Spatial Statistics

AMS 245

Bruno Sansó

www.ams.ucsc.edu/~bruno

Department of Applied Mathematics and Statistics
University of California Santa Cruz

General Description

This course is intended for students that have a background in statistical methods and modeling that includes Bayesian statistics.

The course is focused on models for data that are spatially referenced.

The course will have a strong emphasis on model based geostatistic methods with Bayesian inference.

We will look into the theoretical properties of those models as well as into the computational issues involved in the estimation of their parameters.

Familiarity with R, linear models, Bayesian methods and MCMC methods will be assumed..

General Description

Geostatistics refers to models for random processes that are indexed at fixed locations that are irregularly scattered.

We will look into the theoretical properties of those models as well as into the computational issues involved in the estimation of their parameters.

Familiarity with R, linear models, Bayesian methods and MCMC methods will be assumed.

References

- Hierarchical Modeling and Analysis for Spatial Data, Second Edition, S. Banerjee, B.P. Carlin and A.E. Gelfand. Chapman and Hall.
- Statistics for Spatial Data, N.A.C. Cressie. Wiley.
- Statistics for Spatio-Temporal Data, N.A.C. Cressie and C.K. Wikle. Wiley.
- Model-based Geostatistics, P.J. Diggle and P.J. Ribeiro. Springer
- Handbook of Spatial Statistics, edited by A.E. Gelfand, P.J. Diggle, M. Fuentes and P. Guttorp. CRC Press.

References

- Statistical Analysis of Environmental Space-Time Processes, N.D. Le and J.V. Zidek. Springer.
- Statistical Methods for Spatial Data Analysis, O. Schabenberger and C.A. Gotway. Chapman and Hall/CRC
- Interpolation of Spatial Data, M.L. Stein. Springer.
- Correlation Theory of Stationary and Related Random Functions, A.M. Yaglom. Springer.

Introduction

We are interested in spatial processes $X(s)$, $s \in S \subseteq R^n$ where n is usually small, say 2 or 3. We start by considering the univariate case where $X(s) \in R$.

We need to consider dependencies in the distribution of $X(s)$ that are induced by the indexing variable s

As in the case of time series, we often have substantial amounts of data, but only one realization. We need to overcome this problem by assuming either strong structural or prior information or some 'repeatability' in the form of symmetry or stationarity.

Types of Models

A possible classification of spatial models is as follows:

- **Models for continuous random surfaces**. That is, S is a continuum. It is common to assume that X(s) is Gaussian or (almost equivalently) to model only the first two moments of the process. This area is known as Geostatistics.

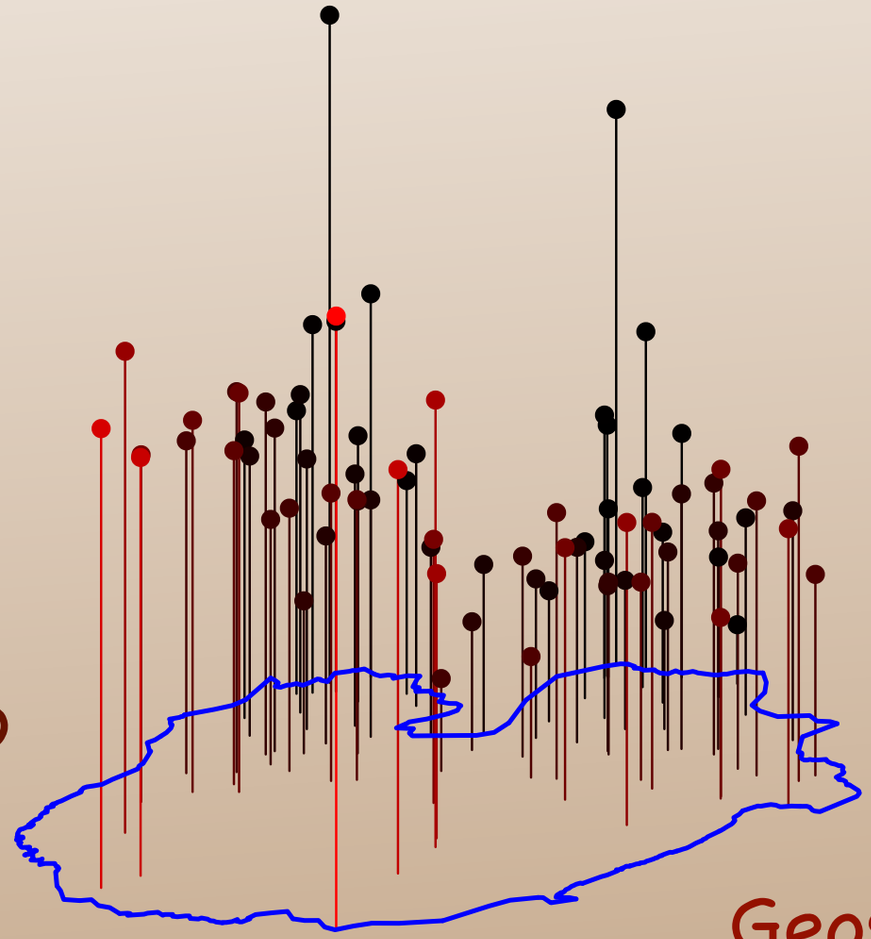
- **Models for mosaic phenomena**, where S is countable (usually finite) and often a lattice. Markov Random Fields are the most popular models of this type.

- **Points and attributes processes,** where $s \in S$ is a random variable. So the inference focuses, at least initially, about the locations where the process is observed. These are Point Processes. When some attributes of the process are also considered then we refer to Marked Point Processes.

