Flexible spatial modelling with INLA and inlabru

Janine B Illian and others

November 3, 2017

Chapter 1

The key ideas

1.1 Introduction – what is this book about

Somebody who opens this book might wonder – it this book for me? What is this book about and will I benefit from reading it? Many of you might have heard about INLA and have picked this book up because it has INLA in the title– and can teach you "How to use INLA" or what models can be fitted with. Others might have come across this book they are familiar with INLA, but would like to see a wide range of models that may be fitted with INLA. The things is– this book is not about INLA.

In this first chapter we illustrate what this book is about, by showing examples of a number of data structures that may be seen a typical special cases that the analysis methods discussed in this book can deal with. When reading this you may find that some of the data sets initially seem to be very different in structure. In this chapter we show the diversity and pointing out similarity.

INLA is a model fitting method – we will discuss it in detail below – and as such a means to an end. Spatial component that has not been traditionally pointed out/discussed.

1.2 Examples – data structures

Lets have a look at a few examples of data structures.

1.2.1 Spatially continuous data

Consider Figure 1.1 which shows the elevation in a rainforest study plot in Panama. In this data structure that takes on values everywhere in the plot; there is not a single location where it would not make sense to consider elevation as the quantity is spatially continuous. In practice, however, there are limitations. While there are there are infinitely many locations in even a small area of space the quantity of interest cannot

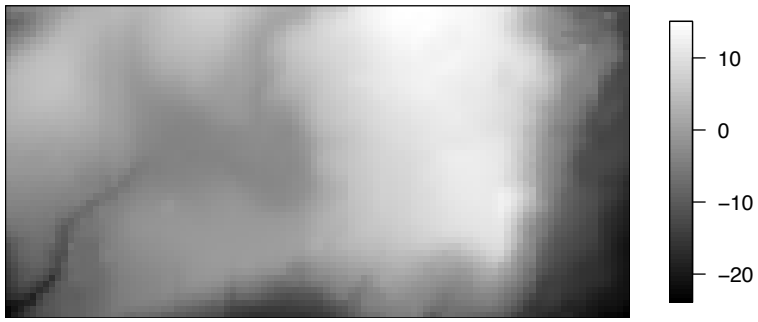

Figure 1.1: Some elevation somewhere...

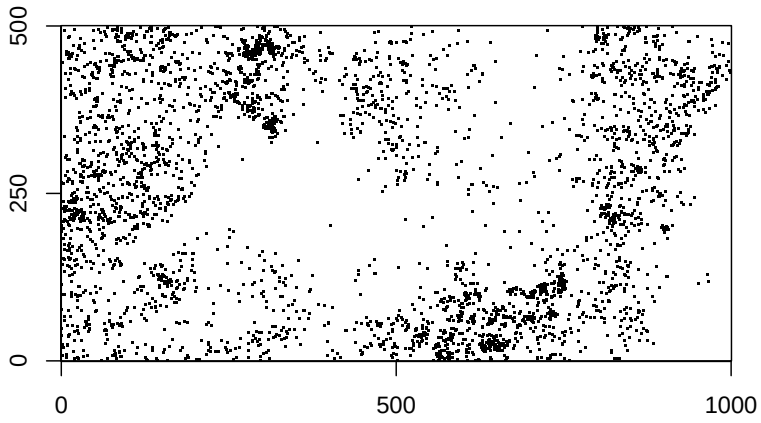

Figure 1.2: Rainforest trees...

be measured everywhere in space – and the quantity of interest cannot be fully observed. It can only be measured at a finite number of locations and these location have typically been chosen as part of the study design. There is an interest in interpolating to locations where no measurements have been taken – referred to as *spatial prediction*. This prediction into other areas of space needs to be done in a clever way, that take into account information from given measurements. In particular, on would like the predicted values to be similar to the values measured in the other locations. The figure is actually based on XX measurements in as many locations randomly distributed across the plot. There might also be an interest in relate one spatially continuous variable to other spatially continuous variables as part of a modelling exercise. These spatial covariates might help to explaining why values are high/low in different areas of the plot, but there might still be spatial structure left that needs to be accounted for.

This data structure is often referred to as *geo-referenced data* or *geostatistical data*.

1.2.2 Spatial point patterns

Figure 1.2 shows the same rainforest study plot as Figure 1.1, but this time the locations of trees of the species *Beilschmidia* have been plotted. A study might be interested in analysing the pattern formed by those trees to understand habitat preferences of the species. Hence, unlike in the previous example the locations have not been deliberately chosen as part of the study but are the object of interest. The main interest is to understand– and hence model the spatial pattern. Again, spatial covariates might explain the spatial pattern but there might still be spatial structure left that needs to be accounted for.

