Spatial Analysis Notes

Francisco Rowe & Dani Arribas-Bel

2020-02-20

Contents

Chapter 1

Spatial Analysis Notes

This book contains computational illustrations on spatial analytical approaches using R.

Chapter 2

Introduction

This chapter[1] introduces R Notebooks, basic functions and data types. These are all important concepts that we will use during the module.

If you are already familiar with R, R notebooks and data types, you may want to jump to Section Read Data and start from there. This section describes how to read and manipulate data using `sf` and `tidyverse` functions, including `mutate()`, `%>%` (known as pipe operator), `select()`, `filter()` and specific packages and functions how to manipulate spatial data.

The chapter is based on:

- Grolemund and Wickham (2019), this book illustrates key libraries, including tidyverse, and functions for data manipulation in R
- Xie et al. (2019), excellent introduction to R markdown!
- Williamson (2018), some examples from the first lecture of ENVS450 are used to explain the various types of random variables.
- Lovelace et al. (2020), a really good book on handling spatial data and historical background of the evolution of R packages for spatial data analysis.

2.1 Dependencies

This tutorial uses the libraries below. Ensure they are installed on your machine[2] before loading them executing the following code chunk:

```
# Data manipulation, transformation and visualisation
library(tidyverse)
# Nice tables
library(kableExtra)
# Simple features (a standardised way to encode vector data ie. points, lines, polygon
library(sf)
# Spatial objects conversion
library(sp)
# Thematic maps
library(tmap)
# Colour palettes
library(RColorBrewer)
# More colour palettes
library(viridis)
```

2.2 Introducing R

R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques. It has gained widespread use in academia and industry. R offers a wider array of functionality than a traditional statistics package, such as SPSS and is composed of core (base) functionality, and is expandable through libraries hosted on CRAN. CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.

Commands are sent to R using either the terminal / command line or the R Console which is installed with R on either Windows or OS X. On Linux, there is no equivalent of the console, however, third party solutions exist. On your own machine, R can be installed from here.

Normally RStudio is used to implement R coding. RStudio is an integrated development environment (IDE) for R and provides a more user-friendly front-end to R than the front-end provided with R.

To run R or RStudio, just double click on the R or RStudio icon. Throughout this module, we will be using RStudio:

[2] You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

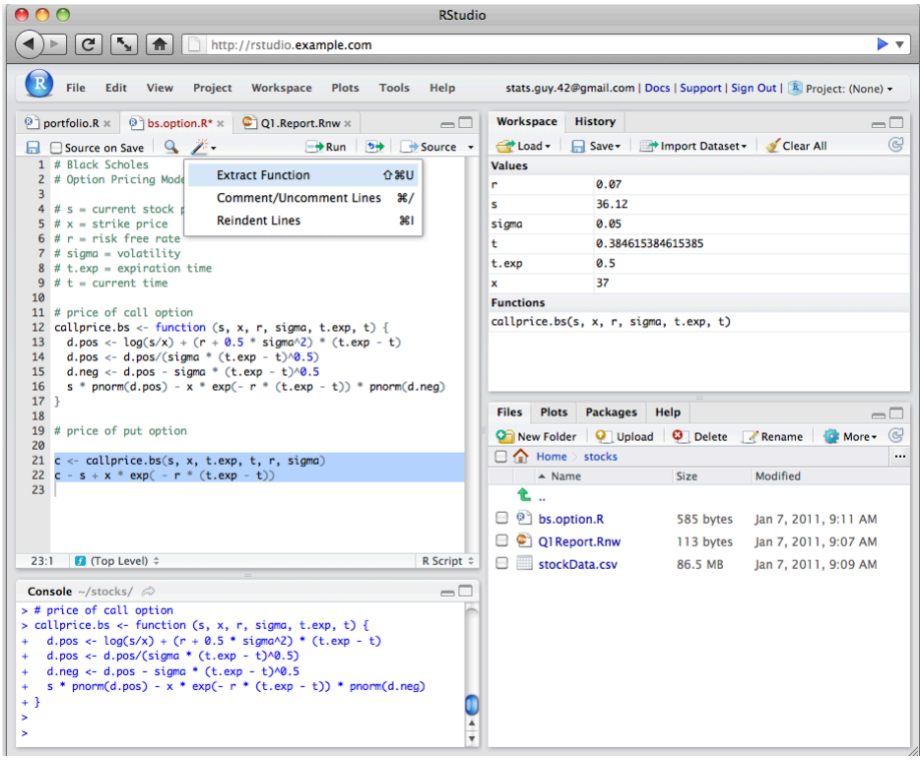


Figure 2.1: Fig. 1. RStudio features.

If you would like to know more about the various features of RStudio, watch this video

2.3 Setting the working directory

Before we start any analysis, ensure to set the path to the directory where we are working. We can easily do that with `setwd()`. Please replace in the following line the path to the folder where you have placed this file -and where the `data` folder lives.

```
#setwd('../data/sar.csv')
#setwd('.')
```

Note: It is good practice to not include spaces when naming folders and files. Use *underscores* or *dots*.

You can check your current working directory by typing:

```
getwd()
```

```
## [1] "/home/jovyan/work"
```

2.4 R Scripts and Notebooks

An *R script* is a series of commands that you can execute at one time and help you save time. So you don't repeat the same steps every time you want to execute the same process with different datasets. An R script is just a plain text file with R commands in it.

To create an R script in RStudio, you need to

- Open a new script file: *File* > *New File* > *R Script*
- Write some code on your new script window by typing eg. `mtcars`
- Run the script. Click anywhere on the line of code, then hit *Ctrl + Enter* (Windows) or *Cmd + Enter* (Mac) to run the command or select the code chunk and click *run* on the right-top corner of your script window. If do that, you should get:

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
```

```
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

- Save the script: *File* > *Save As*, select your required destination folder, and enter any filename that you like, provided that it ends with the file extension *.R*

An *R Notebook* is an R Markdown document with descriptive text and code chunks that can be executed independently and interactively, with output visible immediately beneath a code chunk - see Xie et al. (2019).

To create an R Notebook, you need to:

- Open a new script file: *File* > *New File* > *R Notebook*
- Insert code chunks, either:

1) use the *Insert* command on the editor toolbar;
2) use the keyboard shortcut *Ctrl + Alt + I* or *Cmd + Option + I* (Mac); or,
3) type the chunk delimiters ```` ```{r} ```` and ```` ``` ````

In a chunk code you can produce text output, tables, graphics and write code! You can control these outputs via chunk options which are provided inside the

```
---
title: "My Notebook"
output: html_notebook
---
```

Figure 2.2: Fig. 2. YAML metadata for notebooks.

curly brackets eg.

- Execute code: hit *"Run Current Chunk"*, *Ctrl + Shift + Enter* or *Cmd + Shift + Enter* (Mac)
- Save an R notebook: *File > Save As*. A notebook has a `*.Rmd` extension and when it is saved a `*.nb.html` file is automatically created. The latter is a self-contained HTML file which contains both a rendered copy of the notebook with all current chunk outputs and a copy of the *.Rmd file itself.

Rstudio also offers a *Preview* option on the toolbar which can be used to create pdf, html and word versions of the notebook. To do this, choose from the drop-down list menu `knit to ...`

2.5 Getting Help

You can use `help` or `?` to ask for details for a specific function:

```
help(sqrt) #or ?sqrt
```

And using `example` provides examples for said function:

```
example(sqrt)
```

```
##
## sqrt> require(stats) # for spline
##
## sqrt> require(graphics)
##
## sqrt> xx <- -9:9
##
## sqrt> plot(xx, sqrt(abs(xx)),  col = "red")
##
## sqrt> lines(spline(xx, sqrt(abs(xx)), n=101), col = "pink")
```

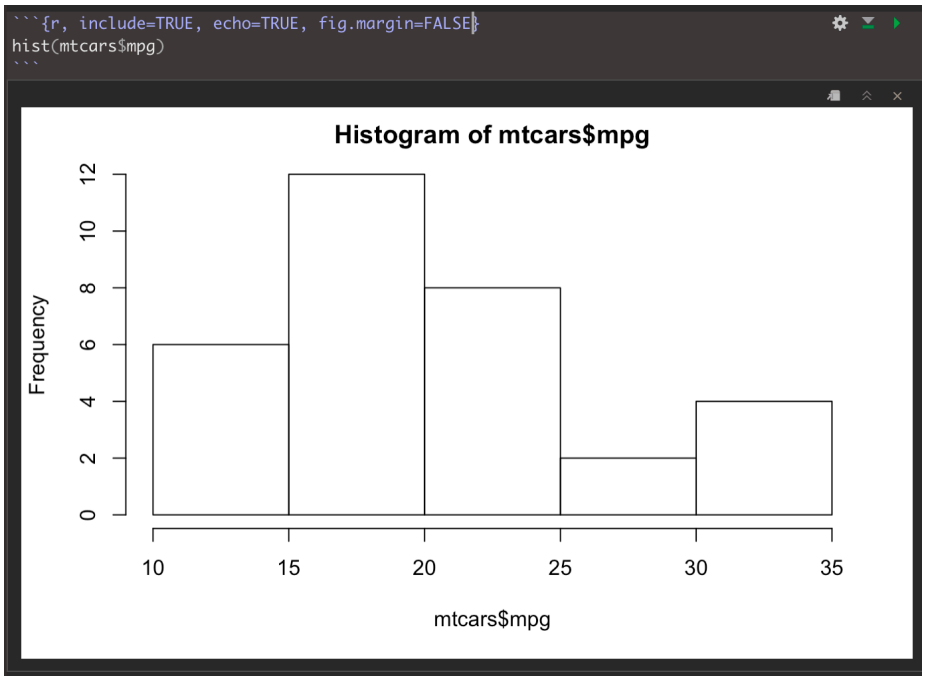


Figure 2.3: Fig. 3. Code chunk example. Details on the various options: https://rmarkdown.rstudio.com/lesson-3.html

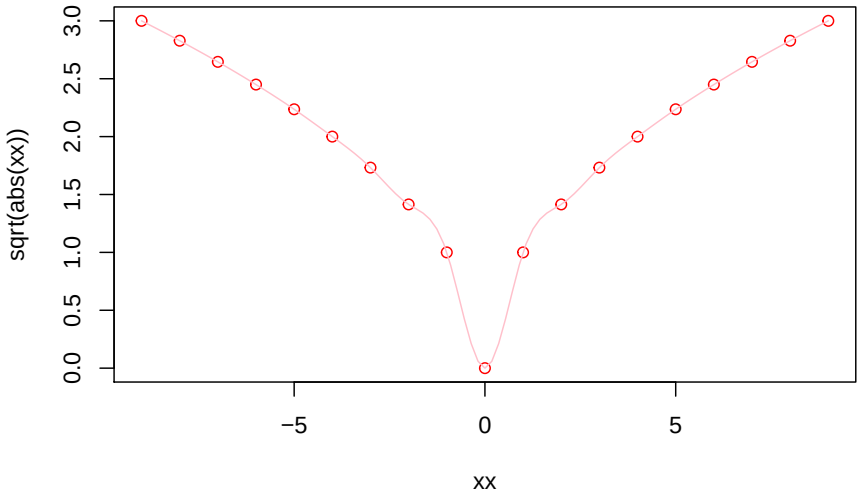


Figure 2.4: Example sqrt

2.6 Variables and objects

An *object* is a data structure having attributes and methods. In fact, everything in R is an object!

A *variable* is a type of data object. Data objects also include list, vector, matrices and text.

- Creating a data object

In R a variable can be created by using the symbol `<-` to assign a value to a variable name. The variable name is entered on the left `<-` and the value on the right. Note: Data objects can be given any name, provided that they start with a letter of the alphabet, and include only letters of the alphabet, numbers and the characters `.` and `_`. Hence AgeGroup, Age_Group and Age.Group are all valid names for an R data object. Note also that R is case-sensitive, to agegroup and AgeGroup would be treated as different data objects.

To save the value *28* to a variable (data object) labelled *age*, run the code:

```
age <- 28
```

- Inspecting a data object

To inspect the contents of the data object *age* run the following line of code:

```
age
```

```
## [1] 28
```

Find out what kind (class) of data object *age* is using:

```
class(age)
```

```
## [1] "numeric"
```

Inspect the structure of the *age* data object:

```
str(age)
```

```
##  num 28
```

- The *vector* data object

What if we have more than one response? We can use the `c( )` function to combine multiple values into one data vector object:

```
age <- c(28, 36, 25, 24, 32)
age
```

```
## [1] 28 36 25 24 32
```

```
class(age) #Still numeric..
```

```
## [1] "numeric"
```

```
str(age) #..but now a vector (set) of 5 separate values
```

```
##  num [1:5] 28 36 25 24 32
```

Note that on each line in the code above any text following the `#` character is ignored by R when executing the code. Instead, text following a `#` can be used to add comments to the code to make clear what the code is doing. Two marks of good code are a clear layout and clear commentary on the code.

2.6.1 Basic Data Types

There are a number of data types. Four are the most common. In R, **numeric** is the default type for numbers. It stores all numbers as floating-point numbers (numbers with decimals). This is because most statistical calculations deal with numbers with up to two decimals.

- Numeric

```
num <- 4.5 # Decimal values
class(num)
```

```
## [1] "numeric"
```

- Integer

```
int <- as.integer(4) # Natural numbers. Note integers are also numerics.
class(int)
```

```
## [1] "integer"
```

- Character

```
cha <- "are you enjoying this?" # text or string. You can also type `as.character("are
class(cha)
```

```
## [1] "character"
```

- Logical

```
log <- 2 < 1 # assigns TRUE or FALSE. In this case, FALSE as 2 is greater than 1
log
```

```
## [1] FALSE
```

```
class(log)
```

```
## [1] "logical"
```

2.6.2 Random Variables

In statistics, we differentiate between data to capture:

- *Qualitative attributes* categorise objects eg.gender, marital status. To measure these attributes, we use *Categorical* data which can be divided into:
  - *Nominal* data in categories that have no inherent order eg. gender
  - *Ordinal* data in categories that have an inherent order eg. income bands
- *Quantitative attributes*:
  - *Discrete* data: count objects of a certain category eg. number of kids, cars
  - *Continuous* data: precise numeric measures eg. weight, income, length.

In R these three types of random variables are represented by the following types of R data object:

variables	objects
nominal	factor
ordinal	ordered factor
discrete	numeric
continuous	numeric

We have already encountered the R data type *numeric*. The next section introduces the *factor* data type.

2.6.2.1 Factor

What is a factor?

A factor variable assigns a numeric code to each possible category (*level*) in a variable. Behind the scenes, R stores the variable using these numeric codes to save space and speed up computing. For example, compare the size of a list of `10,000` *males* and *females* to a list of `10,000` `1s` and `0s`. At the same time R also saves the category names associated with each numeric code (level). These are used for display purposes.

For example, the variable *gender*, converted to a factor, would be stored as a series of `1s` and `2s`, where `1 = female` and `2 = male`; but would be displayed in all outputs using their category labels of *female* and *male*.

Creating a factor

To convert a numeric or character vector into a factor use the `factor( )` function. For instance:

```
gender <- c("female","male","male","female","female") # create a gender variable
gender <- factor(gender) # replace character vector with a factor version
gender
```

```
## [1] female male   male   female female
## Levels: female male
```

```
class(gender)
```

```
## [1] "factor"
```

```
str(gender)
```

```
##  Factor w/ 2 levels "female","male": 1 2 2 1 1
```

Now *gender* is a factor and is stored as a series of `1s` and `2s`, with `1s` representing `females` and `2s` representing `males`. The function `levels( )` lists the levels (categories) associated with a given factor variable:

```
levels(gender)
```

```
## [1] "female" "male"
```

The categories are reported in the order that they have been numbered (starting from 1). Hence from the output we can infer that `females` are coded as 1, and `males` as 2.

2.7 Data Frames

R stores different types of data using different types of data structure. Data are normally stored as a *data.frame*. A data frames contain one row per observation (e.g. wards) and one column per attribute (eg. population and health).

We create three variables wards, population (`pop`) and people with good health (`ghealth`). We use 2011 census data counts for total population and good health for wards in Liverpool.

```
wards <- c("Allerton and Hunts Cross","Anfield","Belle Vale","Central","Childwall","Chu

pop <- c(14853,14510,15004,20340,13908,13974,15272,14045,14503,
                14561,14782,16786,16132,15377,16115,13312,13816,15047,
                16461,17009,17104,18422,12991,20300,16489,16481,14772,
                14382,12921,16746)

ghealth <- c(7274,6124,6129,11925,7219,7461,6403,5930,7094,6992,
                 5517,7879,8990,6495,6662,5981,7322,6529,7192,7953,
                 7636,9001,6450,8973,7302,7521,7268,7013,6025,7717)
```

Note that `pop` and `ghealth` and `wards` contains characters.

2.7.1 Creating A Data Frame

We can create a data frame and examine its structure:

```
df <- data.frame(wards, pop, ghealth)
df # or use view(data)
```

```
##                       wards   pop ghealth
## 1  Allerton and Hunts Cross 14853    7274
## 2                   Anfield 14510    6124
## 3                Belle Vale 15004    6129
## 4                   Central 20340   11925
## 5                 Childwall 13908    7219
## 6                    Church 13974    7461
## 7                  Clubmoor 15272    6403
## 8                    County 14045    5930
## 9                Cressington 14503    7094
## 10                  Croxteth 14561    6992
```

```
## 11                  Everton 14782    5517
## 12               Fazakerley 16786    7879
## 13                Greenbank 16132    8990
## 14 Kensington and Fairfield 15377    6495
## 15                 Kirkdale 16115    6662
## 16               Knotty Ash 13312    5981
## 17             Mossley Hill 13816    7322
## 18             Norris Green 15047    6529
## 19                 Old Swan 16461    7192
## 20                   Picton 17009    7953
## 21             Princes Park 17104    7636
## 22                Riverside 18422    9001
## 23             St Michael's 12991    6450
## 24            Speke-Garston 20300    8973
## 25 Tuebrook and Stoneycroft 16489    7302
## 26                 Warbreck 16481    7521
## 27                Wavertree 14772    7268
## 28               West Derby 14382    7013
## 29                  Woolton 12921    6025
## 30                  Yew Tree 16746    7717
```

```
str(df) # or use glimpse(data)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ wards  : Factor w/ 30 levels "Allerton and Hunts Cross",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ pop    : num  14853 14510 15004 20340 13908 ...
##  $ ghealth: num  7274 6124 6129 11925 7219 ...
```

2.7.2 Referencing Data Frames

Throughout this module, you will need to refer to particular parts of a dataframe - perhaps a particular column (an area attribute); or a particular subset of respondents. Hence it is worth spending some time now mastering this particular skill.

The relevant R function, `[ ]`, has the format `[row,col]` or, more generally, `[set of rows, set of cols]`.

Run the following commands to get a feel of how to extract different slices of the data:

```
df # whole data.frame
df[1, 1] # contents of first row and column
df[2, 2:3] # contents of the second row, second and third columns
df[1, ] # first row, ALL columns [the default if no columns specified]
df[ ,1:2] # ALL rows; first and second columns
df[c(1,3,5), ] # rows 1,3,5; ALL columns
```

```
df[ , 2] # ALL rows; second column (by default results containing only
              #one column are converted back into a vector)
df[ , 2, drop=FALSE] # ALL rows; second column (returned as a data.frame)
```

In the above, note that we have used two other R functions:

- `1:3` The colon operator tells R to produce a list of numbers including the named start and end points.
- `c(1,3,5)` Tells R to combine the contents within the brackets into one list of objects

Run both of these fuctions on their own to get a better understanding of what they do.