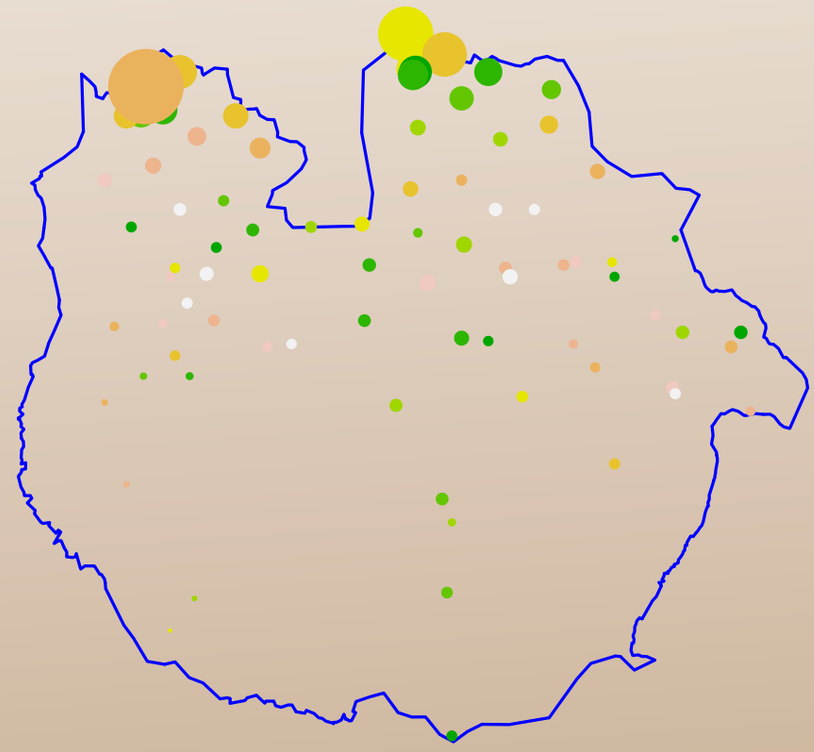
Scattered Observations

This figures show a scatterplot of annual precipitation data, and the locations of the stations with dot sizes proportional to the elevation

Mean Annual Precipitation 1968–1983
Guarico State – Venezuela



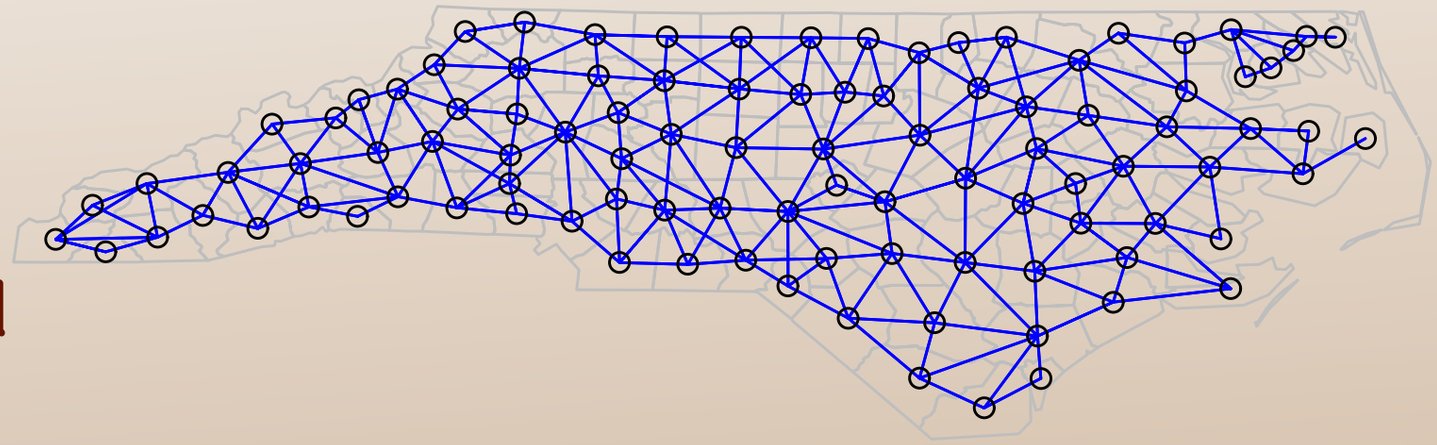
Heights of Guarico Locations



Geostatistical Models are used to create a surface

Lattice Data

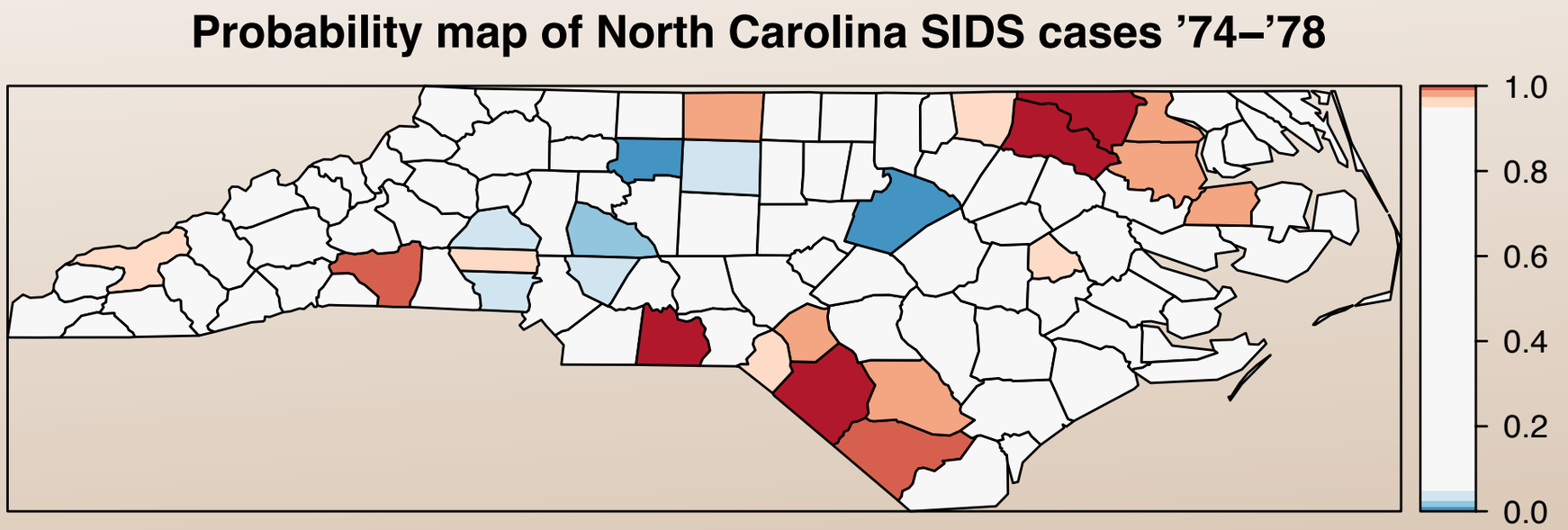
Neighbor Structure for the data on sudden infant deaths in North Carolina. The counties that are directly connected are neighbors.



