Using R for Statistical Seismology

by Ray Brownrigg & David Harte

Statistical Seismology is a relatively new term used to describe the application of statistical methodology to earthquake data. The purpose is to raise new questions about the physical earthquake process, and to describe stochastic components not explained by physical models. Such stochastic quantification allows one to test the validity of various physical hypotheses, and also makes probability forecasts more feasible.

We describe here a suite of R packages, known as the "Statistical Seismology Library" (SSLib). This paper firstly introduces SSLib, providing a little history, and a description of its structure. Then the various components of SSLib are described in more detail and some examples are given. While the packages were developed primarily to analyse earthquake data, two of the packages (Fractal and PtProcess) have more general applicability.

For a general text on seismology, see Lay and Wallace (1995). Further, a large collection of seismological software can be found at http://orfeus.knmi.nl/other.services/software.links.shtml.

Introduction to SSLib

The Statistical Seismology Library (SSLib) is a collection of earthquake hypocentral catalogues (3D location of the rupture initiation point, time and magnitude) and R functions to manipulate, describe and model event data contained in the catalogues. At this stage, analyses include graphical data displays, fitting of point process models, estimation of fractal dimensions, and routines to apply the M8 Algorithm. While we have named it the "Statistical Seismology Library", the type of analyses that are performed really only reflect the research interests of the authors. Part of the rationale was to require our students and collaborators to formally document their programs so that others could determine what they were supposed to do, and to be able to use them after they have possibly moved on. Another aim was to make some of our statistical methods and models more directly available to our seismologist and geophysicist colleagues.

The library is divided into a number of R packages. Details of these and source code can all be found on the Statistical Seismology Library web page (http://homepages.paradise.net.nz/david.harte/SSLib/). Package names with a specifically seismological character are prefixed by "ss" (e.g. the New Zealand catalogue is named ssNZ), whereas those with a more general statistical interest are not (e.g. PtProcess and Fractal). A reference manual (standard R format) for each package can also be found on the web page, along with a Users Guide that contains examples and shows how the different packages relate to each other (Harte, 2005e).

SSLib was started in 1996 as an S-PLUS library (Harte, 1998). After being ported to the R language in 1999, development of SSLib switched to using the R implementation. At this time, SSLib was only available on the Unix and Linux platforms, but in 2004 a Windows version was released.

Earthquake Catalogues

Generally, users will want to use their own earthquake catalogues. However SSLib does contain various earthquake catalogues including the New Zealand catalogue and the PDE catalogue (Preliminary Determinations of Epicentres) which is a worldwide catalogue collated by the US Geological Survey. Further details of the catalogues available in SSLib can be found on the SSLib web page (see http://homepages.paradise.net.nz/david.harte/SSLib/).

The earthquake catalogues are based on raw data available from the World Wide Web. These data are generally collected by national seismological observatories. The raw data appears in many different formats, but the SSLib input routines coerce these different formats into a single common structure. This allows for both a uniformity in analysis, and the ability to make comparisons between the different catalogues. Nevertheless, the data structure within the catalogue packages allows for the inclusion of any number of extra data fields; see Harte (2005e) for further details.

Catalogue Manipulation Utilities

The **ssBase** package (Harte, 2005b) provides catalogue preparation and data manipulation functions. These include functions for date/time formatting, data summaries, printing and data subsetting. This package also contains other functions of a general nature used by the other packages.

A catalogue can be subsetted on any combination of location, time range, depth range or magnitude range, with the location being rectangular (in the latitude/longitude sense), circular (actually cylindrical in 3 dimensions), spherical, or based on an arbitrary polygon on the surface of the earth. Many of the analysis functions will work directly with a subset 'object'.

Exploratory Data Analyses

The **ssEDA** package (Harte, 2005c) consists of functions for exploratory data analysis. In particular these can provide histograms of event depth or event frequency (monthly or yearly), line plots of magnitude over time, maps of the locations of the epicentres, and frequency-magnitude plots to determine a *b*-value (see Figure 2) or to check the completeness of a catalogue. Further, the epicentre maps can identify different magnitudes and depths. Interactive graphics can be used to identify individual events on epicentral plots, and dynamic graphics can be used to show 3-dimensional views with rotation and linked plots.