Three other methods for referencing the contents of a data.frame make direct use of the variable names within the data.frame, which tends to make for easier to read/understand code:

```
df[,"pop"] # variable name in quotes inside the square brackets
df$pop # variable name prefixed with $ and appended to the data.frame name
# or you can use attach
attach(df)
pop # but be careful if you already have an age variable in your local workspace
```

Want to check the variables available, use the `names( )`:

```
names(df)
```

```
## [1] "wards"   "pop"     "ghealth"
```

2.8 Read Data

Ensure your memory is clear

```
rm(list=ls()) # rm for targeted deletion / ls for listing all existing objects
```

There are many commands to read / load data onto R. The command to use will depend upon the format they have been saved. Normally they are saved in *csv* format from Excel or other software packages. So we use either:

- `df <- read.table("path/file_name.csv", header = FALSE, sep =",")`
- `df <- read("path/file_name.csv", header = FALSE)`
- `df <- read.csv2("path/file_name.csv", header = FALSE)`

To read files in other formats, refer to this useful DataCamp tutorial

```
census <- read.csv("data/census/census_data.csv")
head(census)
```

```
##        code                     ward pop16_74 higher_managerial   pop
## 1 E05000886 Allerton and Hunts Cross    10930              1103 14853
## 2 E05000887                  Anfield    10712               312 14510
## 3 E05000888                Belle Vale    10987               432 15004
## 4 E05000889                   Central    19174              1346 20340
## 5 E05000890                 Childwall    10410              1123 13908
## 6 E05000891                    Church    10569              1843 13974
##   ghealth
## 1    7274
## 2    6124
## 3    6129
## 4   11925
## 5    7219
## 6    7461
```

```
# NOTE: always ensure your are setting the correct directory leading to the data.
# It may differ from your existing working directory
```

2.8.1 Quickly inspect the data

1. What class?
2. What R data types?
3. What data types?

```
# 1
class(census)
# 2 & 3
str(census)
```

Just interested in the variable names:

```
names(census)
```

```
## [1] "code"              "ward"              "pop16_74"
## [4] "higher_managerial" "pop"               "ghealth"
```

or want to view the data:

```
View(census)
```

2.9 Manipulation Data

2.9.1 Adding New Variables

Usually you want to add / create new variables to your data frame using existing variables eg. computing percentages by dividing two variables. There are many ways in which you can do this i.e. referecing a data frame as we have done above, or using `$` (e.g. `census$pop`). For this module, we'll use `tidyverse`:

```
census <- census %>% mutate(per_ghealth = ghealth / pop)
```

Note we used a *pipe operator* `%>%`, which helps make the code more efficient and readable - more details, see Grolemund and Wickham (2019). When using the pipe operator, recall to first indicate the data frame before `%>%`.

Note also the use a variable name before the `=` sign in brackets to indicate the name of the new variable after `mutate`.

2.9.2 Selecting Variables

Usually you want to select a subset of variables for your analysis as storing to large data sets in your R memory can reduce the processing speed of your machine. A selection of data can be achieved by using the `select` function:

```
ndf <- census %>% select(ward, pop16_74, per_ghealth)
```

Again first indicate the data frame and then the variable you want to select to build a new data frame. Note the code chunk above has created a new data frame called `ndf`. Explore it.

2.9.3 Filtering Data

You may also want to filter values based on defined conditions. You may want to filter observations greater than a certain threshold or only areas within a certain region. For example, you may want to select areas with a percentage of good health population over 50%:

```
ndf2 <- census %>% filter(per_ghealth < 0.5)
```

You can use more than one variables to set conditions. Use "`,`" to add a condition.

2.9.4 Joining Data Drames

When working with spatial data, we often need to join data. To this end, you need a common unique `id variable`. Let's say, we want to add a data frame containing census data on households for Liverpool, and join the new attributes to one of the existing data frames in the workspace. First we will read the data frame we want to join (ie. `census_data2.csv`).

```
# read data
census2 <- read.csv("data/census/census_data2.csv")
# visualise data structure
str(census2)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ geo_code              : Factor w/ 30 levels "E05000886","E05000887",..: 1 2 3 4 5 6 7 8 9
##  $ households            : int  6359 6622 6622 7139 5391 5884 6576 6745 6317 6024 ...
##  $ socialrented_households: int  827 1508 2818 1311 374 178 2859 1564 1023 1558 ...
```

The variable `geo_code` in this data frame corresponds to the `code` in the existing data frame and they are unique so they can be automatically matched by using the `merge()` function. The `merge()` function uses two arguments: `x` and `y`. The former refers to data frame 1 and the latter to data frame 2. Both of these two data frames must have a `id` variable containing the same information. Note they can have different names. Another key argument to include is `all.x=TRUE` which tells the function to keep all the records in `x`, but only those in `y` that match in case there are discrepancies in the `id` variable.

```
# join data frames
join_dfs <- merge(census, census2, by.x="code", by.y="geo_code", all.x = TRUE)
# check data
head(join_dfs)
```

```
##        code                      ward pop16_74 higher_managerial   pop
## 1 E05000886 Allerton and Hunts Cross    10930              1103 14853
## 2 E05000887                  Anfield    10712               312 14510
## 3 E05000888               Belle Vale    10987               432 15004
## 4 E05000889                  Central    19174              1346 20340
## 5 E05000890                Childwall    10410              1123 13908
## 6 E05000891                   Church    10569              1843 13974
##   ghealth per_ghealth households socialrented_households
## 1    7274   0.4897327       6359                     827
## 2    6124   0.4220538       6622                    1508
## 3    6129   0.4084911       6622                    2818
## 4   11925   0.5862832       7139                    1311
## 5    7219   0.5190538       5391                     374
## 6    7461   0.5339201       5884                     178
```

2.9.5 Saving Data

It may also be convinient to save your R projects. They contains all the objects that you have created in your workspace by using the `save.image( )` function:

```
save.image("week1_envs453.RData")
```

This creates a file labelled "week1_envs453.RData" in your working directory. You can load this at a later stage using the `load( )` function.

```
load("week1_envs453.RData")
```

Alternatively you can save / export your data into a `csv` file. The first argument in the function is the object name, and the second: the name of the csv we want to create.

```
write.csv(join_dfs, "join_censusdfs.csv")
```

2.10 Using Spatial Data Frames

A core area of this module is learning to work with spatial data in R. R has various purposedly designed **packages** for manipulation of spatial data and spatial analysis techniques. Various R packages exist in CRAN eg. `spatial`, `sgeostat`, `splancs`, `maptools`, `tmap`, `rgdal`, `spand` and more recent development of `sf` - see Lovelace et al. (2020) for a great description and historical context for some of these packages.

During this session, we will use `sf`.

We first need to import our spatial data. We will use a shapefile containing data at Output Area (OA) level for Liverpool. These data illustrates the hierarchical structure of spatial data.

2.10.1 Read Spatial Data

```
oa_shp <- st_read("data/census/Liverpool_OA.shp")
```

```
## Reading layer `Liverpool_OA' from data source `/home/jovyan/work/data/census/Liverpo
## Simple feature collection with 1584 features and 18 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 332390.2 ymin: 379748.5 xmax: 345636 ymax: 397980.1
## epsg (SRID):    NA
## proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-1
```

Examine the input data. A spatial data frame stores a range of attributes derived from a shapefile including the **geometry** of features (e.g. polygon shape and location), **attributes** for each feature (stored in the .dbf), projection and coordinates of the shapefile's bounding box - for details, execute:

```
?st_read
```

You can employ the usual functions to visualise the content of the created data frame:

```
# visualise variable names
names(oa_shp)
```

```
##  [1] "OA_CD"    "LSOA_CD"  "MSOA_CD"  "LAD_CD"   "pop"      "H_Vbad"
##  [7] "H_bad"    "H_fair"   "H_good"   "H_Vgood"  "age_men"  "age_med"
## [13] "age_60"   "S_Rent"   "Ethnic"   "illness"  "unemp"    "males"
## [19] "geometry"
```

```
# data structure
str(oa_shp)
```

```
## Classes 'sf' and 'data.frame':   1584 obs. of  19 variables:
##  $ OA_CD   : Factor w/ 1584 levels "E00032987","E00032988",..: 1547 485 143 1567 980 1186 1122
##  $ LSOA_CD : Factor w/ 298 levels "E01006512","E01006513",..: 291 96 33 124 187 231 220 175 17
##  $ MSOA_CD : Factor w/ 61 levels "E02001347","E02001348",..: 59 12 19 23 29 20 31 14 30 10 ...
##  $ LAD_CD  : Factor w/ 1 level "E08000012": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pop     : int  185 281 208 200 321 187 395 320 316 214 ...
##  $ H_Vbad  : int  1 2 3 7 4 4 5 9 5 4 ...
##  $ H_bad   : int  2 20 10 8 10 25 19 22 25 17 ...
##  $ H_fair  : int  9 47 22 17 32 70 42 53 55 39 ...
##  $ H_good  : int  53 111 71 52 112 57 131 104 104 53 ...
##  $ H_Vgood : int  120 101 102 116 163 31 198 132 127 101 ...
##  $ age_men : num  27.9 37.7 37.1 33.7 34.2 ...
##  $ age_med : num  25 36 32 29 34 53 23 30 34 29 ...
##  $ age_60  : num  0.0108 0.1637 0.1971 0.1 0.1402 ...
##  $ S_Rent  : num  0.0526 0.176 0.0235 0.2222 0.0222 ...
##  $ Ethnic  : num  0.3514 0.0463 0.0192 0.215 0.0779 ...
##  $ illness : int  185 281 208 200 321 187 395 320 316 214 ...
##  $ unemp   : num  0.0438 0.121 0.1121 0.036 0.0743 ...
##  $ males   : int  122 128 95 120 158 123 207 164 157 94 ...
##  $ geometry:sfc_MULTIPOLYGON of length 1584; first list element: List of 1
##   ..$ :List of 1
##   .. ..$ : num [1:14, 1:2] 335106 335130 335164 335173 335185 ...
##   ..- attr(*, "class")= chr  "XY" "MULTIPOLYGON" "sfg"
##  - attr(*, "sf_column")= chr "geometry"
##  - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",..: NA NA NA NA NA NA NA NA NA NA
##   ..- attr(*, "names")= chr  "OA_CD" "LSOA_CD" "MSOA_CD" "LAD_CD" ...
```

```
# see first few observations
head(oa_shp)
```

```
## Simple feature collection with 6 features and 18 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 335071.6 ymin: 389876.7 xmax: 339426.9 ymax: 394479
## epsg (SRID):    NA
## proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-1
##       OA_CD   LSOA_CD   MSOA_CD    LAD_CD pop H_Vbad H_bad H_fair H_good
## 1 E00176737 E01033761 E02006932 E08000012 185      1     2      9     53
## 2 E00033515 E01006614 E02001358 E08000012 281      2    20     47    111
## 3 E00033141 E01006546 E02001365 E08000012 208      3    10     22     71
## 4 E00176757 E01006646 E02001369 E08000012 200      7     8     17     52
## 5 E00034050 E01006712 E02001375 E08000012 321      4    10     32    112
## 6 E00034280 E01006761 E02001366 E08000012 187      4    25     70     57
##   H_Vgood  age_men age_med     age_60     S_Rent     Ethnic illness
## 1     120 27.94054      25 0.01081081 0.05263158 0.35135135     185
## 2     101 37.71174      36 0.16370107 0.17600000 0.04626335     281
## 3     102 37.08173      32 0.19711538 0.02352941 0.01923077     208
## 4     116 33.73000      29 0.10000000 0.22222222 0.21500000     200
## 5     163 34.19003      34 0.14018692 0.02222222 0.07788162     321
## 6      31 56.09091      53 0.44919786 0.88524590 0.11764706     187
##        unemp males                       geometry
## 1 0.04379562   122 MULTIPOLYGON (((335106.3 38...
## 2 0.12101911   128 MULTIPOLYGON (((335810.5 39...
## 3 0.11214953    95 MULTIPOLYGON (((336738 3931...
## 4 0.03597122   120 MULTIPOLYGON (((335914.5 39...
## 5 0.07428571   158 MULTIPOLYGON (((339325 3914...
## 6 0.44615385   123 MULTIPOLYGON (((338198.1 39...
```

TASK:

- What are the geographical hierarchy in these data?
- What is the smallest geography?
- What is the largest geography?

2.10.2 Basic Mapping

Again, many functions exist in CRAN for creating maps:

- `plot` to create static maps
- `tmap` to create static and interactive maps
- `leaflet` to create interactive maps
- `mapview` to create interactive maps
- `ggplot2` to create data visualisations, including static maps

Figure 2.5: OAs of Livepool

- `shiny` to create web applications, including maps

Here this notebook demonstrates the use of `plot` and `tmap`. First `plot` is used to map the spatial distribution of non-British-born population in Liverpool. First we only map the geometries on the right,

2.10.2.1 Using `plot`

```
# mapping geometry
plot(st_geometry(oa_shp))
```

and then:

```
# map attributes, adding intervals
plot(oa_shp["Ethnic"], key.pos = 4, axes = TRUE, key.width = lcm(1.3), key.length = 1.,
     breaks = "jenks", lwd = 0.1, border = 'grey')
```

TASK:

- What is the key pattern emerging from this map?

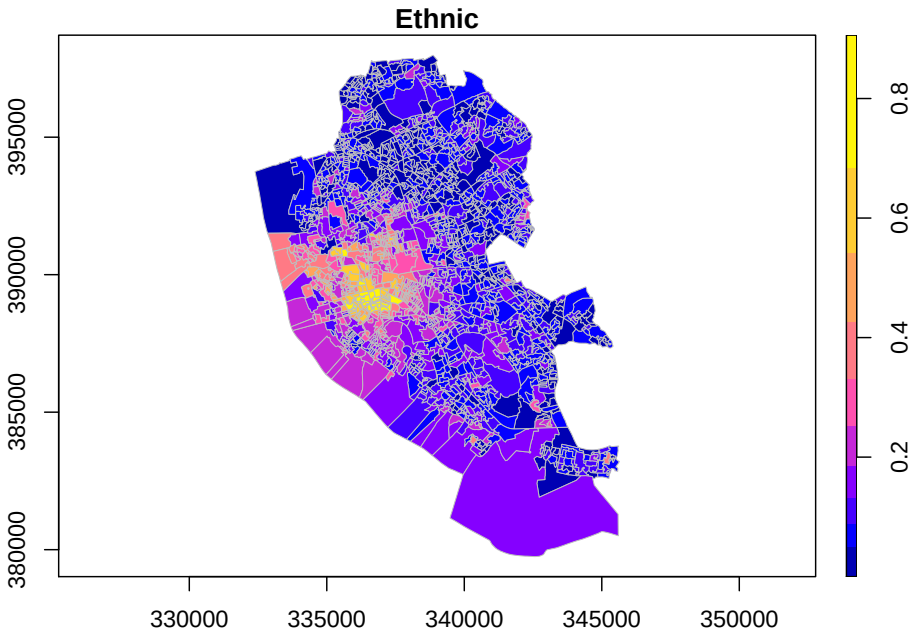


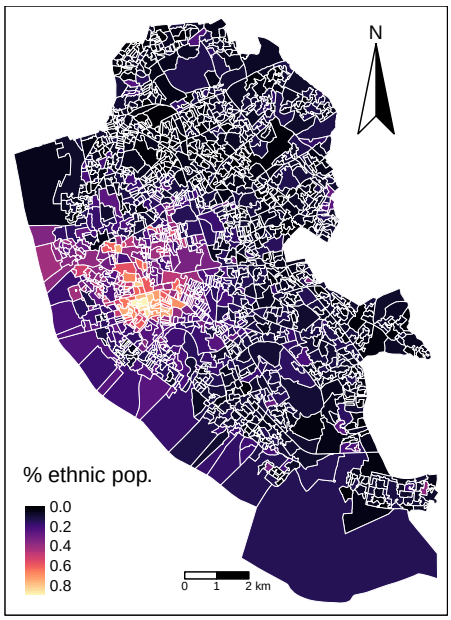
Figure 2.6: Spatial distribution of ethnic groups, Liverpool

2.10.2.2 Using `tmap`

Similar to `ggplot2`, `tmap` is based on the idea of a 'grammar of graphics' which involves a separation between the input data and aesthetics (i.e. the way data are visualised). Each data set can be mapped in various different ways, including location as defined by its geometry, colour and other features. The basic building block is `tm_shape()` (which defines input data), followed by one or more layer elements such as `tm_fill()` and `tm_dots()`.

```
# ensure geometry is valid
oa_shp = lwgeom::st_make_valid(oa_shp)

# map
legend_title = expression("% ethnic pop.")
map_oa = tm_shape(oa_shp) +
  tm_fill(col = "Ethnic", title = legend_title, palette = magma(256), style = "cont") +
  tm_borders(col = "white", lwd = .01)  + # add borders
  tm_compass(type = "arrow", position = c("right", "top") , size = 4) + # add compass
  tm_scale_bar(breaks = c(0,1,2), text.size = 0.5, position =  c("center", "bottom")) +
map_oa
```



Note that the operation + is used to add new layers. You can set style themes by `tm_style`. To visualise the existing styles use `tmap_style_catalogue()`, and you can also evaluate the code chunk below if you would like to create an interactive map.

```
tmap_mode("view")
map_oa
```

TASK:

- Try mapping other variables in the spatial data frame. Where do population aged 60 and over concentrate?

2.10.3 Comparing geographies

If you recall, one of the key issues of working with spatial data is the modifiable area unit problem (MAUP) - see lecture notes. To get a sense of the effects of MAUP, we analyse differences in the spatial patterns of the ethnic population in Liverpool between Middle Layer Super Output Areas (MSOAs) and OAs. So we map these geographies together.

```
# read data at the msoa level
msoa_shp <- st_read("data/census/Liverpool_MSOA.shp")
```

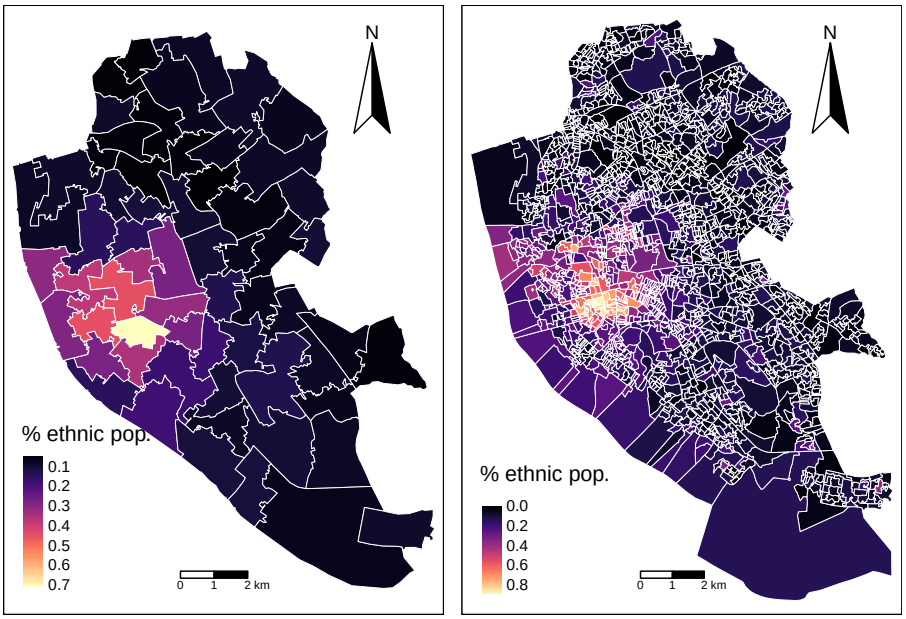
```
## Reading layer `Liverpool_MSOA' from data source `/home/jovyan/work/data/census/Liverpool_MSOA.
## Simple feature collection with 61 features and 16 fields
```

```
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 333086.1 ymin: 381426.3 xmax: 345636 ymax: 397980.1
## epsg (SRID):    NA
## proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-10
```

```r
# ensure geometry is valid
msoa_shp = lwgeom::st_make_valid(msoa_shp)

# create a map
map_msoa = tm_shape(msoa_shp) +
  tm_fill(col = "Ethnic", title = legend_title, palette = magma(256), style = "cont") +
  tm_borders(col = "white", lwd = .01)  +
  tm_compass(type = "arrow", position = c("right", "top") , size = 4) +
  tm_scale_bar(breaks = c(0,1,2), text.size = 0.5, position =  c("center", "bottom"))

# arrange maps
tmap_arrange(map_msoa, map_oa)
```



TASK:

- What differences do you see between OAs and MSOAs?
- Can you identify areas of spatial clustering? Where are they?