This data structure is referred to as a *spatial point pattern*.

1.2.3 Data collected on transects

1.2.4 Distance sampling data

1.3 Synergy – common spatial structure

Datasets collected in different ways – but the all have a spatial structure. Spatial structure is relevant, of interest or might impact on inference.

How do we represent this spatial structure? This is how

1.3.1 The random field – intuition

1.3.2 Issues with spatial data analysis

Computationally complex. Hence need to be computationally efficient.

Need to approximate.
Space is complex (sphere, holes) – need to be flexible. Grids are rigid...

1.4 What to expect from the book

Finding a rigorous yet flexible way of representing the spatial structure and linking this with computationally efficient model fitting strategies that allow us to fit relevant and realistically complex models.

1.5 Structure of the book

Next Chapter explores key concepts by referring to the different data structures we have seen here. It ends with a roadmap of this book.

Chapter after that formalises these concepts

Chapter 2

The key concepts

2.1 Introduction – random fields, a gentle introduction

We have seen in the first chapter that many data structures have one thing in common–they life in space where there is spatial autocorrelation. This autocorrelation is reflected in a random field.

A spatial random field is a *random variable* that represents spatially continuous phenomena in 2 dimensions. Since this is a difficult concept to grasp this chapter slowly builds up to these by first considering simple univariate random variables familiar from basic statistics courses, then moving to 1-dimensional random fields and eventually the 2-dimensional random fields that will be prominent throughout this book.

2.1.1 A simple univariate random variable

Most readers will have come across *random variables* in other contexts and much of what is discussed in this section will be familiar. Random variables are variables that are assumed to follow some probability distribution, i.e. they take on different values or values within a certain interval with different probabilities. A very familiar example is a random variable that follows a normal distribution. This one-dimensional continuous random variable can take on any value, but values close to the mean are particularly likely, resulting in the famous bell curve. As an example lets assume that IQ values in a country are normally distributed with mean 100 and standard deviation 15. The probability density of this distribution is plotted in Figure 2.1, black line indicating that values around the mean rather likely and those further away are less likely.

In practice, we usually we do not know the true distribution for a given population but use a *sample* to estimate the parameters of the distribution from this sample, for example using maximum likelihood estimation methods, as discussed in standard statistics textbooks. For the example in Figure 2.1 a sample of size 20 resulted in an estimate of mean 95.03 and standard deviation 12.67; the estimated density is shown as the red line in the same figure. However, a much bigger sample of size 200 provided a much better estimate with a mean of 99.84 and standard deviation 13.95, represented by the blue line in Figure 2.1. This improvement in precision is not surprising as we gain more *independent information* on a population from a sample of a bigger size and hence our estimates become more precise.

An important point here is that typical estimation and inference approaches assume that any sample consists of **independent observations** or **replicates**. Dependent samples are likely to tend to be similar and hence to not provide much new information on the population. This will eventually introduce bias. In practice, this implies that observations have to be made independently of each other, and the assumption is made that there are no similarities or relationships among any of the observations that might result from the sampling process. For instance, for the estimation of IQ values in a country one would have to sample across the population and not just from a specific subgroup of the population, say university students, as this would bias the estimation. In this case there is a similarity or relationship among the resulting observations as the observation have all be been made from members of the same subgroup. We will see

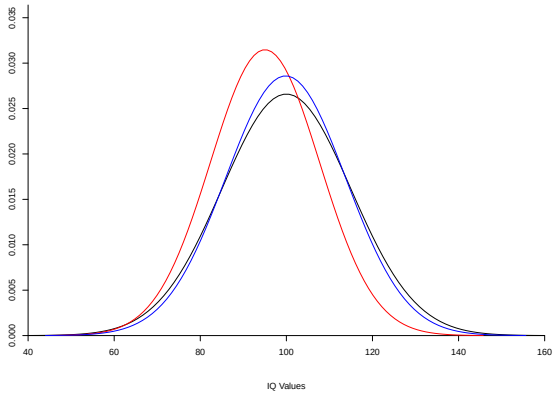

Figure 2.1: The probability density of the normal distribution with mean 100 and sd = 15 (black line), the density estimated from a sample of size 20 (red line) and the density estimated from a sample of size 200 (blue line).

further below that in the central concept of the book–in random fields there is no assumption of independence among measurements; they are explicitly assumed to show **systematic dependence** among measurements and specific assumption is made about the nature of this dependence. We will start thinking about this dependence in the next section. We will initially look at a one dimensional random field even though most of the random fields discussed in this book are spatial random fields, i.e. random fields in 2D.