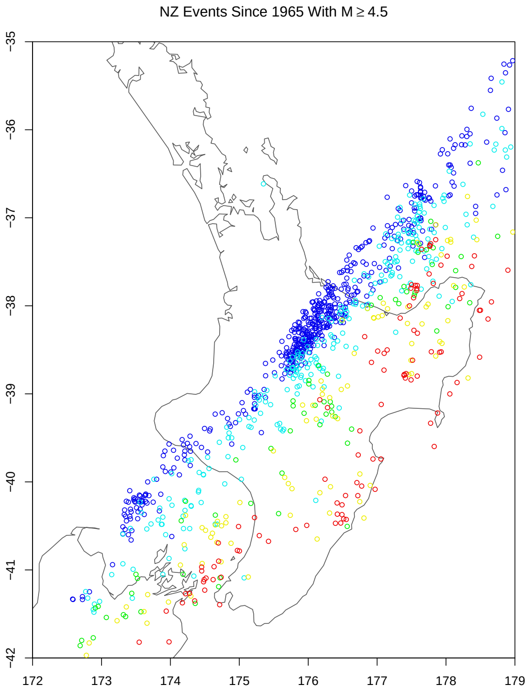


Figure 1: Epicentral plot of events from the NZ catalogue since 1965 with magnitude ≥ 4.5 (1777 events). The Pacific tectonic plate to the east is subducting the Australian plate to the west.

Figure 1 shows a map of the epicentres of a subset of events from the New Zealand catalogue. The colour of the point depicts the depth of the event, with the most shallow events at the red end of the spectrum and the deep events at the blue end of the spectrum. The R commands required to display this plot follow.

```
library(ssNZ)
library(ssEDA)
b <- subset.rect(NZ, minday=julian(1, 1, 1965),
                 minmag=4.5, minlat=-42,
                 maxlat=-35, minlong=172,
                 maxlong=179, mindepth=40)
epicentres(b, depth=c(40, 60, 80, 100, 150,
           200, Inf), mapname="nz",
           criteria=FALSE, cex=0.8,
           usr=c(172, 179, -42, -35))
title(expression(paste("NZ Events Since 1965
      With ", M >= 4.5)))
```

Figure 2 is a frequency-magnitude plot for the same subset as used in Figure 1 and displays the Gutenberg-Richter law (Lay and Wallace, 1995). This law says that the logarithm of the cumulative number of events with magnitude greater than *m* is linear as a function of *m*. The R commands required to present this graph (given the call to `subset.rect` from the previous figure) follow.

```
freq.magnitude(b)
title("Frequency Magnitude Power-Law
      Relationship")
```

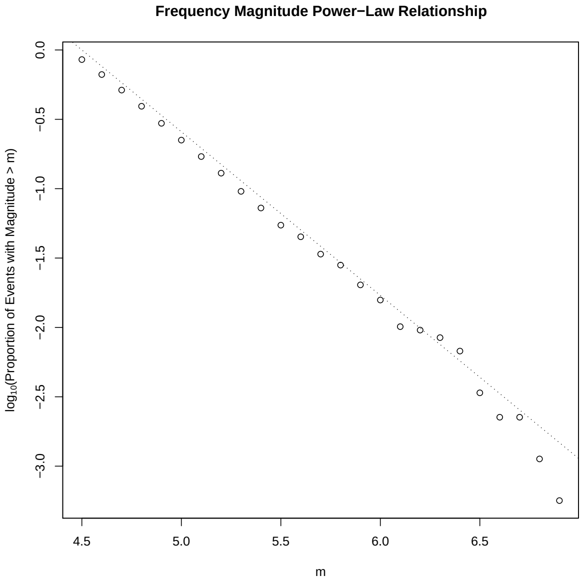


Figure 2: Plot of events from the NZ catalogue since 1965 with magnitude ≥ 4.5 showing the Gutenberg-Richter law. The absolute value of the slope is referred to as the *b*-value and is typically about one.

The interactive graphics is provided by the `epicentres.identify` function through use of the `identify` R function. The dynamic graphics is provided through the `threeD` function, which links to an external viewer provided by the `xgobi` software (Swayne et al., 1998) via the `xgobi` R function. Unfortunately `xgobi` is not easily usable on a Windows system, so work is in progress to use `ggobi` (GGobi) instead.