Statistical models for these data are based on modeling the distribution of the attributes of a given county conditional on the neighbors. This has the advantage of producing sparse structures.

Lattice Data

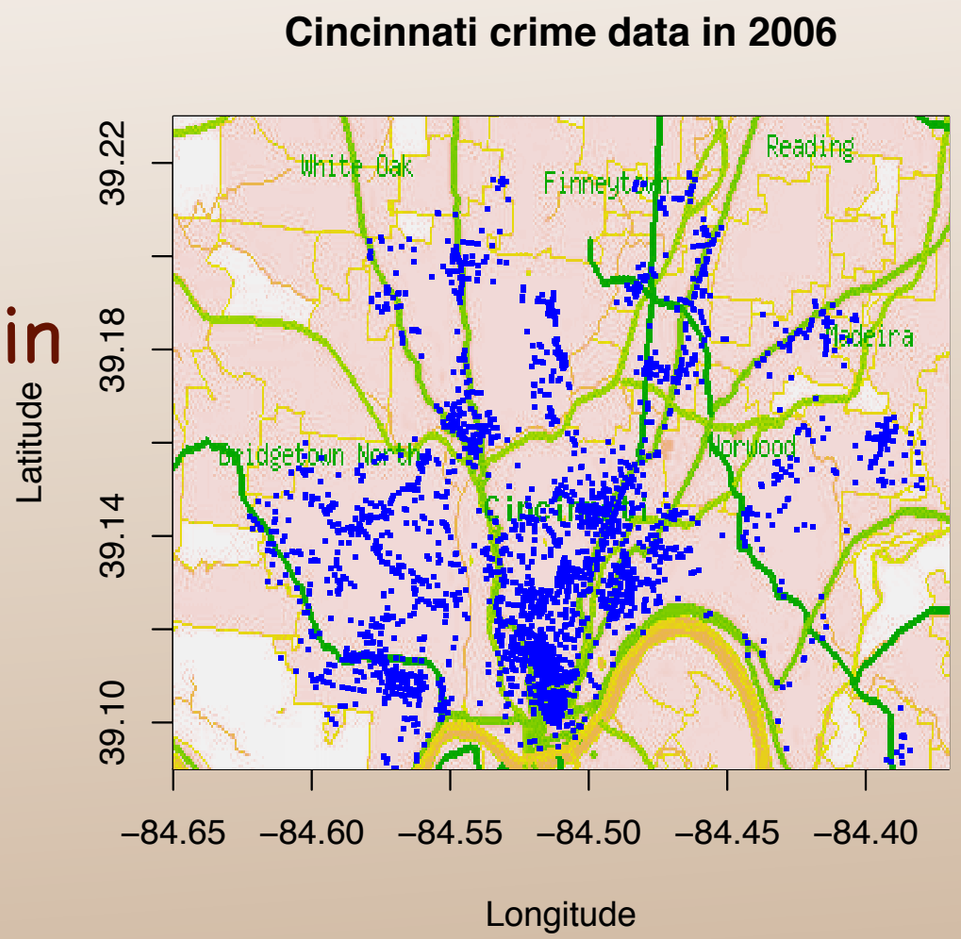
For count data, like the SIDS, we are interested in rates or probabilities of observing an event of interest in a given cell.



The purpose of the modeling is to obtain a smooth version of the raw data that accounts for the neighbor structure.

Point Processes

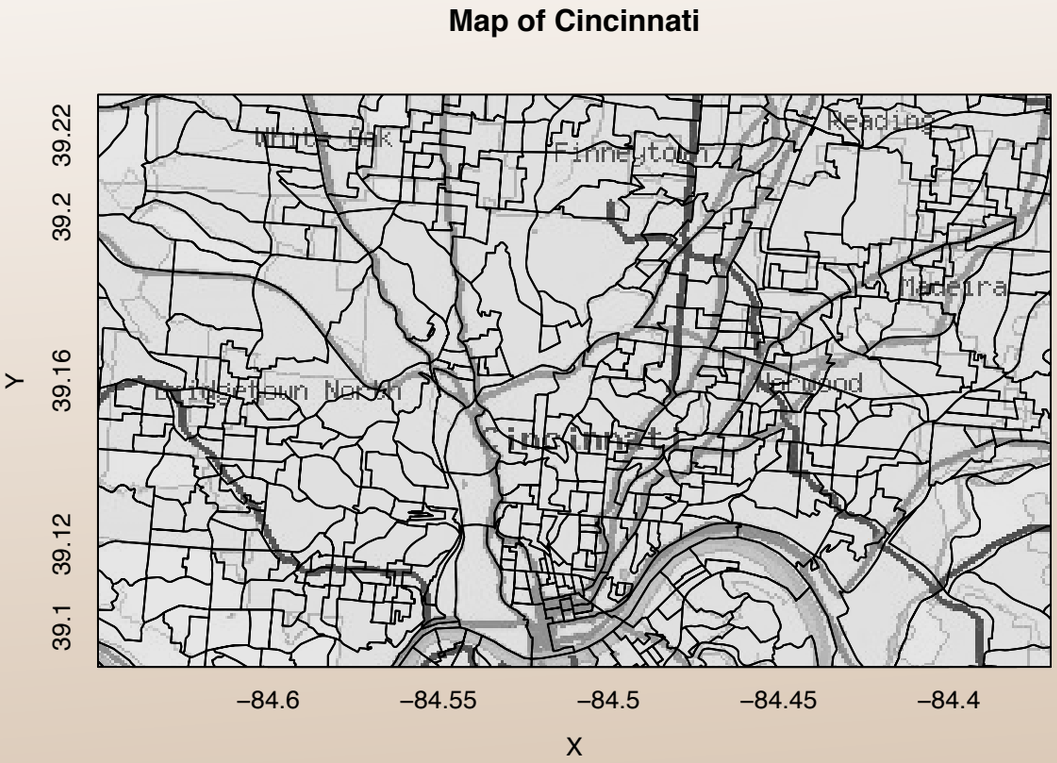
Locations of major crimes in Cincinnati in 2006.



We are interested in a description of the intensity of the process, and the possibility of associating non-homegeneities to spatially-varying covariates.

Marks allow to assign attributes to the process occurrences, producing a marked point process

Point Processes of Lattice Models?