2.11 Useful Functions

Function	Description
read.csv()	read csv files into data frames
str()	inspect data structure
mutate()	create a new variable
filter()	filter observations based on variable values
%>%	pipe operator - chain operations
select()	select variables
merge()	join dat frames
st_read	read spatial data (ie. shapefiles)
plot()	create a map based a spatial data set
tm_shape(), tm_fill(), tm_borders()	create a map using tmap functions
tm_arrange	display multiple maps in a single “metaplot”

Chapter 3

Points

This chapter[1] is based on the following references, which are great follow-up's on the topic:

- Lovelace and Cheshire (2014) is a great introduction.
- Chapter 6 of Brunsdon and Comber (2015), in particular subsections 6.3 and 6.7.
- Bivand et al. (2013) provides an in-depth treatment of spatial data in R.

This chapter is part of Spatial Analysis Notes, a compilation hosted as a GitHub repository that you can access it in a few ways:

- As a download of a `.zip` file that contains all the materials.
- As an html website.
- As a pdf document
- As a GitHub repository.

3.1 Dependencies

This tutorial relies on the following libraries that you will need to have installed on your machine to be able to interactively follow along[2]. Once installed, load them up with the following commands:

```
# Layout
library(tufte)
```

[1] This chapter is part of Spatial Analysis Notes Points – Kernel Density Estimation and Spatial interpolation by Dani Arribas-Bel is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

[2] You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

```
# For pretty table
library(knitr)
# Spatial Data management
library(rgdal)
```

```
## Loading required package: sp
```

```
## rgdal: version: 1.4-4, (SVN revision 833)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: /usr/share/gdal/2.2
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: (autodetected)
##  Linking to sp version: 1.3-1
```

```
# Pretty graphics
library(ggplot2)
# Thematic maps
library(tmap)
# Pretty maps
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
# Various GIS utilities
library(GISTools)
```

```
## Loading required package: maptools
```

```
## Checking rgeos availability: TRUE
```

```
## Loading required package: RColorBrewer
```

```
## Loading required package: MASS
```

```
## Loading required package: rgeos
```

```
## rgeos version: 0.5-1, (SVN revision 614)
##  GEOS runtime version: 3.6.2-CAPI-1.10.2
##  Linking to sp version: 1.3-1
##  Polygon checking: TRUE
```

```
# For all your interpolation needs
library(gstat)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
# For data manipulation
library(plyr)
```

Before we start any analysis, let us set the path to the directory where we are working. We can easily do that with `setwd()`. Please replace in the following line the path to the folder where you have placed this file -and where the `house_transactions` folder with the data lives.

```
#setwd('/media/dani/baul/AAA/Documents/teaching/u-lvl/2016/envs453/code')
setwd('.')
```

3.2 Data

For this session, we will use a subset of residential property transaction data for the city of Liverpool. These are provided by the Land Registry (as part of their Price Paid Data) but have been cleaned and re-packaged by Dani Arribas-Bel.

Let us start by reading the data, which comes in a shapefile:

```
db <- readOGR(dsn = 'data/house_transactions', layer = 'liv_house_trans')
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/home/jovyan/work/data/house_transactions", layer: "liv_house_trans"
## with 6324 features
## It has 18 fields
## Integer64 fields read as strings:  price
```

Before we forget, let us make sure `price` is considered a number, not a factor:

```
db@data$price <- as.numeric(as.character((db@data$price)))
```

The dataset spans the year 2014:

```
# Format dates
dts <- as.Date(db@data$trans_date)
# Set up summary table
tab <- summary(dts)
tab
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2014-01-02" "2014-04-11" "2014-07-09" "2014-07-08" "2014-10-03"
##         Max.
## "2014-12-30"
```

We can then examine the elements of the object with the `summary` method:

```
summary(db)
```

```
## Object of class SpatialPointsDataFrame
```

```
## Coordinates:
##              min    max
## coords.x1 333536 345449
## coords.x2 382684 397833
## Is projected: TRUE
## proj4string :
## [+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000
## +y_0=-100000 +datum=OSGB36 +units=m +no_defs +ellps=airy
## +towgs84=446.448,-125.157,542.060,0.1502,0.2470,0.8421,-20.4894]
## Number of points: 6324
## Data attributes:
##       pcds                                          id
##  L1 6LS : 126   {00029226-80EF-4280-9809-109B8509656A}:   1
##  L8 5TE :  63   {00041BD2-4A07-4D41-A5AE-6459CD5FD37C}:   1
##  L1 5AQ :  34   {0005AE67-9150-41D4-8D56-6BFC868EECA3}:   1
##  L24 1WA:  31   {00183CD7-EE48-434B-8A1A-C94B30A93691}:   1
##  L17 6BT:  26   {003EA3A5-F804-458D-A66F-447E27569456}:   1
##  L3 1EE :  24   {00411304-DD5B-4F11-9748-93789D6A000E}:   1
##  (Other):6020   (Other)                                :6318
##      price                    trans_date   type     new      duration
##  Min.   :    1000   2014-06-27 00:00: 109   D: 505   N:5495   F:3927
##  1st Qu.:   70000   2014-12-19 00:00: 109   F:1371   Y: 829   L:2397
##  Median :  110000   2014-02-28 00:00: 105   O: 119
##  Mean   :  144310   2014-10-31 00:00:  95   S:1478
##  3rd Qu.:  160000   2014-03-28 00:00:  94   T:2851
##  Max.   :26615720   2014-11-28 00:00:  94
##                     (Other)         :5718
##       paon               saon                  street
##  3      : 203   FLAT 2     :  25   CROSSHALL STREET: 133
##  11     : 151   FLAT 3     :  25   STANHOPE STREET :  63
##  14     : 148   FLAT 1     :  24   PALL MALL       :  47
##  5      : 146   APARTMENT 4:  23   DUKE STREET     :  41
##  4      : 140   APARTMENT 2:  21   MANN ISLAND     :  41
##  8      : 128   (Other)    : 893   OLD HALL STREET :  39
##  (Other):5408   NA's       :5313   (Other)         :5960
##           locality           town           district           county
##  WAVERTREE   : 126   LIVERPOOL:6324   KNOWSLEY :  12   MERSEYSIDE:6324
##  MOSSLEY HILL: 102                    LIVERPOOL:6311
##  WALTON      :  88                    WIRRAL   :   1
##  WEST DERBY  :  71
##  WOOLTON     :  66
##  (Other)     : 548
##  NA's        :5323
##  ppd_cat  status         lsoa11          LSOA11CD
##  A:5393   A:6324   E01033762: 144   E01033762: 144
##  B: 931            E01033756:  98   E01033756:  98
```

```
##                    E01033752:  93   E01033752:  93
##                    E01033750:  71   E01033750:  71
##                    E01006518:  68   E01006518:  68
##                    E01033755:  65   E01033755:  65
##                    (Other)   :5785   (Other)   :5785
```

See how it contains several pieces, some relating to the spatial information, some relating to the tabular data attached to it. We can access each of the separately if we need it. For example, to pull out the names of the columns in the `data.frame`, we can use the `@data` appendix:

```
colnames(db@data)
```

```
##  [1] "pcds"       "id"         "price"      "trans_date" "type"
##  [6] "new"        "duration"   "paon"       "saon"       "street"
## [11] "locality"   "town"       "district"   "county"     "ppd_cat"
## [16] "status"     "lsoa11"     "LSOA11CD"
```

The rest of this session will focus on two main elements of the shapefile: the spatial dimension (as stored in the point coordinates), and the house price values contained in the `price` column. To get a sense of what they look like first, let us plot both. We can get a quick look at the non-spatial distribution of house values with the following commands:

```
# Create the histogram
hist <- qplot(data=db@data,x=price)
hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

This basically shows there is a lot of values concentrated around the lower end of the distribution but a few very large ones. A usual transformation to *shrink* these differences is to take logarithms:

```
# Create log and add it to the table
logpr <- log(as.numeric(db@data$price))
db@data['logpr'] <- logpr
# Create the histogram
hist <- qplot(x=logpr)
hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

To obtain the spatial distribution of these houses, we need to turn away from the `@data` component of `db`. The easiest, quickest (and also dirtiest) way to get a sense of what the data look like over space is using `plot`:

```
plot(db)
```

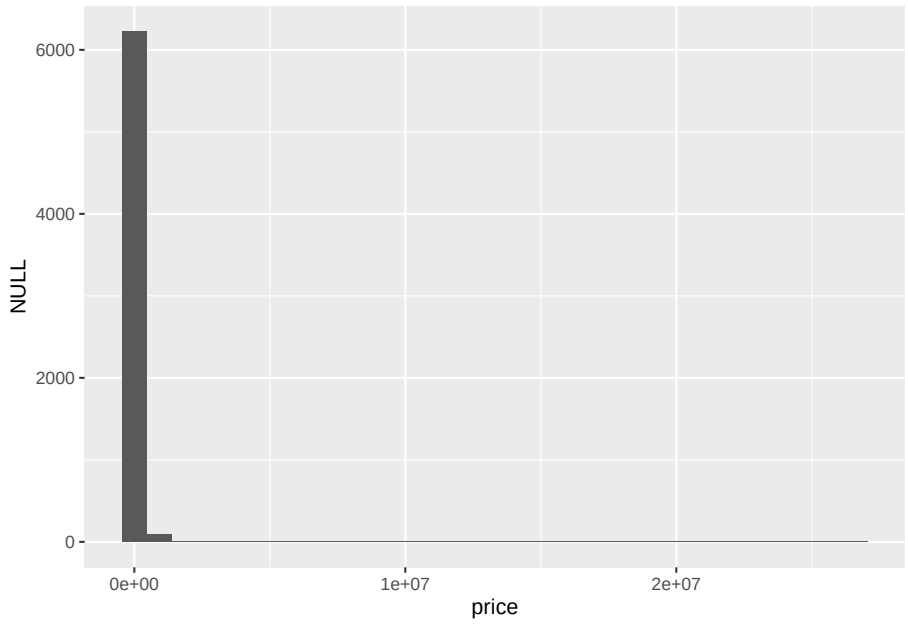


Figure 3.1: Raw house prices in Liverpool

3.3 KDE

Kernel Density Estimation (KDE) is a technique that creates a *continuous* representation of the distribution of a given variable, such as house prices. Although theoretically it can be applied to any dimension, usually, KDE is applied to either one or two dimensions.

3.3.1 One-dimensional KDE

KDE over a single dimension is essentially a contiguous version of a histogram. We can see that by overlaying a KDE on top of the histogram of logs that we have created before:

```
# Create the base
base <- ggplot(db@data, aes(x=logpr))
# Histogram
hist <- base +
  geom_histogram(bins=50, aes(y=..density..))
# Overlay density plot
kde <- hist +
```

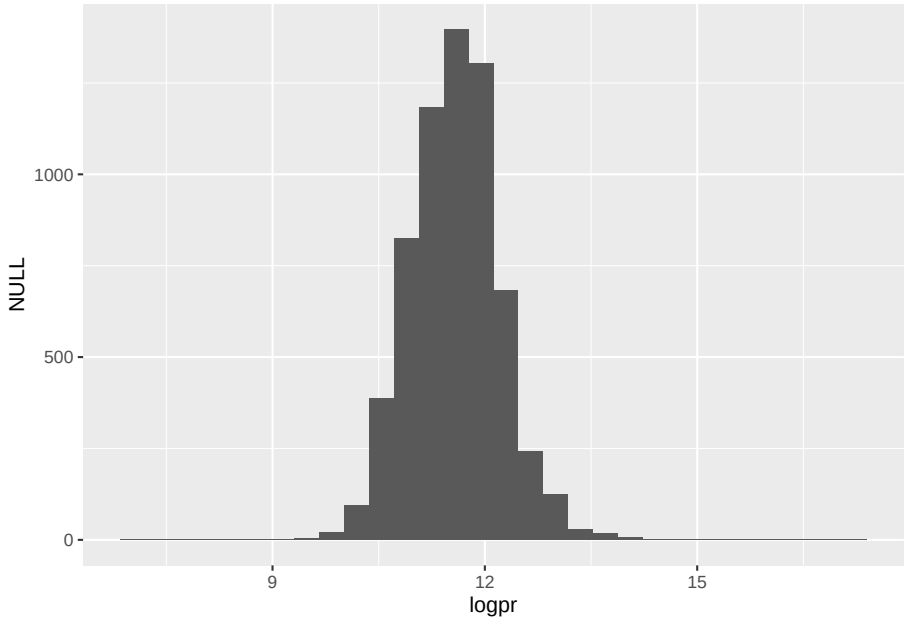


Figure 3.2: Log of house price in Liverpool

```
  geom_density(fill="#FF6666", alpha=0.5, colour="#FF6666")
kde
```

The key idea is that we are smoothing out the discrete binning that the histogram involves. Note how the histogram is exactly the same as above shapewise, but it has been rescalend on the Y axis to reflect probabilities rather than counts.

3.3.2 Two-dimensional (spatial) KDE

Geography, at the end of the day, is usually represented as a two-dimensional space where we locate objects using a system of dual coordinates, `X` and `Y` (or latitude and longitude). Thanks to that, we can use the same technique as above to obtain a smooth representation of the distribution of a two-dimensional variable. The crucial difference is that, instead of obtaining a curve as the output, we will create a *surface*, where intensity will be represented with a color gradient, rather than with the second dimension, as it is the case in the figure above.

To create a spatial KDE in R, there are several ways. If you do not want to necessarily acknowledge the spatial nature of your data, or you they are not

Figure 3.3: Spatial distribution of house transactions in Liverpool

stored in a spatial format, you can plot them using `ggplot2`. Note we first need to convert the coordinates (stored in the spatial part of `db`) into columns of X and Y coordinates, then we can plot them:

```
# Attach XY coordinates
db@data['X'] <- db@coords[, 1]
db@data['Y'] <- db@coords[, 2]
# Set up base layer
base <- ggplot(data=db@data, aes(x=X, y=Y))
# Create the KDE surface
kde <- base + stat_density2d(aes(x = X, y = Y, alpha = ..level..),
               size = 0.01, bins = 16, geom = 'polygon') +
            scale_fill_gradient()
kde
```

Or, we can use a package such as the `GISTools`, which allows to pass a spatial object directly:

```
# Compute the KDE
kde <- kde.points(db)
# Plot the KDE
level.plot(kde)
```

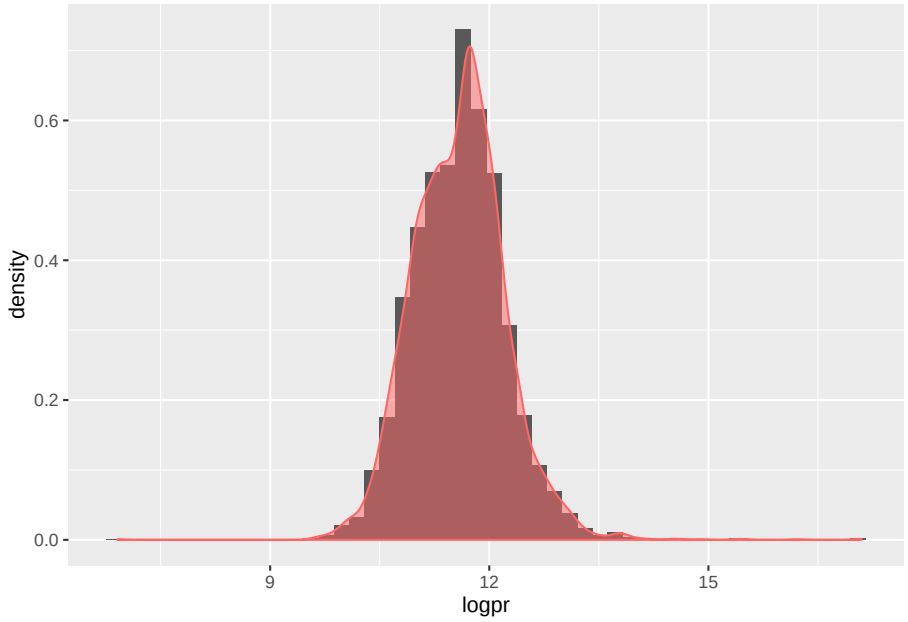


Figure 3.4: Histogram and KDE of the log of house prices in Liverpool

Either of these approaches generate a surface that represents the density of dots, that is an estimation of the probability of finding a house transaction at a given coordinate. However, without any further information, they are hard to interpret and link with previous knowledge of the area. To bring such context to the figure, we can plot an underlying basemap, using a cloud provider such as Google Maps or, as in this case, OpenStreetMap. To do it, we will leverage the library `ggmap`, which is designed to play nicely with the `ggplot2` family (hence the seemingly counterintuitive example above). Before we can plot them with the online map, we need to reproject them though.

```
# Reproject coordinates
wgs84 <- CRS("+proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0")
db_wgs84 <- spTransform(db, wgs84)
db_wgs84@data['lon'] <- db_wgs84@coords[, 1]
db_wgs84@data['lat'] <- db_wgs84@coords[, 2]
xys <- db_wgs84@data[c('lon', 'lat')]
# Bounding box
liv <- c(left = min(xys$lon), bottom = min(xys$lat),
         right = max(xys$lon), top = max(xys$lat))
# Download map tiles
basemap <- get_stamenmap(liv, zoom = 12,
```

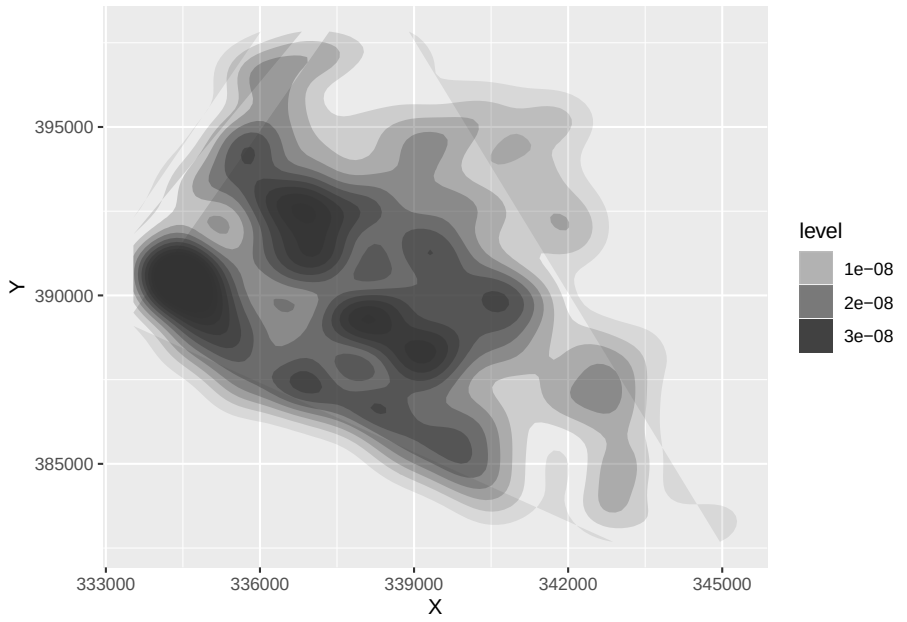


Figure 3.5: KDE of house transactions in Liverpool

```
                          maptype = "toner-lite")
```

```
## Source : http://tile.stamen.com/toner-lite/12/2013/1325.png
## Source : http://tile.stamen.com/toner-lite/12/2014/1325.png
## Source : http://tile.stamen.com/toner-lite/12/2015/1325.png
## Source : http://tile.stamen.com/toner-lite/12/2013/1326.png
## Source : http://tile.stamen.com/toner-lite/12/2014/1326.png
## Source : http://tile.stamen.com/toner-lite/12/2015/1326.png
## Source : http://tile.stamen.com/toner-lite/12/2013/1327.png
## Source : http://tile.stamen.com/toner-lite/12/2014/1327.png
## Source : http://tile.stamen.com/toner-lite/12/2015/1327.png
```

```
# Overlay KDE
final <- ggmap(basemap, extent = "device",
               maprange=FALSE) +
  stat_density2d(data = db_wgs84@data,
                 aes(x = lon, y = lat,
```

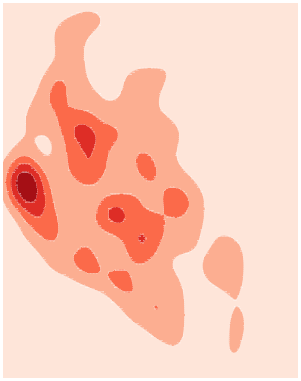


Figure 3.6: KDE of house transactions in Liverpool

```
                    alpha=..level..,
                    fill = ..level..),
                size = 0.01, bins = 16,
                geom = 'polygon',
                show.legend = FALSE) +
  scale_fill_gradient2("Transaction\nDensity",
                       low = "#fffff8",
                       high = "#8da0cb")
final
```

The plot above[3] allows us to not only see the distribution of house transactions, but to relate it to what we know about Liverpool, allowing us to establish many more connections than we were previously able. Mainly, we can easily see that the area with a highest volume of houses being sold is the city centre, with a "hole" around it that displays very few to no transactions and then several pockets further away.