Point Process Modelling

The **PtProcess** package (Harte, 2004), which can be used independently of the other packages, provides a framework for point process modelling. This includes parameter estimation, model evaluation and simulation. This package is likely to be of most interest to a general statistical audience.

Our **PtProcess** package has a slightly different emphasis to the point process packages **spatstat** (available on CRAN) and **ptproc** (http://sandybox.typepad.com/software/ptproc/index.html). Our point process models are for *events* strictly ordered by time, and are conditional on the history of the process up to time t. This could be thought of as the *ground process* (Daley and Vere-Jones, 2003). Other event characteristics could be added as *marks*, e.g. earthquake event magnitude. The **spatstat** package (Baddeley and Turner, 2005) has an emphasis on the spatial location of items, presumably at the same point in time, e.g. tree locations or animals in a forest, etc. Emphasis is then on modelling the spatial intensity. Here marks can be added, e.g. the particular species of tree. The **ptproc** package (Peng, 2003) is derived from our package, and has extended the procedure into multidimensional processes. However, these additional variables are not treated as marks, and hence provides an alternative direction to our intended direction.

While the conditional intensity functions provided within the package have a distinctly seismological flavour, the general methodology and structure of the package is probably applicable to a reasonably large class of point process models. The models fitted are essentially marked point processes (Daley and Vere-Jones, 2003), where the mark distribution has been explicitly built into the conditional intensity function. Our next task is to separate the mark distribution and ground intensity function, and further generalise the framework so that any number of mark distributions can be attached to a given ground intensity function.

Currently the conditional intensity function is the most basic "building block". The conditional intensity function, $\lambda(t|\mathcal{H}_t)$, can be thought of as an instantaneous value of the Poisson rate parameter at time t and is conditional on the history of the process up to but not including t. We have given each conditional intensity function a suffix ".cif". There are a number of "generic"-like functions which perform some operation given an intensity function, for example, simulate a process, perform a goodness of fit evaluation, etc.

As an example, consider the ETAS (Epidemic Type Aftershock Sequence) model. This assumes that earthquake events behave in a similar way to an epidemic, where each event reproduces a number of aftershocks. The larger the event, the more aftershocks that will be reproduced. If various criticality conditions are satisfied, the aftershock sequence will eventually die out. See Harte (2004) and Utsu and Ogata (1997) for further technical details about the ETAS model.

The package contains an example dataset provided by Yosihiko Ogata. These data were simulated over the time interval $(0, 800)$. Using these data and approximate maximum likelihood solutions for the parameters contained in p (the ETAS model contains 5 parameters), the conditional intensity function can be plotted as follows (see Figure 3).

```
library(PtProcess)
data(Ogata)

p <- c(0.02, 70.77, 0.47, 0.002, 1.25)
ti <- seq(0, 800, 0.5)

plot(ti, log(etas.cif(Ogata, ti, p)),
     ylab=expression(paste("log ", lambda(t))),
     xlab="Time", type="l", xlim=c(0, 800),
     main="Conditional Intensity Function")
```

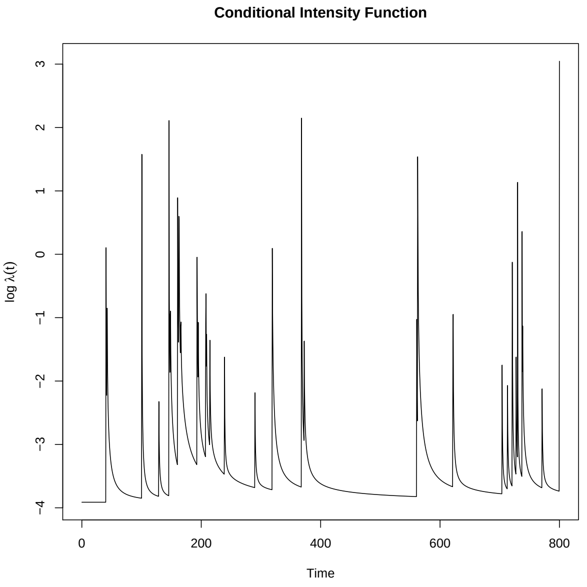


Figure 3: The spikes occur at event times and their heights are proportional to the event magnitudes, with an "Omori" decay (Lay and Wallace, 1995) in the rate after each event.