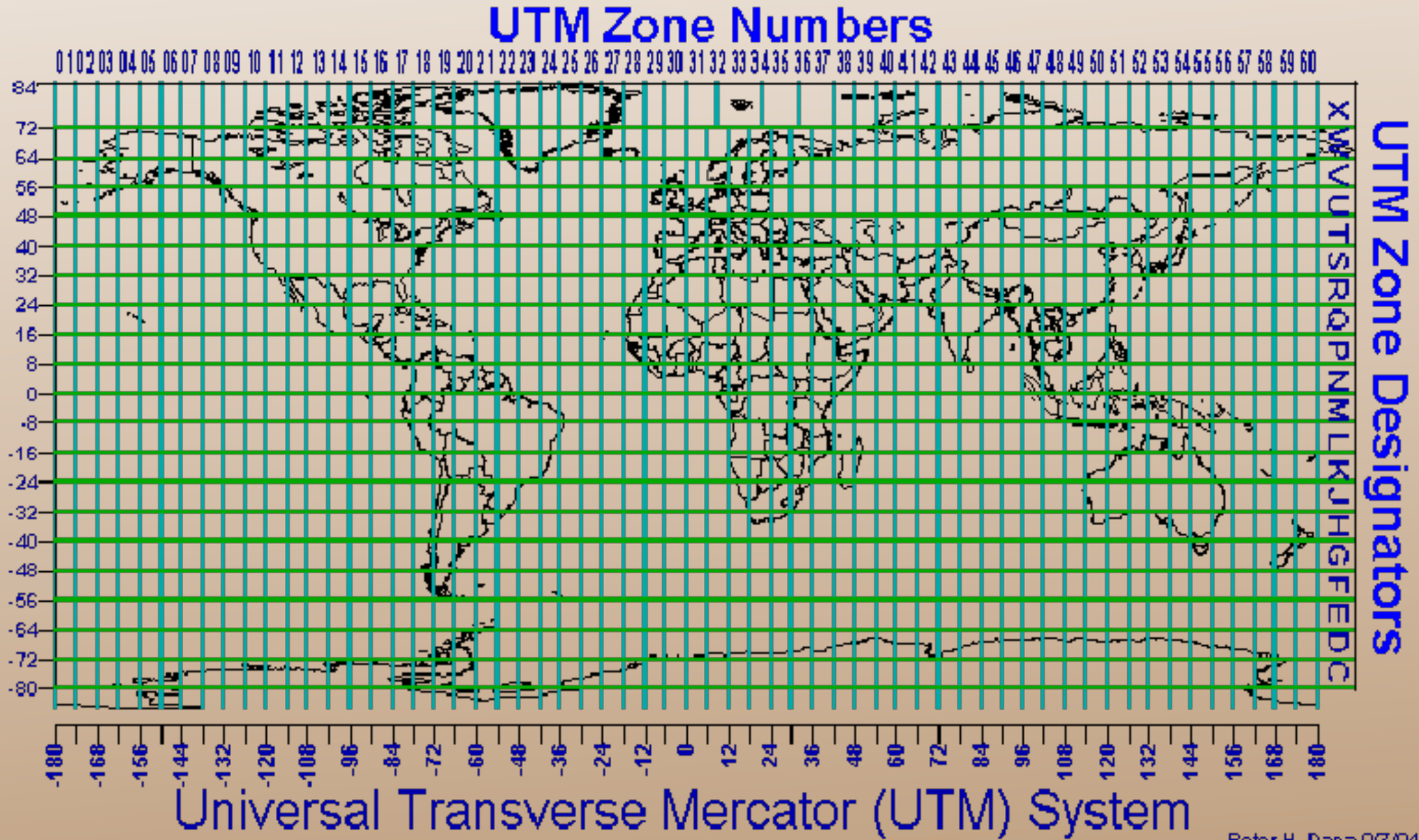


Lattice data can be obtained from point processes discretized over predefined spatial units. Thus, the number of counts per unit is a Poisson random variable.

Using a discretized version of the model allows for the use of generalized linear models. But, spatial continuity is lost, and overdispersion may be artificially induced by the discretization.