[3] **EXERCISE** The map above uses the Stamen map `toner-lite`. Explore additional tile styles on their website and try to recreate the plot above.

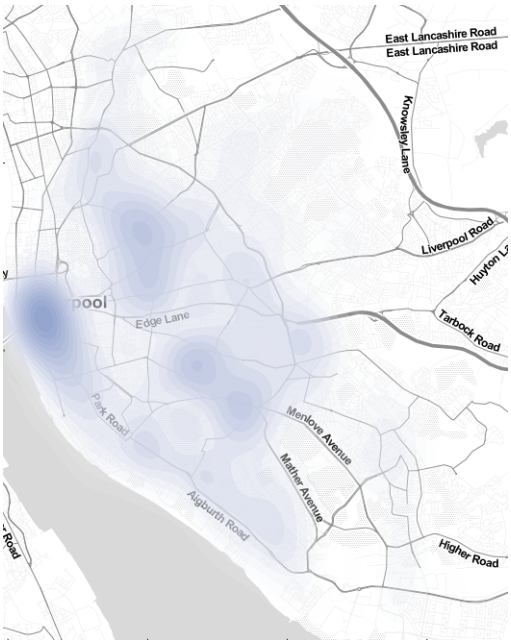


Figure 3.7: KDE of house transactions in Liverpool

3.4 Spatial Interpolation

The previous section demonstrates how to visualize the distribution of a set of spatial objects represented as points. In particular, given a bunch of house transactions, it shows how one can effectively visualize their distribution over space and get a sense of the density of occurrences. Such visualization, because it is based on KDE, is based on a smooth continuum, rather than on a discrete approach (as a choropleth would do, for example).

Many times however, we are not particularly interested in learning about the density of occurrences, but about the distribution of a given value attached to each location. Think for example of weather stations and temperature: the location of the stations is no secret and rarely changes, so it is not of particular interest to visualize the density of stations; what we are usually interested instead is to know how temperature is distributed over space, given we only measure it in a few places. One could argue the example we have been working with so far, house price transactions, fits into this category as well: although where house are sold may be of relevance, more often we are interested in finding out what the "surface of price" looks like. Rather than *where are most houses being sold?* we usually want to know *where the most expensive or most affordable* houses are located.

In cases where we are interested in creating a surface of a given value, rather than a simple density surface of occurrences, KDE cannot help us. In these cases, what we are interested in is *spatial interpolation*, a family of techniques that aim at exactly that: creating continuous surfaces for a particular phenomenon (e.g. temperature, house prices) given only a finite sample of observations. Spatial interpolation is a large field of research that is still being actively developed and that can involve a substantial amount of mathematical complexity in order to obtain the most accurate estimates possible[4]. In this session, we will introduce the simplest possible way of interpolating values, hoping this will give you a general understanding of the methodology and, if you are interested, you can check out further literature. For example, Banerjee et al. (2014) or Cressie (2015) are hard but good overviews.

3.4.1 Inverse Distance Weight (IDW) interpolation

The technique we will cover here is called *Inverse Distance Weighting*, or IDW for convenience. Brunsdon and Comber (2015) offer a good description:

> In the *inverse distance weighting* (IDW) approach to interpolation, to estimate the value of z at location x a weighted mean of nearby observations is taken [...]. To accommodate the idea that observations of z at points closer to x should be given more importance in the interpolation, greater weight is given to these points [...]
>
> — Page 204

The math[5] is not particularly complicated and may be found in detail elsewhere (the reference above is a good starting point), so we will not spend too much time on it. More relevant in this context is the intuition behind. Essentially, the idea is that we will create a surface of house price by smoothing many values arranged along a regular grid and obtained by interpolating from the known locations to the regular grid locations. This will give us full and equal coverage to soundly perform the smoothing.

Enough chat, let's code.

From what we have just mentioned, there are a few steps to perform an IDW spatial interpolation:

1. Create a regular grid over the area where we have house transactions.

[4]There is also an important economic incentive to do this: some of the most popular applications are in the oil and gas or mining industries. In fact, the very creator of this technique, Danie G. Krige, was a mining engineer. His name is usually used to nickname spatial interpolation as *kriging*.

[5]Essentially, for any point x in space, the IDW estimate for value z is equivalent to $\hat{z}(x) = \frac{\sum_i w_i z_i}{\sum_i w_i}$ where i are the observations for which we do have a value, and w_i is a weight given to location i based on its distance to x.

2. Obtain IDW estimates for each point in the grid, based on the values of k nearest neighbors.
3. Plot a smoothed version of the grid, effectively representing the surface of house prices.

Let us go in detail into each of them[6]. First, let us set up a grid:

```
liv.grid <- spsample(db, type='regular', n=25000)
```

That's it, we're done! The function `spsample` hugely simplifies the task by taking a spatial object and returning the grid we need. Not a couple of additional arguments we pass: `type` allows us to get a set of points that are *uniformly* distributed over space, which is handy for the later smoothing; `n` controls how many points we want to create in that grid.

On to the IDW. Again, this is hugely simplified by `gstat`:

```
idw.hp <- idw(price ~ 1, locations=db, newdata=liv.grid)
```

```
## [inverse distance weighted interpolation]
```

Boom! We've got it. Let us pause for a second to see how we just did it. First, we pass `price ~ 1`. This specifies the formula we are using to model house prices. The name on the left of `~` represents the variable we want to explain, while everything to its right captures the *explanatory* variables. Since we are considering the simplest possible case, we do not have further variables to add, so we simply write `1`. Then we specify the original locations for which we do have house prices (our original `db` object), and the points where we want to interpolate the house prices (the `liv.grid` object we just created above). One more note: by default, `idw.hp` uses all the available observations, weighted by distance, to provide an estimate for a given point. If you want to modify that and restrict the maximum number of neighbors to consider, you need to tweak the argument `nmax`, as we do above by using the 150 neares observations to each point[7].

The object we get from `idw` is another spatial table, just as `db`, containing the interpolated values. As such, we can inspect it just as with any other of its kind. For example, to check out the top of the estimated table:

```
head(idw.hp@data)
```

```
##   var1.pred var1.var
## 1  158043.1       NA
## 2  158152.6       NA
## 3  158264.8       NA
## 4  158379.9       NA
## 5  158498.0       NA
```

[6] For the relevant calculations, we will be using the `gstat` library.
[7] Have a play with this because the results do change significantly. Can you reason why?

```
## 6  158619.2       NA
```

The column we will pay attention to is `var1.pred`. And to see the locations for which those correspond:

```
head(idw.hp@coords)
```

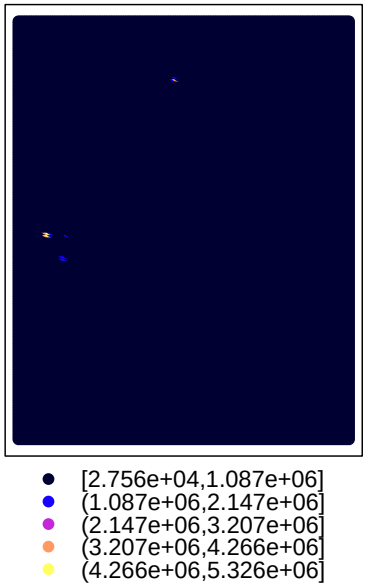
```
##            x1       x2
## [1,] 333552.3 382721.4
## [2,] 333637.3 382721.4
## [3,] 333722.3 382721.4
## [4,] 333807.2 382721.4
## [5,] 333892.2 382721.4
## [6,] 333977.2 382721.4
```

So, for a hypothetical house sold at the location in the first row of `idw.hp@coords` (expressed in the OSGB coordinate system), the price we would guess it would cost, based on the price of houses sold nearby, is the first element of column `var1.pred` in `idw.hp@data`.

3.4.2 A surface of housing prices

Once we have the IDW object computed, we can plot it to explore the distribution, not of house transactions in this case, but of house price over the geography of Liverpool. The easiest way to do this is by quickly calling the command `spplot`:

```
spplot(idw.hp['var1.pred'])
```



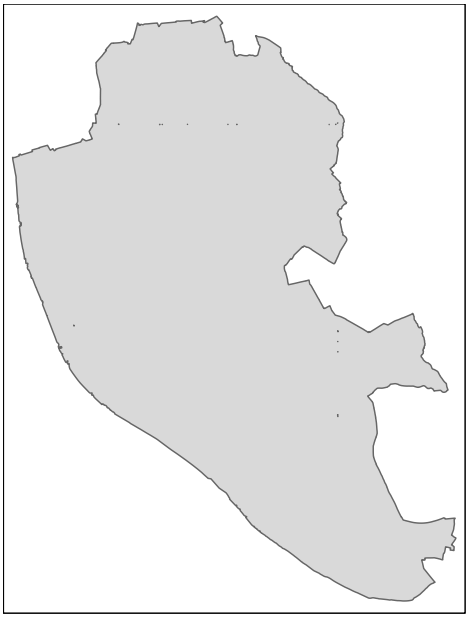
However, this is not entirely satisfactory for a number of reasons. Let us get an equivalen plot with the package `tmap`, which streamlines some of this and makes more aesthetically pleasant maps easier to build as it follows a "ggplot-y" approach.

```
# Load up the layer
liv.otl <- readOGR('data/house_transactions', 'liv_outline')
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/home/jovyan/work/data/house_transactions", layer: "liv_outline"
## with 1 features
## It has 1 fields
```

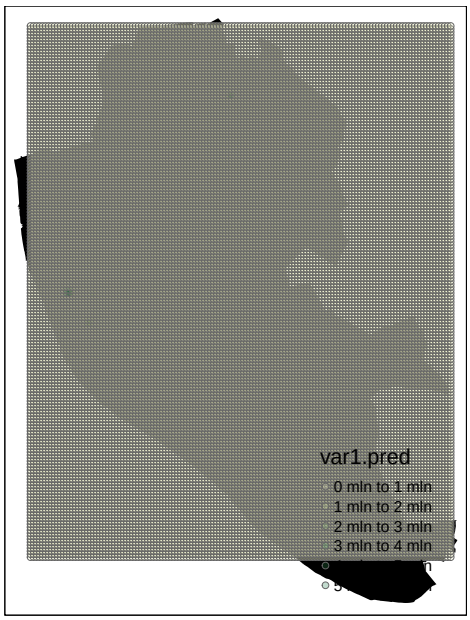
The shape we will overlay looks like this:

```
qtm(liv.otl)
```



Now let's give it a first go!

```
#
p = tm_shape(liv.otl) + tm_fill(col='black', alpha=1) +
  tm_shape(idw.hp) +
  tm_symbols(col='var1.pred', size=0.1, alpha=0.25,
             border.lwd=0., palette='YlGn')
p
```



The last two plots, however, are not really a surface, but a representation of the points we have just estimated. To create a surface, we need to do an interim transformation to convert the spatial object `idw.hp` into a table that a "surface plotter" can understand.

```
xyz <- data.frame(x=coordinates(idw.hp)[, 1],
                  y=coordinates(idw.hp)[, 2],
                  z=idw.hp$var1.pred)
```

Now we are ready to plot the surface as a contour:

```
base <- ggplot(data=xyz, aes(x=x, y=y))
surface <- base + geom_contour(aes(z=z))
surface
```

Which can also be shown as a filled contour:

```
base <- ggplot(data=xyz, aes(x=x, y=y))
surface <- base + geom_raster(aes(fill=z))
surface
```

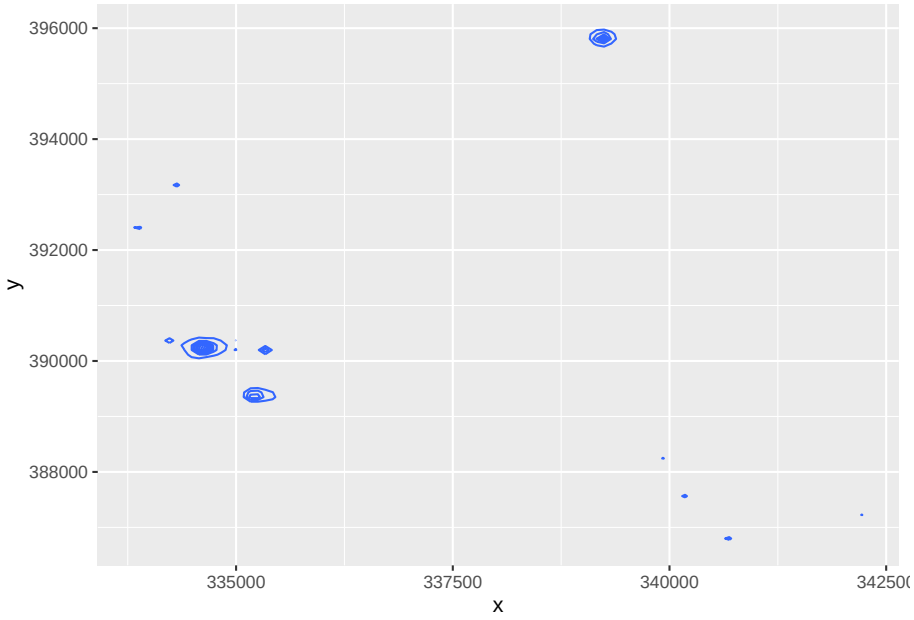
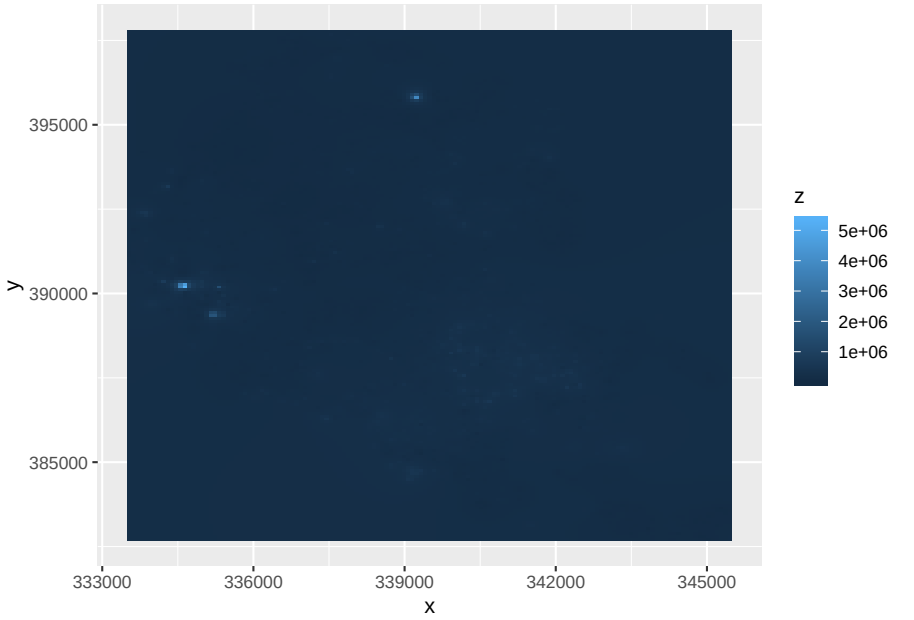


Figure 3.8: Contour of prices in Liverpool



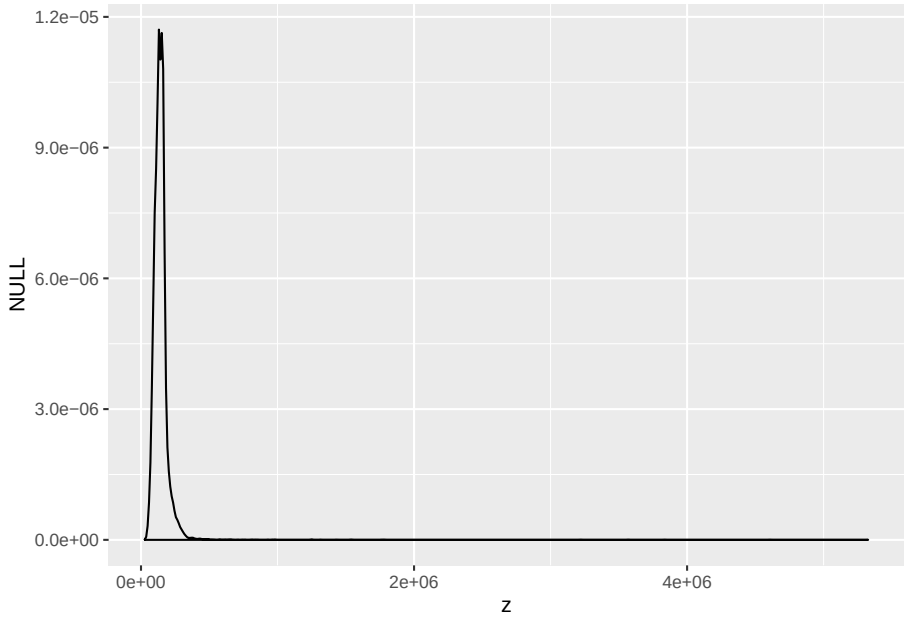


Figure 3.9: Skewness of prices in Liverpool

The problem here, when compared to the KDE above for example, is that a few values are extremely large:

```
qplot(data=xyz, x=z, geom='density')
```

Let us then take the logarithm before we plot the surface:

```
xyz['lz'] <- log(xyz$z)
base <- ggplot(data=xyz, aes(x=x, y=y))
surface <- base +
           geom_raster(aes(fill=lz),
                       show.legend = F)
surface
```

Now this looks better. We can start to tell some patterns. To bring in context, it would be great to be able to add a basemap layer, as we did for the KDE. This is conceptually very similar to what we did above, starting by reprojecting the points and continuing by overlaying them on top of the basemap. However, technically speaking it is not possible because `ggmap` –the library we have been using to display tiles from cloud providers– does not play well with our own rasters (i.e. the price surface). At the moment, it is surprisingly tricky to get this to work, so we will park it for now. However, developments such as the `sf`

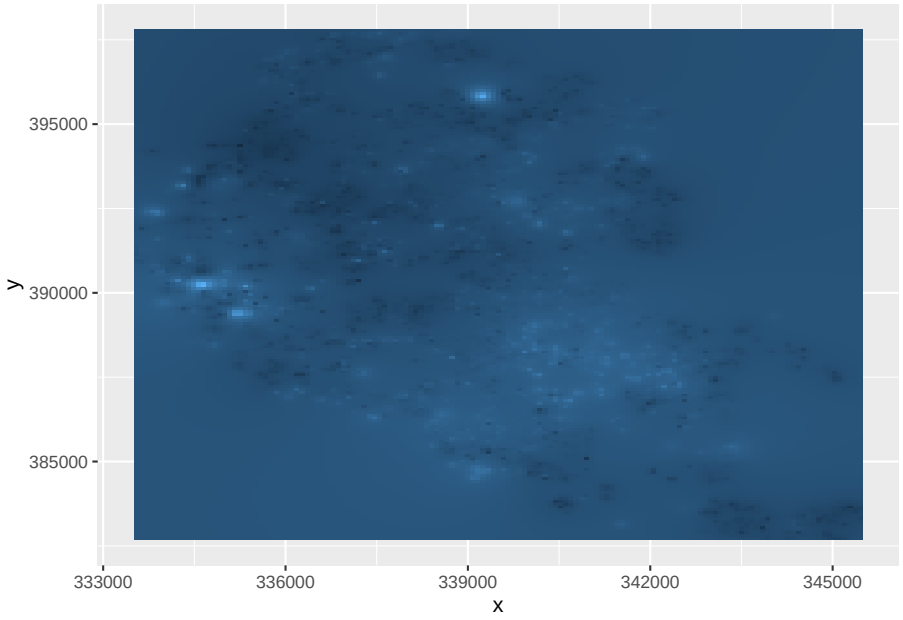


Figure 3.10: Surface of log-prices in Liverpool

project promise to make this easier in the future[8].