Further, the log-likelihood can be calculated as follows.

```
pp.LL(Ogata, etas.cif, p, c(0, 800))
```

Using the vector p as initial values, maximum likelihood parameter estimates can be calculated as follows.

```
posterior <- make.posterior(Ogata, etas.cif,
                            c(0, 800))
neg.posterior <- function(params)
                     (-posterior(params))
p <- c(0.02, 70.77, 0.47, 0.002, 1.25)
z <- nlm(neg.posterior, p, hessian=TRUE,
         iterlim=1000, typsize=p)
```

The function make.posterior creates a log-likelihood function to be evaluated on the time interval $(0, 800)$, and also gives one the option of enforcing constraints on the parameters (none here). We then create a negative posterior function because the function `nlm` is a minimiser. The maximum likelihood parameter estimates are contained within the object `z`.

The creation of the `posterior` function was partly done so that the function to be "optimised" within S-PLUS had only one argument. This is not necessary in R. Further, the use of priors has not been as useful as was initially thought. Consequently, it is probably best to revise the package so that the optimisation works more directly on the `pp.LL` function.

One way to test for the goodness of fit is to calculate the transformed residual process. This effectively creates a new time variable which magnifies or contracts the original process, assuming that the fitted model is correct, in such a manner that the resultant process is a homogeneous Poisson process with rate parameter one. A plot of the transformed event times versus the event number should roughly follow the line $x = y$. Large deviations from this line indicate a poorly fitting model. This can be achieved with the following code.

```
tau <- pp.resid(Ogata, z$estimate, etas.cif)
n <- nrow(Ogata)
plot(1:n, tau, type="l", xlab="Event Number",
     ylab="Transformed Time",
     xlim=c(0, n), ylim=c(0, n))
abline(a=0, b=1, lty=2, col="red")
```

Using the maximum likelihood parameter estimates, one can simulate the process over the subsequent interval $(800, 1200)$, say. This is achieved as follows.

```
x <- pp.sim(Ogata, z$estimate, etas.cif,
            TT=c(800, 1200))
```

The object `x` contains the original `Ogata` data, with the new simulated events appended. One may be interested in forecasting the time taken for an event with magnitude ≥ 6, say, to occur after time 800. One would then perform many such simulations, and determine the empirically simulated distribution for the given event of interest.

As can be seen, the conditional intensity function is the essential ingredient in each of the above analyses, and hence an entity like this is an essential component in a more object oriented setup for these models. It is a relatively simple step to set this up in a more object oriented manner, however, we have held off with this until we have disentangled the conditional intensity into its ground process and a general number of mark distributions.

Other Analyses

The **ssM8** package (Harte, 2005d) implements the Keilis-Borok & Kossobokov M8 algorithm (Keilis-Borok and Kossobokov, 1990). It is an empirically based algorithm that attempts to deduce "times of increased probability". We have been involved in projects that have attempted to test the efficacy of this algorithm.

The **Fractal** package (Harte, 2005a) has been used to calculate various fractal dimensions based on earthquake hypocentral locations, for example, see Harte (2001).

Problems and Future Development

We have already mentioned a number of extensions to the Point Process package: separating the conditional intensity into a ground intensity and a general number of mark distributions, writing more object oriented code, and determining if there is still a need for the `make.posterior` function. To implement more object oriented code, the naming conventions would clearly need to be changed.

There is a difference with the functions contained in the **chron** package (Ripley and Hornik, 2001; Grothendieck and Petzoldt, 2004) and our "datetimes" format in **ssBase** (Harte, 2005b). We would prefer to use the **chron** functions, however, we would like the `format.times` function within **chron** to have greater flexibility, including fractional numbers of seconds. For example, we would like the ability to specify a times format as `hh:mm:ss.s`, or `hh:mm:ss.ss`. Further, many historical earthquake events do not have all date-time components available, and our "datetimes" format has a mechanism to deal with this. Note that the `POSIX` functions are inappropriate, as the event times are generally stored as UTC times.

A feature that we have included in all of our packages is a "changes" manual page. This documents all recent changes made to the package with the date. We have found this particularly useful when old analyses have been repeated and different answers are produced!

