc1,cc2)

[1] TRUE

> cc1

                          Corner  Line1  Line2
1  2.810050e-07-9.542936e-06i 485189 406443
2  2.809866e-07+2.087826e-06i 406443 131558
3 -2.109878e-06+2.087819e-06i 131558  84141
4 -2.109895e-06-9.542950e-06i  84141 485189
5  2.810050e-07-9.542936e-06i 485189 406443
```

A real example with 4390 lines



6 The winds of change

> "Language is a primary element of culture, and stasis in the arts is tantamount to death." – Charles Marsh

> "Merely quantitative differences, beyond a certain point, pass into qualitative changes." – Karl Marx, *Das Kapital, Kritik der politischen Ökonomie*, Vol. 1.

Change is necessary, inevitable and to be welcomed, but not all change will prove to be steps in the right direction.

The **R** community is one of its greatest strengths but all communities can experience problems, particularly at times of change.

Some causes for concern?

Collaboration can be "messy" (JMC).

Four types of package

- *Bad or inept packages*: purveyors of methods that are ideologically driven or objectively wrong, or simply misleading; badly coded and implemented or using protocols that are foreign to **S** usage conventions.
- *A home for the refugee*: packages that seek to emulate another system to ease the passage for newcomers.
- *Empires*: large interlocking suites of packages that seek to implement a variant philosophy of how to work with **R**,
- *GUis*: packages which ease the learning curve, but lead to a dead end only part of the way up the hill.

```
function (x) {
  start <- 1
  end <- length(x) + 1
  while (start < end) {
    y <- x[start]
    if (y > 0) {
      if (start == 1) {
        result = TRUE
      } else result <- c(result, TRUE)
    } else {
      if (start == 1) {
        result = FALSE
      } else result <- c(result, FALSE)
    }
    start <- start + 1
  }
  return(result)
}
```

This masterpiece of obfuscation is equivalent to[a]

```
function(x) x > 0
```

- the same package contains a function to produce a list of primes which is obscure, slow and wrong.
- because it is so slow, a list of "primes" less than 1e+7 is included as a data set, complete with errors, (e.g. 1, 133, ...)
- the package is described as being appropriate for mathematics teaching in schools
- several attempts have been made to have the package either fixed or withdrawn, all without success.

Can we have a mechanism to protect the unwary against this kind of trap without violating the spirit of CRAN or creating a lot of work for someone or more?

[a] except that the latter is more general: it works on zero-length vectors.

Roger D. Peng: I don't think anyone actually believes that R is designed to make *everyone* happy. For me, **R** does about 99% of the things I need to do, but sadly, when I need to order a pizza, I still have to pick up the telephone.

Douglas Bates: There are several chains of pizzerias in the U.S. that provide for Internet-based ordering (e.g.`www.papajohnsonline.com`) so, with the Internet modules in **R**, it's only a matter of time before you will have a pizza-ordering function available.

Brian D. Ripley: Indeed, the `GraphApp` toolkit . . . provides one . . . .

—*Roger D. Peng, Douglas Bates, and Brian D. Ripley*
R-help (June 2004)

Conclusions

- As a training tool **R** should remain a potent device for introducing young people to the excitement of interactive data analysis - the detective work of statistics - as well as training them in the logical discipline and mental focus of *programming*.
- As an *interactive environment for data analysis and graphics*, **R** should have a bright future, essentially as it is, for some time to come,
- As a *programming language* **R** may be reaching it's logical boundaries, within which it can productively remain also, for many years to come,
- The **R** community should nevertheless be open to new languages and environments that can work *alongside* **R** in a coherent way, to meet the software development challenges of the immediate and distant future.

- Perhaps **R** needs to have a more carefully worked out and explicit *succession plan* to replace the old guard, (like me), with younger people keen to meet those challenges and to keep the **R** community flourishing.

References

Becker, G. (2012). *SearchTrees: Spatial Search Trees*. CRAN. **R** package version 0.5.1.

Becker, R. A. and J. M. Chambers (1984). *S: An Interactive Environment for Data Analysis and Graphics*. Pacific Grove CA, USA: Wadsworth & Brooks/Cole. ISBN 0-534-03313-X.

Becker, R. A. and J. M. Chambers (1985). *Extending the S System*. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole. ISBN 0-534-05016-6.

Becker, R. A., J. M. Chambers, and A. R. Wilks (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole. ISBN 0-534-09192-X.

Chambers, J. M. (1998). *Programming with Data: A Guide to the* **S** *Language*. New York: Springer-Verlag.

Chambers, J. M. (2008). *Software for Data Analysis: Programming with* **R**. New York: Springer-Verlag.

Chambers, J. M. (2009). Facets of **R**. *The* **R** *Journal 1*, 5–8.

Chambers, J. M. and T. J. Hastie (Eds.) (1991). *Statistical Models in S*. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole. ISBN 0-412-05291-1.

Cook, J. D. (2012). Why and how people use **R**. `http://channel9.msdn.com/Events/Lang-NEXT/Lang-NEXT-2012/Why-and-How-People-Use-R`. See also `http://lambda-the-ultimate.org/node/4503` and `http://lambda-the-ultimate.org/node/4507`.

Hastie, T. (2011). *gam: Generalized Additive Models*. CRAN. R package version 1.06.2.

Marschner, I. C. (2011). *glm2: Fitting Generalized Linear Models*. CRAN. **R** package version 1.0.

Morandat, F., B. Hill, L. Osvald, and J. Vitek (2012). Evaluating the design of the **R** language: Objects and functions for data analysis. `http://www.cs.purdue.edu/homes/jv/pubs/ecoop12.pdf`. An ECOOP 2012 paper.

Tukey, J. W. (1969). Analysing data: Sanctification or Detective work? *American Psychologist 24*(2), 83–91.

Venables, W. N. and C. M. Dichmont (2004). A generalized linear model for catch allocation: an example from Australia's Northern Prawn Fisherh. *Fisheries Research 70*, 409–426.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with* **S** (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Wang, J., R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, and K. Konis. (2012). *robust: Insightful Robust Library*. CRAN. **R** package version 0.3-19.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B) 65*, 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*, 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with* **R**. Chapman & Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73*, 3–36.

Session information

- R version 2.15.0 (2012-03-30), `i386-pc-mingw32`
- Locale: `LC_COLLATE=English_Australia.1252`, `LC_CTYPE=English_Australia.1252`, `LC_MONETARY=English_Australia.1252`, `LC_NUMERIC=C`, `LC_TIME=English_Australia.1252`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: lattice 0.20-6, SOAR 0.99-10
- Loaded via a namespace (and not attached): grid 2.15.0, Matrix 1.0-6, mgcv 1.7-17, nlme 3.1-104, tools 2.15.0