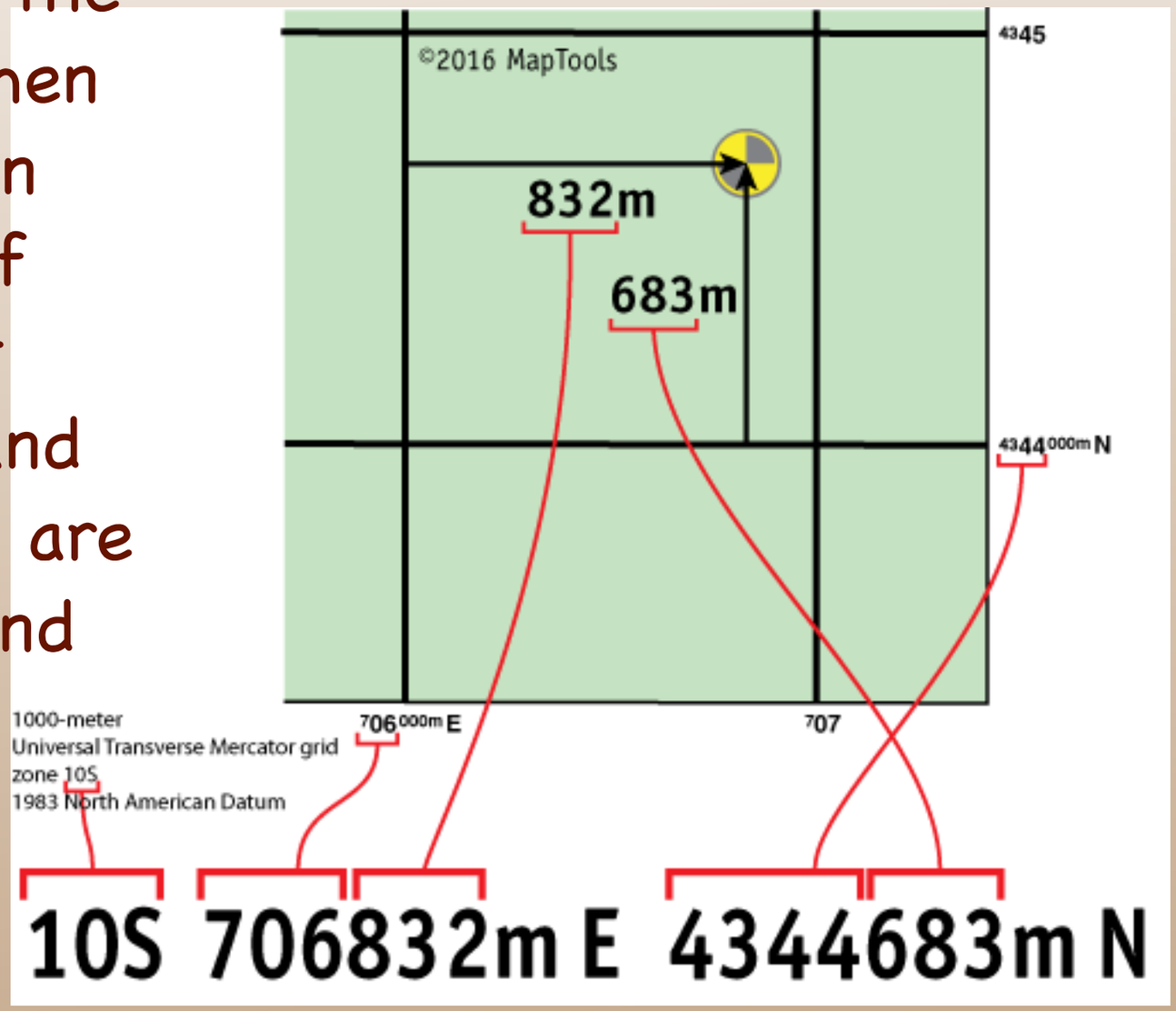
Coordinates and Distances

Spatially referenced data need a reference coordinate system to index the location of the observation. A popular system of coordinates is the Universal Transverse Mercator (UTM), which is based on 60 zones of 6 degrees each.



Easting and Northing

This is based on superimposing a grid on the geographical area and then measuring the distance in meters from the point of interest and the nearest grid lines to the south and west. So the coordinates are referred to as Easting and Northing.



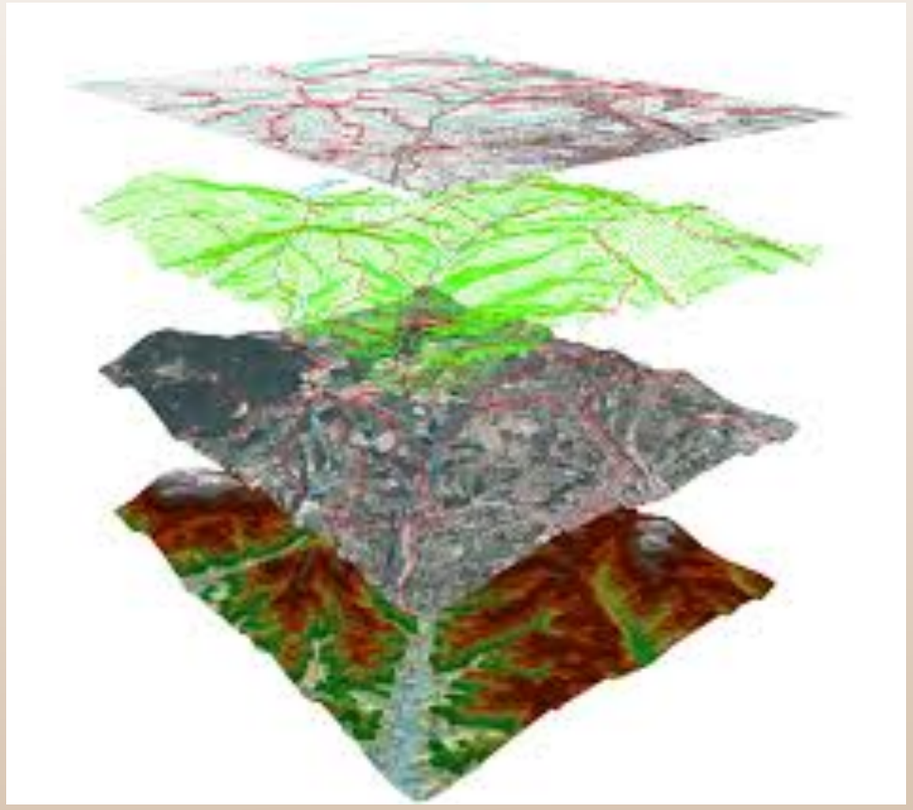
R Packages

There are a number of R packages that are relevant for the analysis of spatial data. A comprehensive discussion of the available packages is presented in the CRAN Task View: **Analysis of Spatial Data** http://cran.r-project.org/web/views/Spatial.html.

The ones that I most familiar with are geoR; fields and spBayes.

Exploring Spatial Fields

As with any statistical modeling problem, the first step is to perform a carefully descriptive analysis of the data. We are looking for clues to explain the spatial variability; the existence of anomalous observations; the shape of the surface; the existence of obvious long range trends, among other things



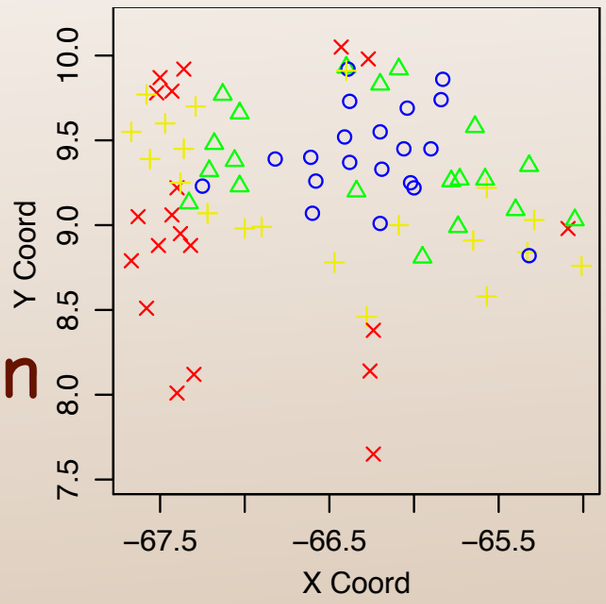
Geographical Information Systems (GIS) are of great help to perform these tasks. There is a substantial number of R packages that provide GIS capabilities.