3.4.3 *"What should the next house's price be?"*

The last bit we will explore in this session relates to prediction for new values. Imagine you are a real state data scientist and your boss asks you to give an estimate of how much a new house going into the market should cost. The only information you have to make such a guess is the location of the house. In this case, the IDW model we have just fitted can help you. The trick is realizing that, instead of creating an entire grid, all we need is to obtain an estimate of a single location.

Let us say, the house is located on the coordinates `x=340000, y=390000` as expressed in the GB National Grid coordinate system. In that case, we can do as follows:

```
pt <- SpatialPoints(cbind(x=340000, y=390000),
                    proj4string = db@proj4string)
idw.one <- idw(price ~ 1, locations=db, newdata=pt)
```

[8] **BONUS** if you can figure out a way to do it yourself!

```
## [inverse distance weighted interpolation]
```

```
idw.one
```

```
## class       : SpatialPointsDataFrame
## features    : 1
## extent      : 340000, 340000, 390000, 390000  (xmin, xmax, ymin, ymax)
## crs         : +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100
## variables   : 2
## names       :        var1.pred, var1.var
## value       : 157099.029513871,       NA
```

And, as show above, the estimated value is GBP157,099[9].

Using this predictive logic, and taking advantage of Google Maps and its geocoding capabilities, it is possible to devise a function that takes an arbitrary address in Liverpool and, based on the transactions occurred throughout 2014, provides an estimate of what the price for a property in that location could be.

```
how.much.is <- function(address, print.message=TRUE){
  # Convert the address into Lon/Lat coordinates
  # NOTE: this now requires an API key
  # https://github.com/dkahle/ggmap#google-maps-and-credentials
  ll.pt <- geocode(address)
  # Process as spatial table
  wgs84 <- CRS("+proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0")
  ll.pt <- SpatialPoints(cbind(x=ll.pt$lon, y=ll.pt$lat),
                         proj4string = wgs84)
  # Transform Lon/Lat into OSGB
  pt <- spTransform(ll.pt, db@proj4string)
  # Obtain prediction
  idw.one <- idw(price ~ 1, locations=db, newdata=pt)
  price <- idw.one@data$var1.pred
  # Return predicted price
  if(print.message==T){
    writeLines(paste("\n\nBased on what surrounding properties were sold",
                     "for in 2014 a house located at", address, "would",
                     "cost",  paste("GBP", round(price), ".", sep=''), "\n\n"))
  }
  return(price)
}
```

Ready to test!

```
address <- "74 Bedford St S, Liverpool, L69 7ZT, UK"
#p <- how.much.is(address)
```

[9] **PRO QUESTION** Is that house expensive or cheap, as compared to the other houses sold in this dataset? Can you figure out where the house is?

Chapter 4

Flows

This chapter[1] covers spatial interaction flows. Using open data from the city of San Francisco about trips on its bikeshare system, we will estimate spatial interaction models that try to capture and explain the variation in the amount of trips on each given route. After visualizing the dataset, we begin with a very simple model and then build complexity progressively by augmenting it with more information, refined measurements, and better modeling approaches. Throughout the chapter, we explore different ways to grasp the predictive performance of each model. We finish with a prediction example that illustrates how these models can be deployed in a real-world application.

Content is based on the following references, which are great follow-up's on the topic:

- Singleton (2017), an online short course on R for Geographic Data Science and Urban Analytics. In particular, the section on mapping flows is specially relevant here.
- The predictive checks section draws heavily from Gelman and Hill (2006), in particular Chapters 6 and 7.

This tutorial is part of Spatial Analysis Notes, a compilation hosted as a GitHub repository that you can access in a few ways:

- As a download of a `.zip` file that contains all the materials.
- As an html website.
- As a pdf document
- As a GitHub repository.

[1] This chapter is part of Spatial Analysis Notes Flows – Exploring flows visually and through spatial interaction by Dani Arribas-Bel is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

4.1 Dependencies

This tutorial relies on the following libraries that you will need to have installed on your machine to be able to interactively follow along[2]. Once installed, load them up with the following commands:

```
# Layout
library(tufte)
# Spatial Data management
library(rgdal)
```

```
## Loading required package: sp
```

```
## rgdal: version: 1.4-4, (SVN revision 833)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: /usr/share/gdal/2.2
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: (autodetected)
##  Linking to sp version: 1.3-1
```

```
# Pretty graphics
library(ggplot2)
# Thematic maps
library(tmap)
# Pretty maps
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
# Simulation methods
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /home/jovyan/work
```

[2] You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

Before we start any analysis, let us set the path to the directory where we are working. We can easily do that with `setwd()`. Please replace in the following line the path to the folder where you have placed this file -and where the `sf_bikes` folder with the data lives.

```
setwd('.')
```

4.2 Data

In this note, we will use data from the city of San Francisco representing bike trips on their public bike share system. The original source is the SF Open Data portal (link) and the dataset comprises both the location of each station in the Bay Area as well as information on trips (station of origin to station of destination) undertaken in the system from September 2014 to August 2015 and the following year. Since this note is about modeling and not data preparation, a cleanly reshaped version of the data, together with some additional information, has been created and placed in the `sf_bikes` folder. The data file is named `flows.geojson` and, in case you are interested, the (Python) code required to created from the original files in the SF Data Portal is also available on the `flows_prep.ipynb` notebook [url], also in the same folder.

Let us then directly load the file with all the information necessary:

```
db <- readOGR('./data/sf_bikes/flows.geojson')
```

```
## OGR data source with driver: GeoJSON
## Source: "/home/jovyan/work/data/sf_bikes/flows.geojson", layer: "flows"
## with 1722 features
## It has 9 fields
```

```
rownames(db@data) <- db$flow_id
db@data$flow_id <- NULL
```

Note how the interface is slightly different since we are reading a `GeoJSON` file instead of a shapefile.

The data contains the geometries of the flows, as calculated from the Google Maps API, as well as a series of columns with characteristics of each flow:

```
head(db@data)
```

```
##       dest orig straight_dist street_dist total_down total_up trips15
## 39-41   41   39      1452.201   1804.1150  11.205753 4.698162      68
## 39-42   42   39      1734.861   2069.1557  10.290236 2.897886      23
## 39-45   45   39      1255.349   1747.9928  11.015596 4.593927      83
## 39-46   46   39      1323.303   1490.8361   3.511543 5.038044     258
## 39-47   47   39       715.689    769.9189   0.000000 3.282495     127
## 39-48   48   39      1996.778   2740.1290  11.375186 3.841296      81
```

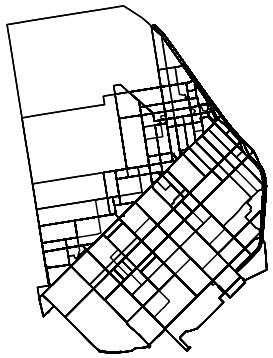


Figure 4.1: Potential routes

```
##       trips16
## 39-41      68
## 39-42      29
## 39-45      50
## 39-46     163
## 39-47      73
## 39-48      56
```

where `orig` and `dest` are the station IDs of the origin and destination, `street/straight_dist` is the distance in metres between stations measured along the street network or as-the-crow-flies, `total_down/up` is the total downhil and climb in the trip, and `tripsXX` contains the amount of trips undertaken in the years of study.

4.3 "*Seeing*" flows

The easiest way to get a quick preview of what the data looks like spatially is to make a simple plot:

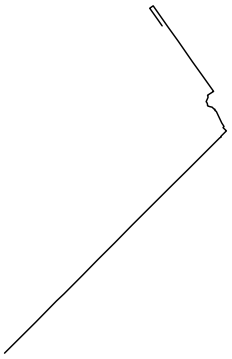


Figure 4.2: Trip from station 39 to 48

```
plot(db)
```

Equally, if we want to visualize a single route, we can simply subset the table. For example, to get the shape of the trip from station `39` to station `48`, we can:

```
one39to48 <- db[ which(
          db@data$orig == 39 & db@data$dest == 48
          ) , ]
plot(one39to48)
```

or, for the most popular route, we can:

```
most_pop <- db[ which(
          db@data$trips15 == max(db@data$trips15)
          ) , ]
plot(most_pop)
```

These however do not reveal a lot: there is no geographical context (*why are there so many routes along the NE?*) and no sense of how volumes of bikers are allocated along different routes. Let us fix those two.

The easiest way to bring in geographical context is by overlaying the routes on top of a background map of tiles downloaded from the internet. Let us download

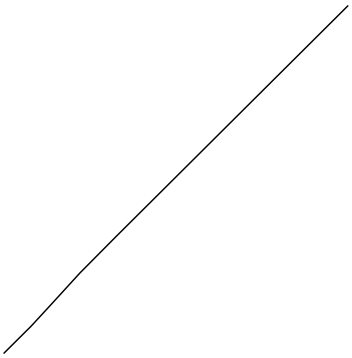


Figure 4.3: Most popular trip

this using ggmap:

```
sf_bb <- c(left=db@bbox['x', 'min'],
           right=db@bbox['x', 'max'],
           bottom=db@bbox['y', 'min'],
           top=db@bbox['y', 'max'])
SanFran <- get_stamenmap(sf_bb,
                         zoom = 14,
                         maptype = "toner-lite")
```

```
## Source : http://tile.stamen.com/toner-lite/14/2620/6330.png
## Source : http://tile.stamen.com/toner-lite/14/2621/6330.png
## Source : http://tile.stamen.com/toner-lite/14/2622/6330.png
## Source : http://tile.stamen.com/toner-lite/14/2620/6331.png
## Source : http://tile.stamen.com/toner-lite/14/2621/6331.png
## Source : http://tile.stamen.com/toner-lite/14/2622/6331.png
## Source : http://tile.stamen.com/toner-lite/14/2620/6332.png
## Source : http://tile.stamen.com/toner-lite/14/2621/6332.png
```

```
## Source : http://tile.stamen.com/toner-lite/14/2622/6332.png
## Source : http://tile.stamen.com/toner-lite/14/2620/6333.png
## Source : http://tile.stamen.com/toner-lite/14/2621/6333.png
## Source : http://tile.stamen.com/toner-lite/14/2622/6333.png
```

and make sure it looks like we intend it to look:

```
ggmap(SanFran)
```

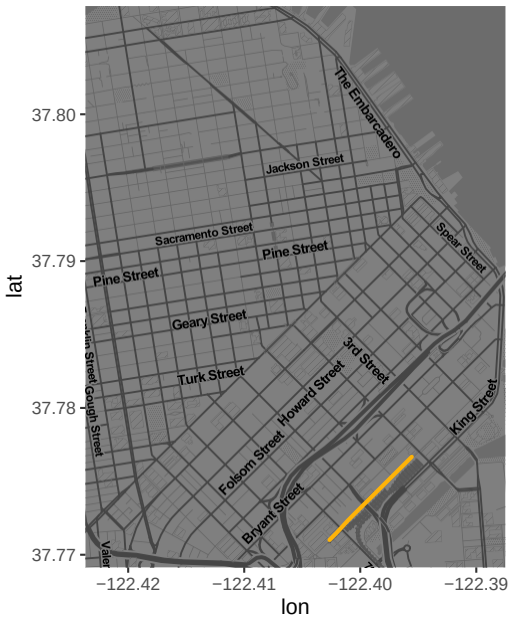
Now to combine tiles and routes, we need to pull out the coordinates that make up each line. For the route example above, this would be:

```
xys1 <- as.data.frame(coordinates(most_pop))
```

Now we can plot the route[3] (note we also dim down the background to focus the attention on flows):

```
ggmap(SanFran, darken=0.5) +
  geom_path(aes(x=X1, y=X2),
            data=xys1,
            size=1,
            color=rgb(0.996078431372549, 0.7019607843137254, 0.03137254901960784),
            lineend='round')
```

[3] **EXERCISE**: *can you plot the route for the largest climb?*



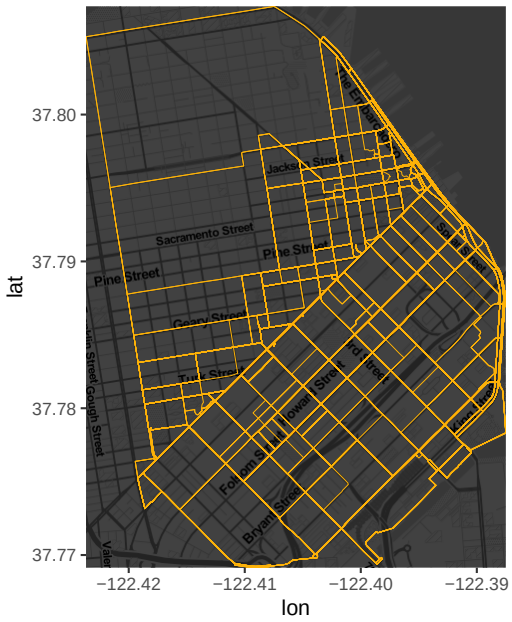
Now we can plot all of the lines by using a short `for` loop to build up the table:

```
# Set up shell data.frame
lines <- data.frame(lat = numeric(0),
                    lon = numeric(0),
                    trips = numeric(0),
                    id = numeric(0)
                    )
# Run loop
for(x in 1:nrow(db)){
  # Pull out row
  r <- db[x, ]
  # Extract lon/lat coords
  xys <- as.data.frame(coordinates(r))
  names(xys) <- c('lon', 'lat')
  # Insert trips and id
  xys['trips'] <- r@data$trips15
  xys['id'] <- x
  # Append them to `lines`
  lines <- rbind(lines, xys)
}
```

Now we can go on and plot all of them:

```
ggmap(SanFran, darken=0.75) +
  geom_path(aes(
```

```
                x=lon,
                y=lat,
                group=id
                ),
            data=lines,
            size=0.1,
            color=rgb(0.996078431372549, 0.7019607843137254, 0.03137254901960784),
            lineend='round')
```



Finally, we can get a sense of the distribution of the flows by associating a color gradient to each flow based on its number of trips:

```
ggmap(SanFran, darken=0.75) +
  geom_path(aes(
                x=lon,
                y=lat,
                group=id,
                colour=trips
                ),
            data=lines,
            size=log1p(lines$trips / max(lines$trips)),
            lineend='round') +
  scale_colour_gradient(low='grey',
                        high='#07eda0') +
  theme(axis.text.x = element_blank(),
```

```
      axis.text.y = element_blank(),
      axis.ticks = element_blank()
      )
```

Note how we transform the size so it's a proportion of the largest trip and then it is compressed with a logarithm.

4.4 Modelling flows

Now we have an idea of the spatial distribution of flows, we can begin to think about modeling them. The core idea in this section is to fit a model that can capture the particular characteristics of our variable of interest (the volume of trips) using a set of predictors that describe the nature of a given flow. We will start from the simplest model and then progressively build complexity until we get to a satisfying point. Along the way, we will be exploring each model using concepts from Gelman and Hill (2006) such as predictive performance checks[4] (PPC)

Before we start running regressions, let us first standardize the predictors so we can interpret the intercept as the average flow when all the predictors take the average value, and so we can interpret the model coefficients as changes in standard deviation units:

[4]For a more elaborate introduction to PPC, have a look at Chapters 7 and 8.

```
# Scale all the table
db_std <- as.data.frame(scale(db@data))
# Reset trips as we want the original version
db_std$trips15 <- db@data$trips15
db_std$trips16 <- db@data$trips16
# Reset origin and destination station and express them as factors
db_std$orig <- as.factor(db@data$orig)
db_std$dest <- as.factor(db@data$dest)
```

Baseline model

One of the simplest possible models we can fit in this context is a linear model that explains the number of trips as a function of the straight distance between the two stations and total amount of climb and downhill. We will take this as the baseline on which we can further build later:

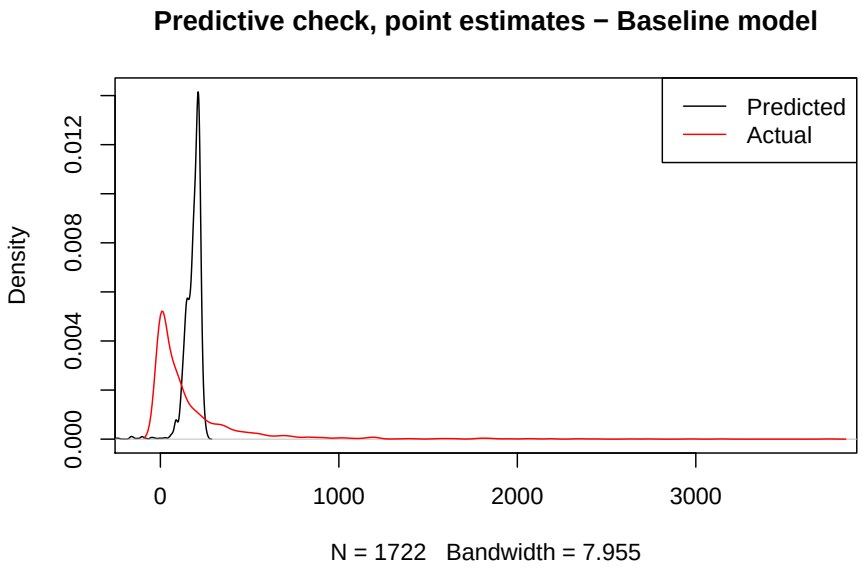
```
m1 <- lm('trips15 ~ straight_dist + total_up + total_down', data=db_std)
summary(m1)
```

```
##
## Call:
## lm(formula = "trips15 ~ straight_dist + total_up + total_down",
##     data = db_std)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -261.9 -168.3 -102.4   30.8 3527.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    182.070      8.110  22.451  < 2e-16 ***
## straight_dist   17.906      9.108   1.966   0.0495 *
## total_up       -44.100      9.353  -4.715 2.61e-06 ***
## total_down     -20.241      9.229  -2.193   0.0284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 336.5 on 1718 degrees of freedom
## Multiple R-squared:  0.02196,    Adjusted R-squared:  0.02025
## F-statistic: 12.86 on 3 and 1718 DF,  p-value: 2.625e-08
```

To explore how good this model is, we will be comparing the predictions the model makes about the number of trips each flow should have with the actual number of trips. A first approach is to simply plot the distribution of both variables:

```
plot(density(m1$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
```

```
     main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Predicted', 'Actual'),
       col=c('black', 'red'),
       lwd=1)
title(main="Predictive check, point estimates - Baseline model")
```



The plot makes pretty obvious that our initial model captures very few aspects of the distribution we want to explain. However, we should not get too attached to this plot just yet. What it is showing is the distribution of predicted *point* estimates from our model. Since our model is not deterministic but inferential, there is a certain degree of uncertainty attached to its predictions, and that is completely absent from this plot.

Generally speaking, a given model has two sources of uncertainty: *predictive*, and *inferential*. The former relates to the fact that the equation we fit does not capture all the elements or in the exact form they enter the true data generating process; the latter has to do with the fact that we never get to know the true value of the model parameters only guesses (estimates) subject to error and uncertainty. If you think of our linear model above as

$$T_{ij} = X_{ij}\beta + \epsilon_{ij}$$

where T_{ij} represents the number of trips undertaken between station i and j, X_{ij} is the set of explanatory variables (length, climb, descent, etc.), and ϵ_{ij} is an error term assumed to be distributed as a normal distribution $N(0, \sigma)$; then predictive uncertainty comes from the fact that there are elements to some extent relevant for y that are not accounted for and thus subsummed into ϵ_{ij}. Inferential uncertainty comes from the fact that we never get to know β but only an estimate of it which is also subject to uncertainty itself.