2.1.2 A random field in 1D

Random field in !D is a random variable that the realisations of which take on values in one-dimensional space. Lets think about this through an example. Say, the 25 children in a primary school class in Helsinki are set to learn about temperatures, its measurement and how it fluctuates on their home town during a day in spring. In other words the random variable of interest is "temperature throughout school day in spring in Helsinki". Continuous in time. To estimate this random variable on one specific day in May, they the children are each given a different random time between 8:00 am and 4:00 pm by their teacher and are asked to check and write down the temperature at the thermometer in their classroom that measures the outside temperature at exactly that time.

When comparing these data to the IQ data we discussed earlier we realise a number of things. Clearly, these temperature measurements are *not independent*–it is likely that the measurements taken at as 12:45 and at 12:46 are not very different from each other. The measurements tell us about the temperature throughout a specific school day in one place in Helsinki, but they do not tell us much about the temperature and its fluctuations elsewhere in the country or the world. Even if each child was given two random times to measure the temperature and the number of observations was doubled, these would not improve our knowledge about the temperature levels and fluctuations that day elsewhere.

However, this is not what the children where meant to learn anyway –they were meant to get a picture of the behaviour of temperature throughout a day in spring in Helsinki. Now, they have only measured it at some points in time, not continuously through time. However, if one makes some general reasonable assumptions about how temperature behaves over time, it is possible to get an idea of the temperature at times when the temperature has not been measured, based on the given values. Intuitively, one would want to assume

a) that temperature values measured within a few minutes of each other are similar to each other,

b) that the temperature does not jump to some high value and back down frequently between measurements.

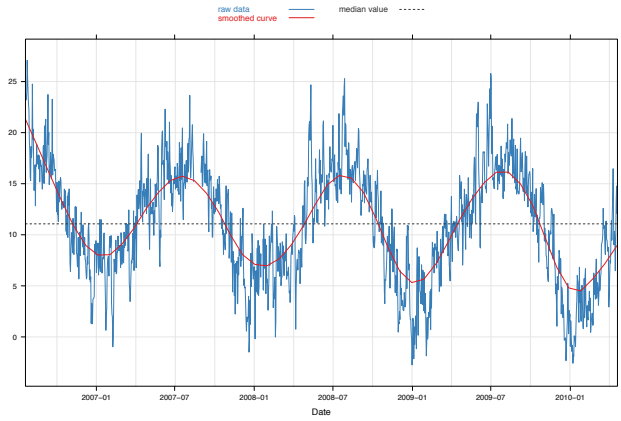

Figure 2.2: Temperature over time... this is a bad plot....

With these simple assumptions one could just connect the observations by straight line, as in Figure 2.2; doing this means we are simply taking the average between the values. But we know that temperature does not behave like this and that it is more smooth. Hence we would like some smooth function to describe the relationship between two observations–and this smooth function depends on distance between the observations and describes their similarity.

Note that the temperatures in this example have only been observed on one single day, while there is an interest in temperatures in Helsinki in spring. This is inplies Values would be different if different day was chosen. only one realisation.

While the school children in Helsinki could probably repeat the measurement procedure on several –randomly chosen– days throughout spring, in many realistic examples This has important consequences

Note that measured only at very few points and this realisation has not been fully observed

This is what a random field in 1D does, make assumptions on this functional form random variable

Note there is actually another set of assumptions at work here- one that concerns the measuring process.

c) that temperature is measured at different times during the day and year and month

d) and that the person does not get into their car more frequently when it is warmer outside than when it is cold.

[example does not work very well; fewer values during the night, but wanted non-regular measurements that one would get from some measurement device...]

With these assumption we implicitly make assumptions about the nature of the dependence among the observations.

- realise that observations are not independent through time
- prediction hence needs to take the other values into account
- interpolated values should, on average, be similar to all values measured
- interpolated values should be similar to values measured close in time

random field has properties that describe average behaviour and on distance in time

2.1.3 The same thing in 2D

Lets now move to 2 dimensional space.

key points:

- realise that observations are not independent through space
- interpolated values should, on average, be similar to all values measured
- interpolated values should be similar values close in space

2.2 Random fields – and how they appear everywhere

Lets have a look at a few examples of data structures.

2.2.1 Spatially continuous data as random fields

2.2.2 Spatial point processes, conditional on random fields

2.2.3 Transect data thinned spatial point processes, conditional on random fields

2.3 Random fields – a definition

2.4 Spatial point processes

2.5 Thinned point processes

2.5.1 Known thinning process

2.5.2 Unknown thinning process

Chapter 3

Key concept – spatial point prcesses

3.1 Spatial point processes

Model location of objects in space.

3.2 Thinned point processes

3.2.1 Known thinning process

3.2.2 Unknown thinning process