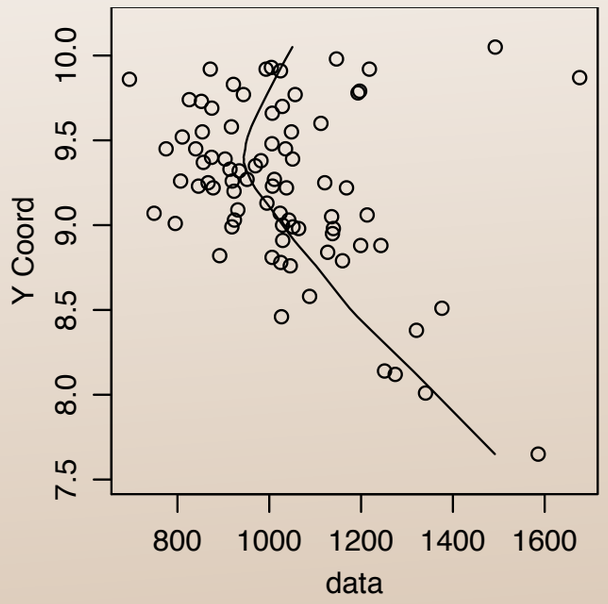
Plotting Spatial Data

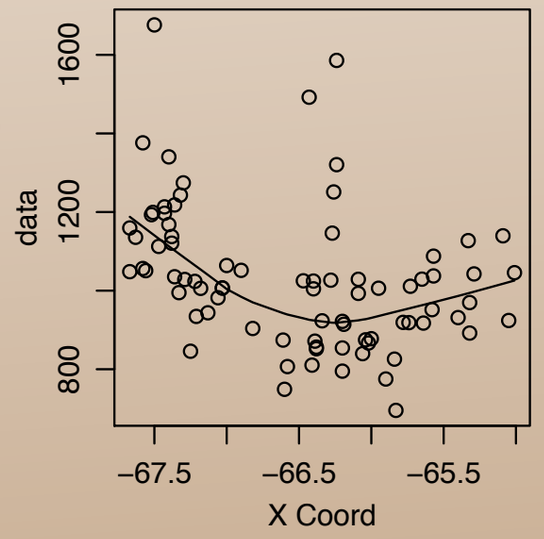
The package geoR offers a simple tool to explore the spatial variability, the depend on coordinates and distribution of a spatial field.

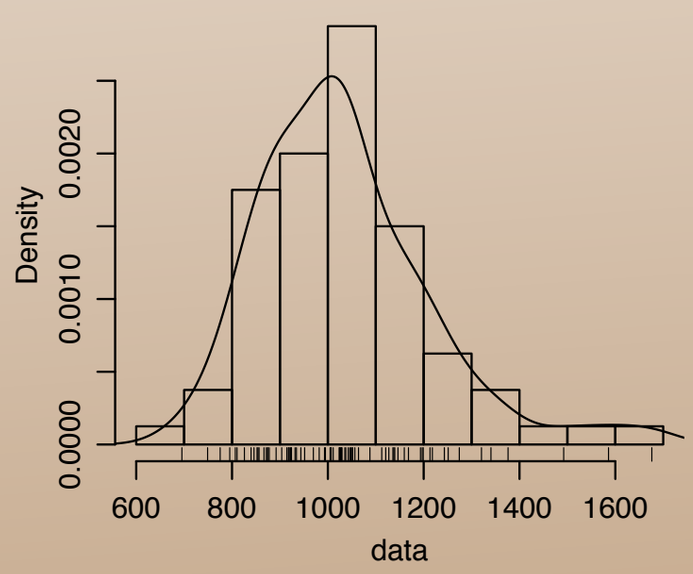
The plot to the right was produced with:

```
> class(lluvia)
[1] “geodata”
> plot(lluvia,lowess=T)
```