Taking these two sources into consideration means that the black line in the plot above represents only the behaviour of our model we expect if the error term is absent (no predictive uncertainty) and the coefficients are the true estimates (no inferential uncertainty). However, this is not necessarily the case as our estimate for the uncertainty of the error term is certainly not zero, and our estimates for each parameter are also subject to a great deal of inferential variability. we do not know to what extent other outcomes would be just as likely. Predictive checking relates to simulating several feasible scenarios under our model and use those to assess uncertainty and to get a better grasp of the quality of our predictions.

Technically speaking, to do this, we need to build a mechanism to obtain a possible draw from our model and then repeat it several times. The first part of those two steps can be elegantly dealt with by writing a short function that takes a given model and a set of predictors, and produces a possible random draw from such model:

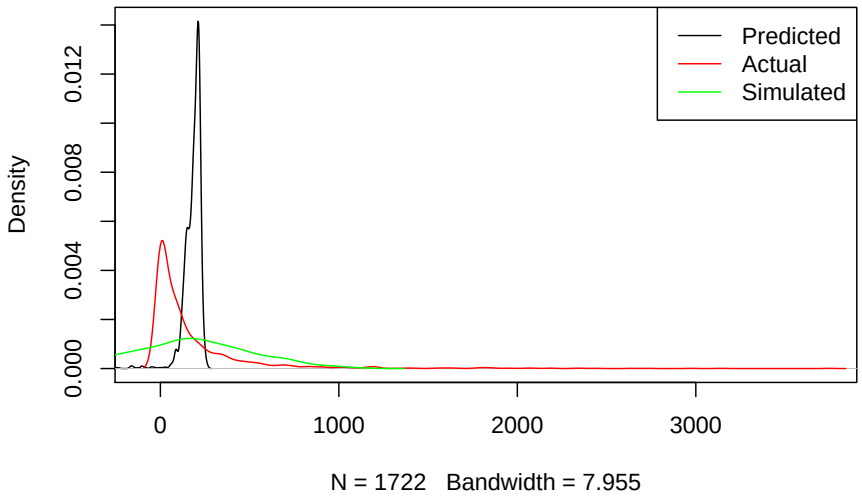
```
generate_draw <- function(m){
  # Set up predictors matrix
  x <- model.matrix(m)
  # Obtain draws of parameters (inferential uncertainty)
  sim_bs <- sim(m, 1)
  # Predicted value
  mu <- x %*% sim_bs@coef[1, ]
  # Draw
  n <- length(mu)
  y_hat <- rnorm(n, mu, sim_bs@sigma[1])
  return(y_hat)
}
```

This function takes a model `m` and the set of covariates `x` used and returns a random realization of predictions from the model. To get a sense of how this works, we can get and plot a realization of the model, compared to the expected one and the actual values:

```
new_y <- generate_draw(m1)

plot(density(m1$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
```

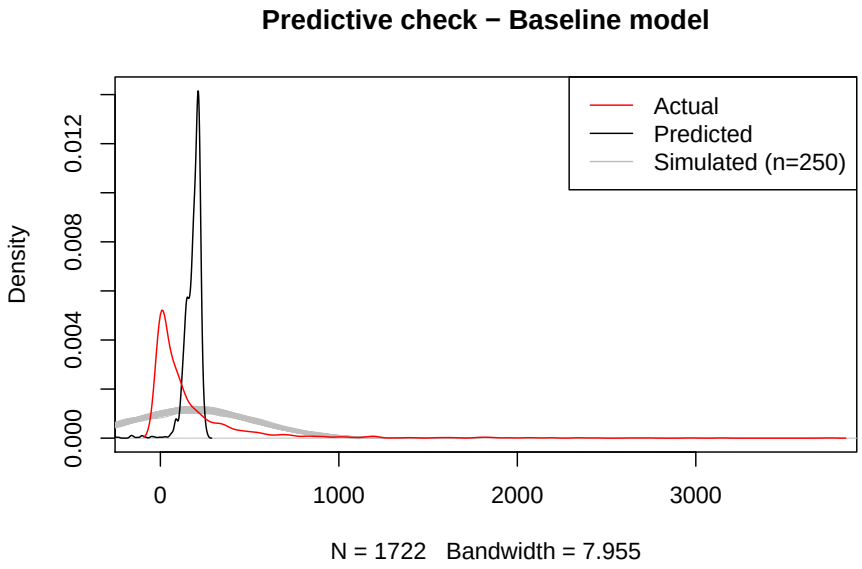
```
                    max(density(m1$fitted.values)$y),
                    max(density(db_std$trips15)$y)
                    )
                  )
            ),
     col='black',
     main='')
lines(density(db_std$trips15),
      col='red',
      main='')
lines(density(new_y),
      col='green',
      main='')
legend('topright',
       c('Predicted', 'Actual', 'Simulated'),
       col=c('black', 'red', 'green'),
       lwd=1)
```



Once we have this "draw engine", we can set it to work as many times as we want using a simple `for` loop. In fact, we can directly plot these lines as compared to the expected one and the trip count:

```
plot(density(m1$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
```

```
                  max(density(m1$fitted.values)$y),
                  max(density(db_std$trips15)$y)
                  )
                )
        ),
     col='white',
     main='')
# Loop for realizations
for(i in 1:250){
  tmp_y <- generate_draw(m1)
  lines(density(tmp_y),
        col='grey',
        lwd=0.1
        )
}
#
lines(density(m1$fitted.values),
      col='black',
      main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Actual', 'Predicted', 'Simulated (n=250)'),
       col=c('red', 'black', 'grey'),
       lwd=1)
title(main="Predictive check - Baseline model")
```



The plot shows there is a significant mismatch between the fitted values, which are much more concentrated around small positive values, and the realizations of our "inferential engine", which depict a much less concentrated distribution of values. This is likely due to the combination of two different reasons: on the one hand, the accuracy of our estimates may be poor, causing them to jump around a wide range of potential values and hence resulting in very diverse predictions (inferential uncertainty); on the other hand, it may be that the amount of variation we are not able to account for in the model[5] is so large that the degree of uncertainty contained in the error term of the model is very large, hence resulting in such a flat predictive distribution.

It is important to keep in mind that the issues discussed in the paragraph above relate only to the uncertainty behind our model, not to the point predictions derived from them, which are a mechanistic result of the minimization of the squared residuals and hence are not subject to probability or inference. That allows them in this case to provide a fitted distribution much more accurate apparently (black line above). However, the lesson to take from this model is that, even if the point predictions (fitted values) are artificially accurate[6], our capabilities to infer about the more general underlying process are fairly limited.

Improving the model

The bad news from the previous section is that our initial model is not great at explaining bike trips. The good news is there are several ways in which

[5] The R^2 of our model is around 2%

[6] which they are not really, in light of the comparison between the black and red lines.

we can improve this. In this section we will cover three main extensions that exemplify three different routes you can take when enriching and augmenting models in general, and spatial interaction ones in particular[7]. These three routes are aligned around the following principles:

1. Use better approximations to model your dependent variable.
2. Recognize the structure of your data.
3. Get better predictors.

- **Use better approximations to model your dependent variable**

Standard OLS regression assumes that the error term and, since the predictors are deterministic, the dependent variable are distributed following a normal (gaussian) distribution. This is usually a good approximation for several phenomena of interest, but maybe not the best one for trips along routes: for one, we know trips cannot be negative, which the normal distribution does not account for[8]; more subtly, their distribution is not really symmetric but skewed with a very long tail on the right. This is common in variables that represent counts and that is why usually it is more appropriate to fit a model that relies on a distribution different from the normal.

One of the most common distributions for this cases is the Poisson, which can be incorporated through a general linear model (or GLM). The underlying assumption here is that instead of $T_{ij} \sim N(\mu_{ij}, \sigma)$, our model now follows:

$$T_{ij} \sim Poisson(\exp^{X_{ij}\beta})$$

As usual, such a model is easy to run in R:

```
m2 <- glm('trips15 ~ straight_dist + total_up + total_down',
          data=db_std,
          family=poisson,
          )
```

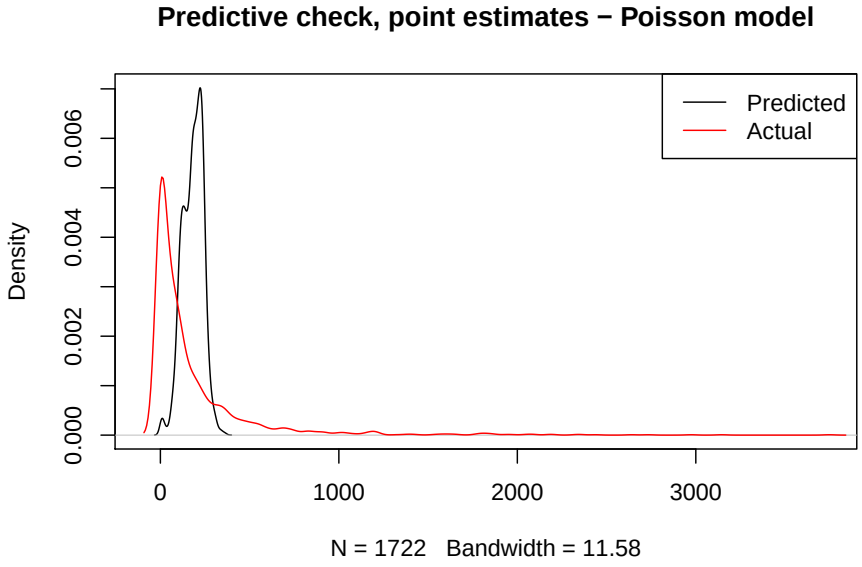
Now let's see how much better, if any, this approach is. To get a quick overview, we can simply plot the point predictions:

```
plot(density(m2$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
                  max(density(m2$fitted.values)$y),
                  max(density(db_std$trips15)$y)
                  )
                )
```

[7] These principles are general and can be applied to pretty much any modeling exercise you run into. The specific approaches we take in this note relate to spatial interaction models

[8] For an illustration of this, consider the amount of probability mass to the left of zero in the predictive checks above.

```
      ),
     col='black',
     main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Predicted', 'Actual'),
       col=c('black', 'red'),
       lwd=1)
title(main="Predictive check, point estimates - Poisson model")
```



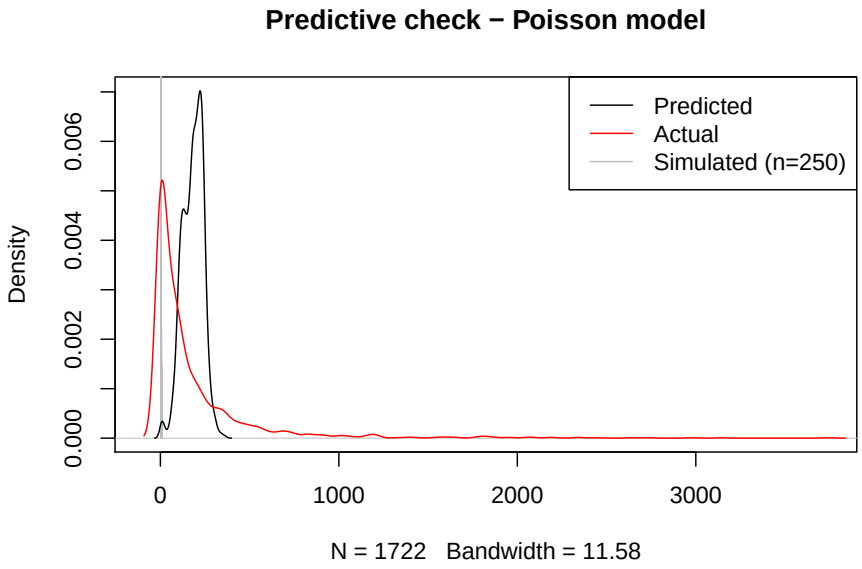
To incorporate uncertainty to these predictions, we need to tweak our `generate_draw` function so it accommodates the fact that our model is not linear anymore.

```
generate_draw_poi <- function(m){
  # Set up predictors matrix
  x <- model.matrix(m)
  # Obtain draws of parameters (inferential uncertainty)
  sim_bs <- sim(m, 1)
  # Predicted value
  xb <- x %*% sim_bs@coef[1, ]
  #xb <- x %*% m$coefficients
  # Transform using the link function
```

```
  mu <- exp(xb)
  # Obtain a random realization
  y_hat <- rpois(n=length(mu), lambda=mu)
  return(y_hat)
}
```

And then we can examine both point predictions an uncertainty around them:

```
plot(density(m2$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
                  max(density(m2$fitted.values)$y),
                  max(density(db_std$trips15)$y)
                  )
                )
      ),
     col='white',
     main='')
# Loop for realizations
for(i in 1:250){
  tmp_y <- generate_draw_poi(m2)
  lines(density(tmp_y),
        col='grey',
        lwd=0.1
        )
}
#
lines(density(m2$fitted.values),
      col='black',
      main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Predicted', 'Actual', 'Simulated (n=250)'),
       col=c('black', 'red', 'grey'),
       lwd=1)
title(main="Predictive check - Poisson model")
```



Voila! Although the curve is still a bit off, centered too much to the right of the actual data, our predictive simulation leaves the fitted values right in the middle. This speaks to a better fit of the model to the actual distribution othe original data follow.

- **Recognize the structure of your data**

So far, we've treated our dataset as if it was flat (i.e. comprise of fully independent realizations) when in fact it is not. Most crucially, our baseline model does not account for the fact that every observation in the dataset pertains to a trip between two stations. This means that all the trips from or to the same station probably share elements which likely help explain how many trips are undertaken between stations. For example, think of trips to an from a station located in the famous Embarcadero, a popular tourist spot. Every route to and from there probably has more trips due to the popularity of the area and we are currently not acknowledging it in the model.

A simple way to incorporate these effects into the model is through origin and destination fixed effects. This approach shares elements with both spatial fixed effects and multilevel modeling and essentially consists of including a binary variable for every origin and destination station. In mathematical notation, this equates to:

$$T_{ij} = X_{ij}\beta + \delta_i + \delta_j + \epsilon_{ij}$$

where δ_i and δ_j are origin and destination station fixed effects[9], and the rest is as above. This strategy accounts for all the unobserved heterogeneity associated with the location of the station. Technically speaking, we simply need to introduce `orig` and `dest` in the the model:

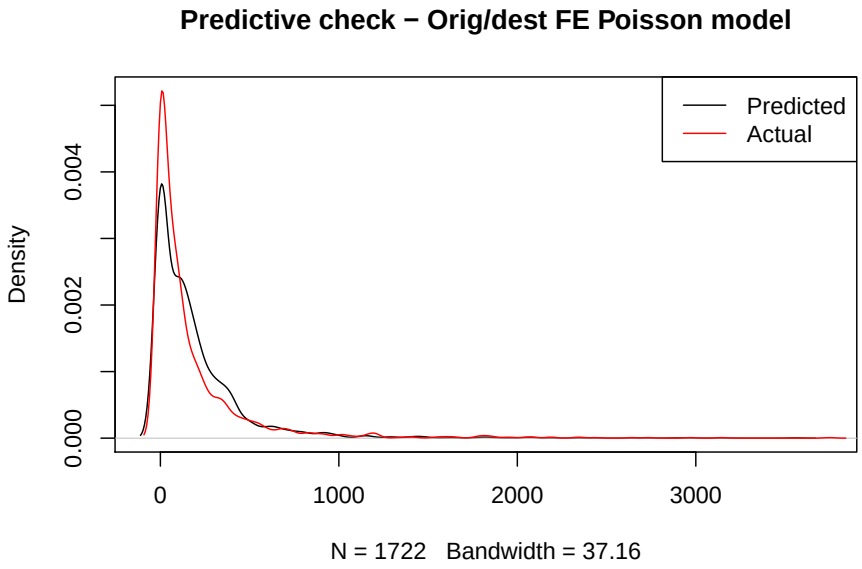
```
m3 <- glm('trips15 ~ straight_dist + total_up + total_down + orig + dest',
          data=db_std,
          family=poisson)
```

And with our new model, we can have a look at how well it does at predicting the overall number of trips[10]:

```
plot(density(m3$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
                  max(density(m3$fitted.values)$y),
                  max(density(db_std$trips15)$y)
                  )
               )
      ),
     col='black',
     main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Predicted', 'Actual'),
       col=c('black', 'red'),
       lwd=1)
title(main="Predictive check - Orig/dest FE Poisson model")
```

[9] In this session, δ_i and δ_j are estimated as independent variables so their estimates are similar to interpret to those in β. An alternative approach could be to model them as random effects in a multilevel framework.

[10] Although, theoretically, we could also include simulations of the model in the plot to get a better sense of the uncertainty behind our model, in practice this seems troublesome. The problems most likely arise from the fact that many of the origin and destination binary variable coefficients are estimated with a great deal of uncertainty. This causes some of the simulation to generate extreme values that, when passed through the exponential term of the Poisson link function, cause problems. If anything, this is testimony of how a simple fixed effect model can sometimes lack accuracy and generate very uncertain estimates. A potential extension to work around these problems could be to fit a multilevel model with two specific levels beyond the trip-level: one for origin and another one for destination stations.



That looks significantly better, doesn't it? In fact, our model now better accounts for the long tail where a few routes take a lot of trips. This is likely because the distribution of trips is far from random across stations and our origin and destination fixed effects do a decent job at accounting for that structure. However our model is still notably underpredicting less popular routes and overpredicting routes with above average number of trips. Maybe we should think about moving beyond a simple linear model.

- **Get better predictors**

The final extension is, in principle, always available but, in practice, it can be tricky to implement. The core idea is that your baseline model might not have the best measurement of the phenomena you want to account for. In our example, we can think of the distance between stations. So far, we have been including the distance measured "as the crow flies" between stations. Although in some cases this is a good approximation (particularly when distances are long and likely route taken is as close to straight as possible), in some cases like ours, where the street layout and the presence of elevation probably matter more than the actual final distance pedalled, this is not necessarily a safe assumption.

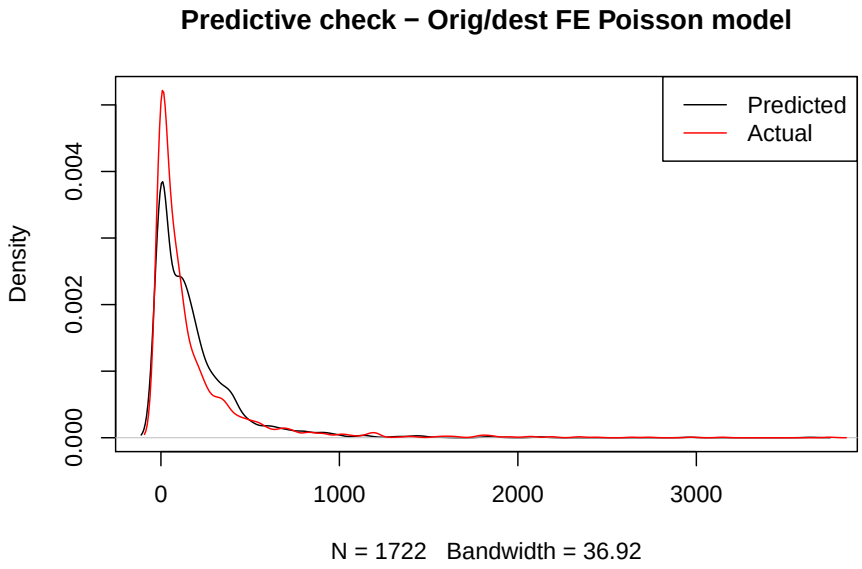
As an exampe of this approach, we can replace the straight distance measurements for more refined ones based on the Google Maps API routes. This is very easy as all we need to do (once the distances have been calculated!) is to swap `straight_dist` for `street_dist`:

```
m4 <- glm('trips15 ~ street_dist + total_up + total_down + orig + dest',
          data=db_std,
```

```
          family=poisson)
```

And we can similarly get a sense of our predictive fitting with:

```
plot(density(m4$fitted.values),
     xlim=c(-100, max(db_std$trips15)),
     ylim=c(0, max(c(
                  max(density(m4$fitted.values)$y),
                  max(density(db_std$trips15)$y)
                  )
                )
      ),
     col='black',
     main='')
lines(density(db_std$trips15),
      col='red',
      main='')
legend('topright',
       c('Predicted', 'Actual'),
       col=c('black', 'red'),
       lwd=1)
title(main="Predictive check - Orig/dest FE Poisson model")
```



Hard to tell any noticeable difference, right? To see if there is any, we can have a look at the estimates obtained:

```
summary(m4)$coefficients['street_dist', ]
```

```
##       Estimate     Std. Error        z value       Pr(>|z|)
##  -9.961619e-02   2.688731e-03  -3.704952e+01  1.828096e-300
```

And compare this to that of the straight distances in the previous model:

```
summary(m3)$coefficients['straight_dist', ]
```

```
##       Estimate     Std. Error        z value       Pr(>|z|)
##  -7.820014e-02   2.683052e-03  -2.914596e+01  9.399407e-187
```

As we can see, the differences exist but ar not massive. Let's use this example to learn how to interpret coefficients in a Poisson model[11]. Effectively, these estimates can be understood as multiplicative effects. Since our model fits

$$T_{ij} \sim Poisson(\exp^{X_{ij}\beta})$$

we need to transform β through an exponential in order to get a sense of the effect of distance on the number of trips. This means that for the street distance, our original estimate is $\beta_{street} = -0.0996$, but this needs to be translated through the exponential into $e^{-0.0996} = 0.906$. In other words, since distance is expressed in standard deviations[12], we can expect a 10% decrease in the number of trips for an increase of one standard deviation (about 1Km) in the distance between the stations. This can be compared with $e^{-0.0782} = 0.925$ for the straight distances, or a reduction of about 8% the number of trips for every increase of a standard deviation (about 720m).