As noted earlier, the type of analyses included in SSLib largely reflects our own research interests. This also means that it is continually being changed and augmented.

Bibliography

A. Baddeley and R. Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. ISSN 1548-7660. URL http://www.jstatsoft.org. 33

D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*, volume I: Elementary Theory and Methods. Springer-Verlag, New York, second edition, 2003. 33

GGobi. GGobi data visualization system. http://www.ggobi.org/, 2003. 32

G. Grothendieck and T. Petzoldt. R help desk: Date and time classes in R. *R News*, 4(1):29–32, 2004. 34

D. Harte. Documentation for the Statistical Seismology Library. Technical Report 98-10, School of Mathematical and Computing Sciences, Victoria University of Wellington, Wellington, New Zealand, 1998. 31

D. Harte. *Multifractals: Theory and Applications*. Chapman and Hall/CRC, Boca Raton, 2001. 34

D. Harte. *Package PtProcess: Time Dependent Point Process Modelling*. Statistics Research Associates, Wellington, New Zealand, 2004. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/pp.pdf. 32, 33

D. Harte. *Package Fractal: Fractal Analysis*. Statistics Research Associates, Wellington, New Zealand, 2005a. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/fractal.pdf. 34

D. Harte. *Package ssBase: Base Functions for SSLib*. Statistics Research Associates, Wellington, New Zealand, 2005b. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/base.pdf. 31, 34

D. Harte. *Package ssEDA: Exploratory Data Analysis for Earthquake Data*. Statistics Research Associates, Wellington, New Zealand, 2005c. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/eda.pdf. 32

D. Harte. *Package ssM8: M8 Earthquake Forecasting Algorithm*. Statistics Research Associates, Wellington, New Zealand, 2005d. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/m8.pdf. 34

D. Harte. *Users Guide for the Statistical Seismology Library*. Statistics Research Associates, Wellington, New Zealand, 2005e. URL http://homepages.paradise.net.nz/david.harte/SSLib/Manuals/guide.pdf. 31

V. Keilis-Borok and V. Kossobokov. Premonitory activation of earthquake flow: algorithm M8. *Phys. Earth & Planet. Int.*, 61:73–83, 1990. 34

T. Lay and T. Wallace. *Modern Global Seismology*. Academic Press, San Diego, 1995. 31, 32, 33

R. Peng. Multi-dimensional point process models in R. *Journal of Statistical Software*, 8(16):1–24, 2003. ISSN 1548-7660. URL http://www.jstatsoft.org. 33

B. Ripley and K. Hornik. Date-time classes. *R News*, 1(2):8–11, 2001. 34

D. F. Swayne, D. Cook, and A. Buja. XGobi: Interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998. ISSN 1061-8600. URL http://www.research.att.com/areas/stat/xgobi/. 32

T. Utsu and Y. Ogata. Statistical analysis of seismicity. In J. Healy, V. Keilis-Borok, and W. Lee, editors, *Algorithms for Earthquake Statistics and Prediction*, pages 13–94. IASPEI, Menlo Park CA, 1997. 33

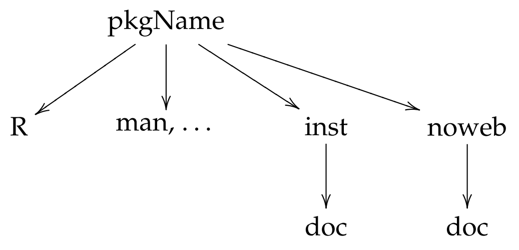
Ray Brownrigg
Victoria University of Wellington
ray@mcs.vuw.ac.nz
David Harte
Statistics Research Associates
david@statsresearch.co.nz

Literate programming for creating and maintaining packages

Jonathan Rougier

Outline

I describe a strategy I have found useful for developing large packages with lots of not-obvious mathematics that needs careful documentation. The basic idea is to combine the 'noweb' literate programming tool with the Unix 'make' utility. The source of the package itself has the usual structure, but with the addition of a `noweb` directory alongside the `R` and `man` directories. For further reference below, the general directory tree structure I adopt is



where '. . . ' denotes other optional directories such as `src` and `data`.

The files in the `noweb` directory are the entry-point