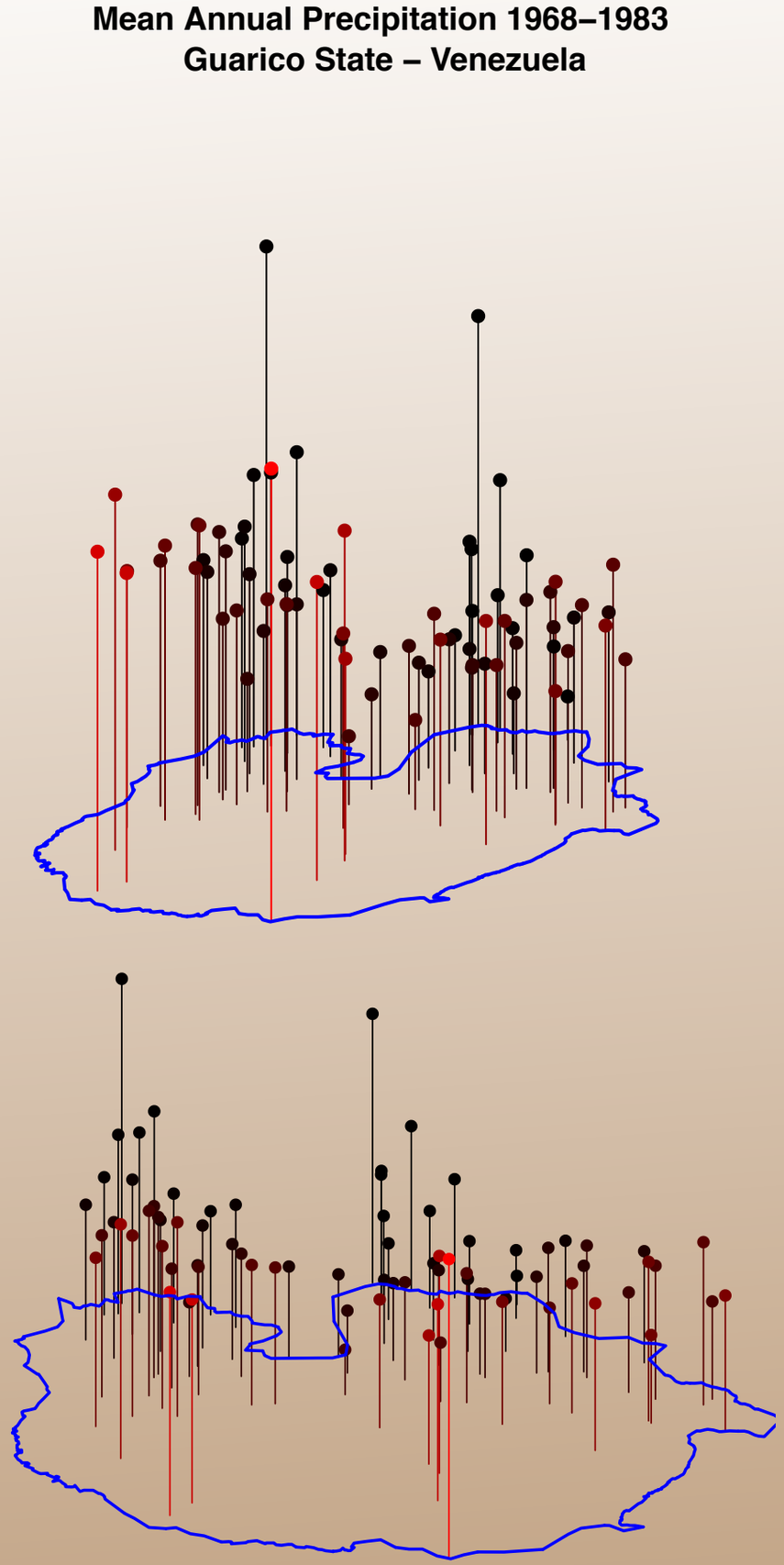








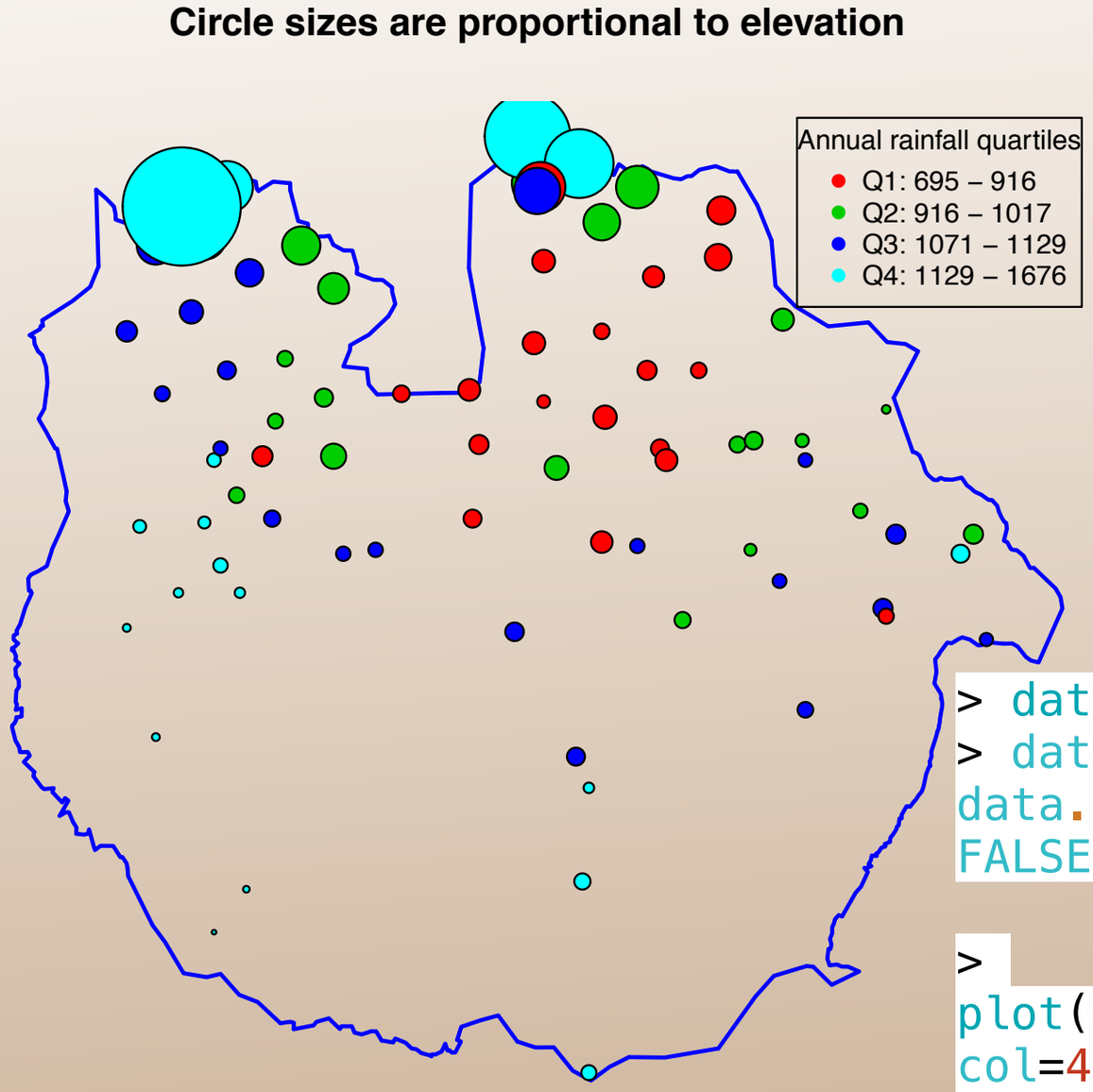
3D plots



The package `scatterplot3d` allows to visualize clouds of points in space

```
> library(scatterplot3d)
rain3d=scatterplot3d(long.lat[1],
long.lat[2],annual.mean,box=F,axis=F,grid
=F,type='h',pch=19,highlight.3d=T)

> rain3d$points3d(guar$x,guar$y,
rep(600,505),type='l',col=4,lwd=2)

#Change the angle

> rain3d=scatterplot3d(long.lat[1],
long.lat[2],annual.mean,box=F,axis=F,grid
=F,type='h',pch=19,highlight3d=T,angle=9)

>rain3d$points3d(guar$x,guar
$y,rep(600,505),type='l',col=4,lwd=2)
```