4.5 Predicting flows

So far we have put all of our modeling efforts in understanding the model we fit and improving such model so it fits our data as closely as possible. This is essential in any modelling exercise but should be far from a stopping point. Once we're confident our model is a decent representation of the data generating process, we can start exploiting it. In this section, we will cover one specific case that showcases how a fitted model can help: out-of-sample forecasts.

It is August 2015, and you have just started working as a data scientist for the bikeshare company that runs the San Francisco system. You join them as they're planning for the next academic year and, in order to plan their operations (re-allocating vans, station maintenance, etc.), they need to get a sense of how many people are going to be pedalling across the city and, crucially, *where* they are going to be pedalling through. What can you do to help them?

[11] See section 6.2 of Gelman and Hill (2006) for a similar treatment of these.
[12] Remember the transformation at the very beginning.

The easiest approach is to say "well, a good guess for how many people will be going between two given stations this coming year is how many went through last year, isn't it?". This is one prediction approach. However, you could see how, even if the same process governs over both datasets (2015 and 2016), each year will probably have some idiosyncracies and thus looking too closely into one year might not give the best possible answer for the next one. Ideally, you want a good stylized synthesis that captures the bits that stay constant over time and thus can be applied in the future and that ignores those aspects that are too particular to a given point in time. That is the rationale behind using a fitted model to obtain predictions.

However good any theory though, the truth is in the pudding. So, to see if a modeling approach is better at producing forecasts than just using the counts from last year, we can put them to a test. The way this is done when evaluating the predictive performance of a model (as this is called in the literature) relies on two basic steps: a) obtain predictions from a given model and b) compare those to the actual values (in our case, with the counts for 2016 in `trips16`) and get a sense of "how off" they are. We have essentially covered a) above; for b), there are several measures to use. We will use one of the most common ones, the root mean squared error (RMSE), which roughly gives a sense of the average difference between a predicted vector and the real deal:

$$RMSE = \sqrt{\sum_{ij} (\hat{T_{ij}} - T_{ij})^2}$$

where $\hat{T_{ij}}$ is the predicted amount of trips between stations i and j. RMSE is straightforward in R and, since we will use it a couple of times, let's write a short function to make our lives easier:

```
rmse <- function(t, p){
  se <- (t - p)^2
  mse <- mean(se)
  rmse <- sqrt(mse)
  return(rmse)
}
```

where `t` stands for the vector of true values, and `p` is the vector of predictions. Let's give it a spin to make sure it works:

```
rmse_m4 <- rmse(db_std$trips16, m4$fitted.values)
rmse_m4
```

```
## [1] 256.2197
```

That means that, on average, predictions in our best model `m4` are 256 trips off. Is this good? Bad? Worse? It's hard to say but, being practical, what we can

say is whether this better than our alternative. Let us have a look at the RMSE of the other models as well as that of simply plugging in last year's counts:[13]

```
rmses <- data.frame(model=c('OLS', 'Poisson', 'Poisson + FE',
                            'Poisson + FE + street dist.',
                            'Trips-2015'
                            ),
                    RMSE=c(rmse(db_std$trips16,
                                m1$fitted.values),
                           rmse(db_std$trips16,
                                m2$fitted.values),
                           rmse(db_std$trips16,
                                m3$fitted.values),
                           rmse(db_std$trips16,
                                m4$fitted.values),
                           rmse(db_std$trips16,
                                db_std$trips15)
                           )
                    )
rmses
```

```
##                         model     RMSE
## 1                         OLS 323.6135
## 2                     Poisson 320.8962
## 3                 Poisson + FE 254.4468
## 4 Poisson + FE + street dist. 256.2197
## 5                   Trips-2015 131.0228
```

The table is both encouraging and disheartning at the same time. On the one hand, all the modeling techniques covered above behave as we would expect: the baseline model displays the worst predicting power of all, and every improvement (except the street distances!) results in notable decreases of the RMSE. This is good news. However, on the other hand, all of our modelling efforts fall short of given a better guess than simply using the previous year's counts. *Why? Does this mean that we should not pay attention to modeling and inference?* Not really. Generally speaking, a model is as good at predicting as it is able to mimic the underlying process that gave rise to the data in the first place. The results above point to a case where our model is not picking up all the factors that determine the amount of trips undertaken in a give route. This could be improved by enriching the model with more/better predictors, as we have seen above. Also, the example above seems to point to a case where those idiosyncracies in 2015 that the model does not pick up seem to be at work in 2016 as well. This is great news for our prediction efforts this time, but we have no idea why this is the case and, for all that matters, it could change the coming year. Besides the elegant quantification of uncertainty, the true advantage of a modeling approach

[13] **EXERCISE**: can you create a single plot that displays the distribution of the predicted values of the five different ways to predict trips in 2016 and the actual counts of trips?

in this context is that, if well fit, it is able to pick up the fundamentals that apply over and over. This means that, if next year we're not as lucky as this one and previous counts are not good predictors but the variables we used in our model continue to have a role in determining the outcome, the data scientist should be luckier and hit a better prediction.

4.6 References

Chapter 5

Spatial Econometrics

This chapter[1] is based on the following references, which are good follow-up's on the topic:

- Session III of Arribas-Bel (2014). Check the "Related readings" section on the session page for more in-depth discussions.
- Anselin (2007), freely available to download [`pdf`].
- The second part of this tutorial assumes you have reviewed Lecture 5 of Arribas-Bel (2019). [html]
- A similar coverage of topics is presented in Chapter 11 of the upcoming book "Geographic Data Science with Python", which is currently in progress. [`url`]

This tutorial is part of Spatial Analysis Notes, a compilation hosted as a GitHub repository that you can access it in a few ways:

- As a download of a `.zip` file that contains all the materials.
- As an html website.
- As a pdf document
- As a GitHub repository.

5.1 Dependencies

The illustration below relies on the following libraries that you will need to have installed on your machine to be able to interactively follow along[2]. Once

[1] Spatial Econometrics by Dani Arribas-Bel is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

[2] You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

installed, load them up with the following commands:

```
# Layout
library(tufte)
# For pretty table
library(knitr)
# Spatial Data management
library(rgdal)
```

```
## Loading required package: sp
```

```
## rgdal: version: 1.4-4, (SVN revision 833)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: /usr/share/gdal/2.2
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: (autodetected)
##  Linking to sp version: 1.3-1
```

```
# Pretty graphics
library(ggplot2)
# Pretty maps
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
# Various GIS utilities
library(GISTools)
```

```
## Loading required package: maptools
```

```
## Checking rgeos availability: TRUE
```

```
## Loading required package: RColorBrewer
```

```
## Loading required package: MASS
```

```
## Loading required package: rgeos
```

```
## rgeos version: 0.5-1, (SVN revision 614)
##  GEOS runtime version: 3.6.2-CAPI-1.10.2
##  Linking to sp version: 1.3-1
##  Polygon checking: TRUE
```

```
# For all your interpolation needs
library(gstat)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
```

```
##   as.zoo.xts zoo
```

```
# For data manipulation
library(plyr)
# Spatial regression
library(spdep)
```

```
## Loading required package: spData
```

```
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
```

```
## Loading required package: sf
```

```
## Linking to GEOS 3.6.2, GDAL 2.2.3, PROJ 4.9.3
```

Before we start any analysis, let us set the path to the directory where we are working. We can easily do that with `setwd()`. Please replace in the following line the path to the folder where you have placed this file -and where the `house_transactions` folder with the data lives.

```
#setwd('/media/dani/baul/AAA/Documents/teaching/u-lvl/2016/envs453/code/GIT/kde_idw_r/')
setwd('.')
```

5.2 Data

To explore ideas in spatial regression, we will be using house price data for the municipality of Liverpool. Our main dataset is provided by the Land Registry (as part of their Price Paid Data) but has been cleaned and re-packaged into a shapefile by Dani Arribas-Bel.

Let us load it up first of all:

```
hst <- readOGR(dsn = 'data/house_transactions', layer = 'liv_house_trans')
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/home/jovyan/work/data/house_transactions", layer: "liv_house_trans"
## with 6324 features
## It has 18 fields
## Integer64 fields read as strings:  price
```

The tabular component of the spatial frame contains the followig variables:

```
names(hst)
```

```
##  [1] "pcds"       "id"         "price"      "trans_date" "type"
##  [6] "new"        "duration"   "paon"       "saon"       "street"
## [11] "locality"   "town"       "district"   "county"     "ppd_cat"
## [16] "status"     "lsoa11"     "LSOA11CD"
```

The meaning for most of the variables can be found in the original Land Registry documentation. The dataset contains transactions that took place during 2,014:

```
# Format dates
dts <- as.Date(hst@data$trans_date)
# Set up summary table
tab <- summary(dts)
tab
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2014-01-02" "2014-04-11" "2014-07-09" "2014-07-08" "2014-10-03"
##         Max.
## "2014-12-30"
```

Although the original Land Registry data contain some characteristics of the house, all of them are categorical: *is the house newly built? What type of property is it?* To bring in a richer picture and illustrate how continuous variables can also be included in a spatial setting, we will augment the original transaction data with Deprivation indices from the CDRC at the Lower Layer Super Output Area (LSOA) level.

Let us read the `csv` in:

```
imd <- read.csv('data/house_transactions/E08000012.csv')
```

The table contains not only the overall IMD score and rank, but some of the component scores, as well as the `LSOA` code:

```
names(imd)
```

```
##  [1] "LSOA11CD"   "imd_rank"   "imd_score"  "income"     "employment"
##  [6] "education"  "health"     "crime"      "housing"    "living_env"
## [11] "idaci"      "idaopi"
```

That bit of information, `LSOA11CD`, is crucial to be able to connect it to each house transaction. To "join" both tables, we can use the base command `merge`, which will assign values from `imd` into `hst` making sure that each house transaction get the IMD data for the LSOA where it is located:

```
db <- merge(hst, imd)
```

The resulting table, `db`, contains variables from both original tables:

```
names(db)
```

```
##  [1] "LSOA11CD"   "pcds"       "id"         "price"      "trans_date"
##  [6] "type"       "new"        "duration"   "paon"       "saon"
## [11] "street"     "locality"   "town"       "district"   "county"
## [16] "ppd_cat"    "status"     "lsoa11"     "imd_rank"   "imd_score"
## [21] "income"     "employment" "education"  "health"     "crime"
## [26] "housing"    "living_env" "idaci"      "idaopi"
```

Since we will heavily rely on `price`, we need to turn it into a numeric column, rather than as a factor, which is how it is picked up:

```
db@data$price <- as.numeric(as.character(db@data$price))
```

For some of our analysis, we will need the coarse postcode of each house, rather than the finely specified one in the original data. This means using the available one

```
head(db@data['pcds'])
```

```
##      pcds
## 62 L1 0AB
## 63 L1 0AB
## 64 L1 0AB
## 65 L1 0AB
## 66 L1 0AB
## 67 L1 0AB
```

to create a new column that only contains the first bit of the postcode (`L1` in the examples above). The following lines of code will do that for us:

```
db$pc <- as.character(lapply(strsplit(as.character(db$pcds), split=" "), "[", 1))
```

Given there are 6,324 transactions in the dataset, a simple plot of the point coordinates implicitly draws the shape of the Liverpool municipality:

```
plot(db)
```

5.3 Non-spatial regression, a refresh

Before we discuss how to explicitly include space into the linear regression framework, let us show how basic regression can be carried out in R, and how you can begin to interpret the results. By no means is this a formal and complete introduction to regression so, if that is what you are looking for, I suggest the first part of Gelman and Hill (2006), in particular chapters 3 and 4.

The core idea of linear regression is to explain the variation in a given (*dependent*) variable as a linear function of a series of other (*explanatory*) variables. For example, in our case, we may want to express/explain the price of a house as a function of whether it is new and the degree of deprivation of the area where it is located. At the individual level, we can express this as:

$$P_i = \alpha + \beta_1 NEW_i + \beta_2 IMD_i + \epsilon_i$$

where P_i is the price of house i, NEW_i is a binary variable that takes one if the house is newly built or zero otherwise and IMD_i is the IMD score of the LSOA

Figure 5.1: Spatial distribution of house transactions in Liverpool

where i is located. The parameters β_1, β_2, and β_3 give us information about in which way and to what extent each variable is related to the price, and α, the constant term, is the average house price when all the other variables are zero. The term ϵ_i is usually referred to as "error" and captures elements that influence the price of a house but are not whether the house is new or the IMD score of its area. We can also express this relation in matrix form, excluding subindices for i[3].

Essentially, a regression can be seen as a multivariate extension of simple bivariate correlations. Indeed, one way to interpret the β_k coefficients in the equation above is as the degree of correlation between the explanatory variable k and the dependent variable, *keeping all the other explanatory variables constant*. When you calculate simple bivariate correlations, the coefficient of a variable is picking up the correlation between the variables, but it is also subsuming into it variation associated with other correlated variables –also called confounding factors[4].

[3] In this case, the equation would look like

$$P = \alpha + \beta_1 NEW + \beta_2 IMD + \epsilon$$

and would be interpreted in terms of vectors and matrices instead of scalar values.

[4] **EXAMPLE** Assume that new houses tend to be built more often in areas with low deprivation. If that is the case, then NEW and IMD will be correlated with each other (as well as with the price of a house, as we are hypothesizing in this case). If we calculate a

Regression allows you to isolate the distinct effect that a single variable has on the dependent one, once we *control* for those other variables.

Practically speaking, running linear regressions in `R` is straightforward. For example, to fit the model specified in the equation above, we only need one line of code:

```
m1 <- lm('price ~ new + imd_score', db)
```

We use the command `lm`, for linear model, and specify the equation we want to fit using a string that relates the dependent variable (`price`) with a set of explanatory ones (`new` and `price`) by using a tilde `~` that is akin the = symbol in the mathematical equation. Since we are using names of variables that are stored in a table, we need to pass the table object (`db`) as well.

In order to inspect the results of the model, the quickest way is to call `summary`:

```
summary(m1)
```

```
##
## Call:
## lm(formula = "price ~ new + imd_score", data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -184254   -59948   -29032    11430 26434741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   235596      13326  17.679  < 2e-16 ***
## newY            4926      19104   0.258    0.797
## imd_score      -2416        308  -7.843 5.12e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 509000 on 6321 degrees of freedom
## Multiple R-squared:  0.009712,   Adjusted R-squared:  0.009398
## F-statistic: 30.99 on 2 and 6321 DF,  p-value: 4.027e-14
```

A full detailed explanation of the output is beyond the scope of this note, so we will focus on the relevant bits for our main purpose. This is concentrated on the `Coefficients` section, which gives us the estimates for the β_k coefficients in our model. Or, in other words, the coefficients are the raw equivalent of

simple correlation between P and IMD, the coefficient will represent the degree of association between both variables, but it will also include some of the association between IMD and NEW. That is, part of the obtained correlation coefficient will be due not to the fact that higher prices tend to be found in areas with low IMD, but to the fact that new houses tend to be more expensive. This is because (in this example) new houses tend to be built in areas with low deprivation and simple bivariate correlation cannot account for that.

the correlation coefficient between each explanatory variable and the dependent one, once the polluting effect of confounding factors has been accounted for[5]. Results are as expected for the most part: houses tend to be significantly more expensive in areas with lower deprivation (an average of GBP2,416 for every additional score); and a newly built house is on average GBP4,926 more expensive, although this association cannot be ruled out to be random (probably due to the small relative number of new houses).

Finally, before we jump into introducing space in our models, let us modify our equation slightly to make it more useful when it comes to interpretating it. Many house price models in the literature is estimated in log-linear terms:

$$\log P_i = \alpha + \beta_1 NEW_i + \beta_2 IMD_i + \epsilon_i$$

This allows to interpret the coefficients more directly: as the percentual change induced by a unit increase in the explanatory variable of the estimate. To fit such a model, we can specify the logarithm of a given variable directly in the formula.

```
m2 <- lm('log(price) ~ new + imd_score', db)
summary(m2)
```

```
##
## Call:
## lm(formula = "log(price) ~ new + imd_score", data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3172 -0.3231 -0.0149  0.3063  5.2769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.2037784  0.0138634  880.28   <2e-16 ***
## newY         0.2456446  0.0198740   12.36   <2e-16 ***
## imd_score   -0.0169702  0.0003204  -52.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5295 on 6321 degrees of freedom
## Multiple R-squared:  0.3101, Adjusted R-squared:  0.3099
## F-statistic:  1421 on 2 and 6321 DF,  p-value: < 2.2e-16
```

Looking at the results we can see a couple of differences with respect to the original specification. First, the estimates are substantially different numbers. This is because, although they consider the same variable, the look at it from

[5]Keep in mind that regression is no magic. We are only discounting the effect of other confounding factors that we include in the model, not of *all* potentially confounding factors.

different angles, and provide different interpretations. For example, the coefficient for the IMD, instead of being interpretable in terms of GBP, the unit of the dependent variable, it represents a percentage: a unit increase in the degree of deprivation is associated with a 0.2% decrease in the price of a house.[6] Second, the variable `new` is significant in this case. This is probably related to the fact that, by taking logs, we are also making the dependent variable look more normal (Gaussian) and that allows the linear model to provide a better fit and, hence, more accurate estimates. In this case, a house being newly built, as compared to an old house, is overall 25% more expensive.

5.4 Spatial regression: a (very) first dip

Spatial regression is about *explicitly* introducing space or geographical context into the statistical framework of a regression. Conceptually, we want to introduce space into our model whenever we think it plays an important role in the process we are interested in, or when space can act as a reasonable proxy for other factors we cannot but should include in our model. As an example of the former, we can imagine how houses at the seafront are probably more expensive than those in the second row, given their better views. To illustrate the latter, we can think of how the character of a neighborhood is important in determining the price of a house; however, it is very hard to identify and quantify "character" perse, although it might be easier to get at its spatial variation, hence a case of space as a proxy.

Spatial regression is a large field of development in the econometrics and statistics literatures. In this brief introduction, we will consider two related but very different processes that give rise to spatial effects: spatial heterogeneity and spatial dependence. For more rigorous treatments of the topics introduced here, the reader is referred to Anselin (2003), Anselin and Rey (2014), and Gibbons et al. (2014).