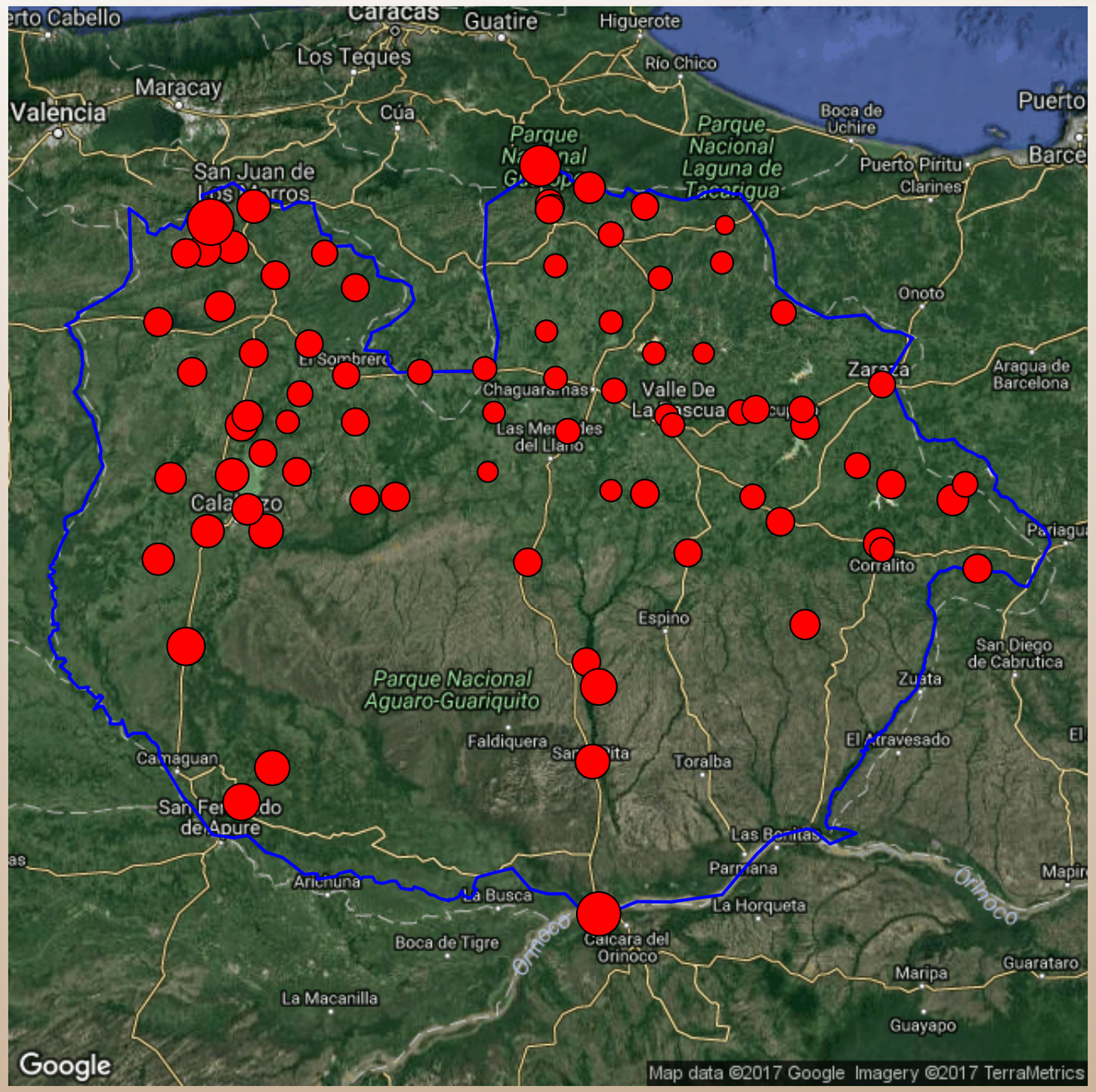
Using Colors and Symbols



Appropriate use of the usual plot function can be used to produce plots with several layers of information

```
> data.breaks=unique(quantile(annual.mean))
> data.cut <- cut(annual.mean, breaks =
data.breaks, include.l = TRUE,labels =
FALSE)

>
plot(guar,axes=F,xlab="",ylab="",type='l',
col=4,lwd=2)
> symbols(long.lat[1],long.lat[2],
circles=long.lat[,3],add=T,inches=.3,
bg=data.cut+1)
> legend(locator(1),legend=c('Q1: 695 -
916',
'Q2: 916 - 1017','Q3: 1071 - 1129','Q4:
1129 - 1676'),pch=rep(19,4),col=2:5,cex=.8,
title='Annual rainfall quartiles')
```

Using Google Maps



The package RgoogleMaps has functionalityes to map information obtained from Google

```
> library(RgoogleMaps)
> Gmap=GetMap.bbox(lonR=guar$range
[1:2],latR=guar$range[3:4])
> PlotOnStaticMap(Gmap)
> convert.points=LatLon2XY.
centered(Gmap,guar$y,guar$x)
> lines(convert.points$newX,
convert.points$newY,col=4,lwd=2)
> convert.points=LatLon2XY.
centered(Gmap,long.lat[,2],long.lat[,
1])
> symbols(convert.points$newX,
convert.points$newY,
circles=annual.mean,
inches=.15,bg=2,add=T)
```

The size of the circles is proportional to mean annual rainfall

Calculating Distances

When calculating distances over the surface of the earth that are far apart, one has to account for the curvature.

Consider two points on the surface given in latitude and longitude, say $P_1 = (\theta_1, \lambda_1)$ and $P_2 = (\theta_2, \lambda_2)$. The distance is given by

$$D = R\phi$$

$2 * \pi * r$

where

$$\cos(\phi) = \sin(\theta_1)\sin(\theta_2) + \cos(\theta_1)\cos(\theta_2)\cos(\lambda_1 - \lambda)$$

and R is the radius of the earth.

Thus, ϕ is the angular distance between $\mathrm{P1}$ and $\mathrm{P2}$.

Calculating Distances

```
> library(fields)
>
round(rdist.earth(long.lat[stations,],miles=F
),2)
       [,1]   [,2]   [,3]   [,4]   [,5]
[1,]   0.00 144.42  68.53 139.54  68.56
[2,] 144.42   0.00 160.65 268.24 199.26
[3,]  68.53 160.65   0.00 110.08 126.03
[4,] 139.54 268.24 110.08   0.00 147.46
[5,]  68.56 199.26 126.03 147.46   0.00
```

```
> round(rdist(planar[stations,])/1000,2)
       [,1]   [,2]   [,3]   [,4]   [,5]
[1,]   0.00 144.44  68.22 139.55  68.21
[2,] 144.44   0.00 160.85 268.45 198.98
[3,]  68.22 160.85   0.00 110.06 125.29
[4,] 139.55 268.45 110.06   0.00 147.11
[5,]  68.21 198.98 125.29 147.11   0.00
```

The package `fields` can be used to calculate distances very efficiently, with or without taking into account curvature

When using UTM coordinates we ignore the earth curvature