5.5 Spatial heterogeneity

Spatial heterogeneity (SH) arises when we cannot safely assume the process we are studying operates under the same "rules" throughout the geography of interest. In other words, we can observe SH when there are effects on the outcome variable that are intrinsically linked to specific locations. A good example of this is the case of seafront houses above: we are trying to model the price of

[6] **EXERCISE** *How does the type of a house affect the price at which it is sold, given whether it is new and the level of deprivation of the area where it is located?* To answer this, fit a model as we have done but including additionally the variable `type`. In order to interpret the codes, check the reference at the Land Registry documentation.

a house and, the fact some houses are located under certain conditions (i.e. by the sea), makes their price behave differently[7].

This somewhat abstract concept of SH can be made operational in a model in several ways. We will explore the following two: spatial fixed-effects (FE); and spatial regimes, which is a generalization of FE.

Spatial FE

Let us consider the house price example from the previous section to introduce a more general illustration that relates to the second motivation for spatial effects ("space as a proxy"). Given we are only including two explanatory variables in the model, it is likely we are missing some important factors that play a role at determining the price at which a house is sold. Some of them, however, are likely to vary systematically over space (e.g. different neighborhood characteristics). If that is the case, we can control for those unobserved factors by using traditional dummy variables but basing their creation on a spatial rule. For example, let us include a binary variable for every two-digit postcode in Liverpool, indicating whether a given house is located within such area (`1`) or not (`0`). Mathematically, we are now fitting the following equation:

$$\log P_i = \alpha_r + \beta_1 NEW_i + \beta_2 IMD_i + \epsilon_i$$

where the main difference is that we are now allowing the constant term, α, to vary by postcode r, α_r.

Programmatically, this is straightforward to estimate:

```
# Include `-1` to eliminate the constant term and include a dummy for every area
m3 <- lm('log(price) ~ pc + new + imd_score - 1', db)
summary(m3)
```

```
##
## Call:
## lm(formula = "log(price) ~ pc + new + imd_score - 1", data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2680 -0.2833 -0.0235  0.2599  5.6714
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## pcL1      11.697166   0.032137  363.98   <2e-16 ***
## pcL10     11.893524   0.076471  155.53   <2e-16 ***
## pcL11     11.902319   0.043194  275.55   <2e-16 ***
## pcL12     12.065776   0.029692  406.37   <2e-16 ***
```

[7] **QUESTION** How would you incorporate this into a regression model that extends the log-log equation of the previous section?

```
## pcL13      11.865523   0.034893  340.06   <2e-16 ***
## pcL14      11.920922   0.044480  268.01   <2e-16 ***
## pcL15      12.039916   0.028555  421.64   <2e-16 ***
## pcL16      12.295535   0.034950  351.81   <2e-16 ***
## pcL17      12.275339   0.027978  438.75   <2e-16 ***
## pcL18      12.396475   0.024534  505.27   <2e-16 ***
## pcL19      12.162630   0.029697  409.56   <2e-16 ***
## pcL2       12.061443   0.100805  119.65   <2e-16 ***
## pcL20      11.928142   0.226932   52.56   <2e-16 ***
## pcL24      11.868363   0.045785  259.22   <2e-16 ***
## pcL25      12.234462   0.028557  428.42   <2e-16 ***
## pcL27      12.035241   0.092832  129.65   <2e-16 ***
## pcL28      11.438267   0.206323   55.44   <2e-16 ***
## pcL3       11.954453   0.029561  404.40   <2e-16 ***
## pcL4       11.718609   0.039575  296.11   <2e-16 ***
## pcL5       12.037267   0.055952  215.14   <2e-16 ***
## pcL6       11.898506   0.042918  277.24   <2e-16 ***
## pcL7       11.855374   0.044681  265.33   <2e-16 ***
## pcL8       12.034093   0.039801  302.36   <2e-16 ***
## pcL9       11.759056   0.033610  349.87   <2e-16 ***
## newY        0.310829   0.020947   14.84   <2e-16 ***
## imd_score -0.012008   0.000517  -23.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5007 on 6298 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9981
## F-statistic: 1.305e+05 on 26 and 6298 DF,  p-value: < 2.2e-16
```

Econometrically speaking, what the postcode FE we have introduced imply is that, instead of comparing all house prices across Liverpool as equal, we only derive variation from within each postcode[8]. Remember that the interpretation of a β_k coefficient is the effect of variable k, *given all the other explanatory variables included remain constant*. By including a single variable for each area, we are effectively forcing the model to compare as equal only house prices that share the same value for each variable; in other words, only houses located within the same area. Introducing FE affords you a higher degree of isolation of the effects of the variables you introduce in your model because you can control for unobserved effects that align spatially with the distribution of the FE you introduce (by postcode, in our case).

[8]Additionally, estimating spatial FE in our particular example also gives you an indirect measure of area *desirability*: since they are simple dummies in a regression explaining the price of a house, their estimate tells us about how much people are willing to pay to live in a given area. However, this interpretation does not necessarily apply in other contexts where introducing spatial FEs does make sense. **EXERCISE** *What is the most desired area to live in Liverpool?*

Spatial regimes

At the core of estimating spatial FEs is the idea that, instead of assuming the dependent variable behaves uniformly over space, there are systematic effects following a geographical pattern that affect its behaviour. In other words, spatial FEs introduce econometrically the notion of spatial heterogeneity. They do this in the simplest possible form: by allowing the constant term to vary geographically. The other elements of the regression are left untouched and hence apply uniformly across space. The idea of spatial regimes (SRs) is to generalize the spatial FE approach to allow not only the constant term to vary but also any other explanatory variable. This implies that the equation we will be estimating is:

$$\log P_i = \alpha_r + \beta_{1r} NEW_i + \beta_{2r} IMD_i + \epsilon_i$$

where we are not only allowing the constant term to vary by region (α_r), but also every other parameter (β_{kr}).

In R terms, this is more straightforward to estimate if `new` is expressed as `0` and `1`, rather than as factors:

```
# Create a new variable `newB` to store the binary form of `new`
db@data$newB <- 1
db[db@data$new=='N', 'newB'] <- 0
```

Also, given we are going to allow *every* coefficient to vary by regime, we will need to explicitly set a constant term that we can allow to vary:

```
db$one <- 1
```

Then, the estimation leverages the capabilities in model description of R formulas:

```
# `:` notation implies interaction variables
m4 <- lm('log(price) ~ 0 +(one + newB + imd_score):(pc)', db)
summary(m4)
```

```
##
## Call:
## lm(formula = "log(price) ~ 0 +(one + newB + imd_score):(pc)",
##     data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2051 -0.2506 -0.0182  0.2198  5.3507
##
## Coefficients: (8 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## one:pcL1        11.983460   0.098675 121.444  < 2e-16 ***
```

```
## one:pcL10       11.911281   0.504210  23.624  < 2e-16 ***
## one:pcL11       11.774865   0.160864  73.198  < 2e-16 ***
## one:pcL12       12.157627   0.068165 178.356  < 2e-16 ***
## one:pcL13       11.821576   0.103946 113.728  < 2e-16 ***
## one:pcL14       11.821895   0.115040 102.763  < 2e-16 ***
## one:pcL15       12.317524   0.052051 236.645  < 2e-16 ***
## one:pcL16       12.761711   0.098709 129.287  < 2e-16 ***
## one:pcL17       12.240652   0.065074 188.103  < 2e-16 ***
## one:pcL18       12.695103   0.056510 224.654  < 2e-16 ***
## one:pcL19       12.489604   0.054177 230.533  < 2e-16 ***
## one:pcL2        12.164848   0.321558  37.831  < 2e-16 ***
## one:pcL20       11.069199   0.211083  52.440  < 2e-16 ***
## one:pcL24       10.876938   0.135326  80.376  < 2e-16 ***
## one:pcL25       12.427878   0.051147 242.982  < 2e-16 ***
## one:pcL27       11.022835   0.435730  25.297  < 2e-16 ***
## one:pcL28       11.300143   1.126387  10.032  < 2e-16 ***
## one:pcL3        11.977708   0.054781 218.647  < 2e-16 ***
## one:pcL4        11.653464   0.109113 106.802  < 2e-16 ***
## one:pcL5        11.544085   0.277000  41.675  < 2e-16 ***
## one:pcL6        11.574617   0.158022  73.247  < 2e-16 ***
## one:pcL7        11.617544   0.128402  90.478  < 2e-16 ***
## one:pcL8        11.586224   0.085036 136.251  < 2e-16 ***
## one:pcL9        11.837544   0.080452 147.139  < 2e-16 ***
## newB:pcL1       -0.298430   0.051078  -5.843 5.40e-09 ***
## newB:pcL10             NA         NA      NA       NA
## newB:pcL11       0.672479   0.089894   7.481 8.40e-14 ***
## newB:pcL12       0.977903   0.114281   8.557  < 2e-16 ***
## newB:pcL13      -0.659910   0.473773  -1.393 0.163705
## newB:pcL14       0.610482   0.132911   4.593 4.45e-06 ***
## newB:pcL15       0.744844   0.139732   5.331 1.01e-07 ***
## newB:pcL16             NA         NA      NA       NA
## newB:pcL17      -0.094580   0.095188  -0.994 0.320449
## newB:pcL18      -0.197115   0.122109  -1.614 0.106521
## newB:pcL19       0.393901   0.089607   4.396 1.12e-05 ***
## newB:pcL2              NA         NA      NA       NA
## newB:pcL20             NA         NA      NA       NA
## newB:pcL24       0.735316   0.072739  10.109  < 2e-16 ***
## newB:pcL25       0.196018   0.144527   1.356 0.175063
## newB:pcL27       0.356678   0.189946   1.878 0.060457 .
## newB:pcL28             NA         NA      NA       NA
## newB:pcL3       -0.261461   0.055228  -4.734 2.25e-06 ***
## newB:pcL4        1.087380   0.079152  13.738  < 2e-16 ***
## newB:pcL5        0.519945   0.094850   5.482 4.38e-08 ***
## newB:pcL6        0.695856   0.062461  11.141  < 2e-16 ***
## newB:pcL7              NA         NA      NA       NA
## newB:pcL8        0.334517   0.059368   5.635 1.83e-08 ***
```

```
## newB:pcL9              NA         NA      NA       NA
## imd_score:pcL1  -0.010489   0.003383  -3.101 0.001938 **
## imd_score:pcL10 -0.012413   0.011381  -1.091 0.275485
## imd_score:pcL11 -0.010637   0.003005  -3.540 0.000404 ***
## imd_score:pcL12 -0.017296   0.002625  -6.588 4.81e-11 ***
## imd_score:pcL13 -0.011004   0.002185  -5.036 4.89e-07 ***
## imd_score:pcL14 -0.010432   0.002469  -4.225 2.42e-05 ***
## imd_score:pcL15 -0.020737   0.001413 -14.681  < 2e-16 ***
## imd_score:pcL16 -0.050522   0.007703  -6.559 5.85e-11 ***
## imd_score:pcL17 -0.009763   0.002214  -4.410 1.05e-05 ***
## imd_score:pcL18 -0.032289   0.003799  -8.499  < 2e-16 ***
## imd_score:pcL19 -0.024629   0.001860 -13.242  < 2e-16 ***
## imd_score:pcL2  -0.016601   0.013654  -1.216 0.224084
## imd_score:pcL20         NA         NA      NA       NA
## imd_score:pcL24  0.004090   0.002391   1.711 0.087139 .
## imd_score:pcL25 -0.020461   0.001981 -10.327  < 2e-16 ***
## imd_score:pcL27  0.006626   0.007561   0.876 0.380911
## imd_score:pcL28 -0.009473   0.020367  -0.465 0.641850
## imd_score:pcL3  -0.006935   0.001747  -3.970 7.28e-05 ***
## imd_score:pcL4  -0.012032   0.001731  -6.952 3.96e-12 ***
## imd_score:pcL5  -0.006051   0.004009  -1.509 0.131280
## imd_score:pcL6  -0.008324   0.002367  -3.517 0.000440 ***
## imd_score:pcL7  -0.007517   0.002342  -3.209 0.001337 **
## imd_score:pcL8  -0.004064   0.001561  -2.603 0.009255 **
## imd_score:pcL9  -0.014124   0.002052  -6.882 6.47e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.472 on 6260 degrees of freedom
## Multiple R-squared:  0.9984, Adjusted R-squared:  0.9983
## F-statistic: 5.966e+04 on 64 and 6260 DF,  p-value: < 2.2e-16
```

As we can see, there are a few `NA` values (e.g. `pcL10`). This has to do with the fact that there are not that many new houses, so some of the buckets in which the regimes split the data to estimate each parameter are empty. This can be readily seen by obtaining a quick cross tabulation of `pc` and `new`:

```
table(db$pc, db$new)
```

```
##
##         N   Y
##   L1  161 189
##   L10  47   0
##   L11 192  34
##   L12 326  18
##   L13 387   1
##   L14 144  19
```

```
##   L15 466  12
##   L16 212   0
##   L17 398  27
##   L18 439  16
##   L19 329  32
##   L2   25   0
##   L20   5   0
##   L24  97  84
##   L25 357  11
##   L27  22  10
##   L28   6   0
##   L3  272 101
##   L4  437  41
##   L5   97  46
##   L6  319  78
##   L7  201   0
##   L8  227 110
##   L9  329   0
```

To illustrate a correct regime estimation, we can focus only on `imd_score`[9]:

```
# `:` notation implies interaction variables
m5 <- lm('log(price) ~ 0 + (one + imd_score):pc', db)
summary(m5)
```

```
##
## Call:
## lm(formula = "log(price) ~ 0 + (one + imd_score):pc", data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4952 -0.2779 -0.0221  0.2477  5.4973
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## one:pcL1        11.8965065  0.1030279 115.469  < 2e-16 ***
## one:pcL10       11.9112807  0.5325447  22.367  < 2e-16 ***
## one:pcL11       11.6203268  0.1684973  68.964  < 2e-16 ***
## one:pcL12       12.2106579  0.0716973 170.309  < 2e-16 ***
## one:pcL13       11.8099973  0.1094358 107.917  < 2e-16 ***
## one:pcL14       12.1229804  0.0998496 121.412  < 2e-16 ***
## one:pcL15       12.3753226  0.0537697 230.154  < 2e-16 ***
## one:pcL16       12.7617106  0.1042555 122.408  < 2e-16 ***
## one:pcL17       12.2248054  0.0666348 183.460  < 2e-16 ***
## one:pcL18       12.7033951  0.0594382 213.724  < 2e-16 ***
```

[9]Note this still returns a `NA` for the IMD estimate in `L20`. This is most likely due to the little amount of houses (five) sold in that area. The regression nevertheless serves the illustration

```
## one:pcL19       12.4767606  0.0571384 218.360  < 2e-16 ***
## one:pcL2        12.1648479  0.3396283  35.818  < 2e-16 ***
## one:pcL20       11.0691989  0.2229447  49.650  < 2e-16 ***
## one:pcL24       11.5214275  0.1260747  91.386  < 2e-16 ***
## one:pcL25       12.4351502  0.0537240 231.464  < 2e-16 ***
## one:pcL27       11.3770205  0.4148632  27.424  < 2e-16 ***
## one:pcL28       11.3001428  1.1896847   9.498  < 2e-16 ***
## one:pcL3        11.8873513  0.0542344 219.185  < 2e-16 ***
## one:pcL4        11.4097020  0.1137109 100.340  < 2e-16 ***
## one:pcL5        10.9973874  0.2729459  40.291  < 2e-16 ***
## one:pcL6        11.1816689  0.1626918  68.729  < 2e-16 ***
## one:pcL7        11.6175445  0.1356173  85.664  < 2e-16 ***
## one:pcL8        11.5086161  0.0886286 129.852  < 2e-16 ***
## one:pcL9        11.8375435  0.0849726 139.310  < 2e-16 ***
## imd_score:pcL1  -0.0131286  0.0035407  -3.708 0.000211 ***
## imd_score:pcL10 -0.0124128  0.0120210  -1.033 0.301837
## imd_score:pcL11 -0.0058373  0.0031009  -1.882 0.059824 .
## imd_score:pcL12 -0.0173734  0.0027727  -6.266 3.95e-10 ***
## imd_score:pcL13 -0.0107902  0.0023022  -4.687 2.83e-06 ***
## imd_score:pcL14 -0.0160870  0.0022604  -7.117 1.23e-12 ***
## imd_score:pcL15 -0.0219205  0.0014734 -14.878  < 2e-16 ***
## imd_score:pcL16 -0.0505217  0.0081354  -6.210 5.63e-10 ***
## imd_score:pcL17 -0.0093980  0.0023061  -4.075 4.65e-05 ***
## imd_score:pcL18 -0.0333885  0.0039476  -8.458  < 2e-16 ***
## imd_score:pcL19 -0.0228257  0.0019160 -11.913  < 2e-16 ***
## imd_score:pcL2  -0.0166013  0.0144213  -1.151 0.249707
## imd_score:pcL20         NA         NA      NA       NA
## imd_score:pcL24 -0.0020546  0.0024420  -0.841 0.400163
## imd_score:pcL25 -0.0205241  0.0020921  -9.810  < 2e-16 ***
## imd_score:pcL27  0.0020943  0.0075687   0.277 0.782014
## imd_score:pcL28 -0.0094733  0.0215112  -0.440 0.659671
## imd_score:pcL3  -0.0061811  0.0018374  -3.364 0.000773 ***
## imd_score:pcL4  -0.0066455  0.0017803  -3.733 0.000191 ***
## imd_score:pcL5   0.0039338  0.0037726   1.043 0.297115
## imd_score:pcL6  -0.0004620  0.0023860  -0.194 0.846480
## imd_score:pcL7  -0.0075165  0.0024737  -3.039 0.002387 **
## imd_score:pcL8  -0.0006921  0.0015228  -0.455 0.649469
## imd_score:pcL9  -0.0141241  0.0021676  -6.516 7.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4985 on 6277 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.9982
## F-statistic: 7.282e+04 on 47 and 6277 DF,  p-value: < 2.2e-16
```

This allows us to get a separate constant term and estimate of the impact of

IMD on the price of a house *for every postcode*[10].

5.6 Spatial dependence

As we have just discussed, SH is about effects of phenomena that are *explicitly linked* to geography and that hence cause spatial variation and clustering of values. This encompasses many of the kinds of spatial effects we may be interested in when we fit linear regressions. However, in other cases, our interest is on the effect of the *spatial configuration* of the observations, and the extent to which that has an effect on the outcome we are considering. For example, we might think that the price of a house not only depends on the level of deprivation where the house is located, but also whether is close to other highly deprived areas. This kind of spatial effect is fundamentally different from SH in that is it not related to inherent characteristics of the geography but relates to the characteristics of the observations in our dataset and, specially, to their spatial arrangement. We call this phenomenon by which the values of observations are related to each other through distance *spatial dependence* (Anselin, 1988).

Spatial Weights

There are several ways to introduce spatial dependence in an econometric framework, with varying degrees of econometric sophistication (see Anselin, 2003, for a good overview). Common to all of them however is the way space is formally encapsulated: through *spatial weights matrices (W)*[11]. These are NxN matrices with zero diagonals and every w_{ij} cell with a value that represents the degree of spatial connectivity/interaction between observations i and j. If they are not connected at all, $w_{ij} = 0$, otherwise $w_{ij} > 0$ and we call i and j neighbors. The exact value in the latter case depends on the criterium we use to define neighborhood relations. These matrices also tend to be row-standardized so the sum of each row equals to one.

A related concept to spatial weight matrices is that of *spatial lag*. This is an operator that multiplies a given variable y by a spatial weight matrix:

$$y_{lag} = Wy$$

If W is row-standardized, y_{lag} is effectively the average value of y in the neighborhood of each observation. The individual notation may help clarify this:

[10] **PRO EXERCISE** *How does the effect of IMD vary over space?* You can answer this by looking at the coefficients of `imd_score` over postcodes, but it would be much clearer if you could create a choropleth of the house locations where each dot is colored based on the value of the `imd_score` estimated for that postcode.

[11] If you need to refresh your knowledge on spatial weight matrices, check Lecture 5 of **?**

$$y_{lag-i} = \sum_j w_{ij} y_j$$

where y_{lag-i} is the spatial lag of variable y at location i, and j sums over the entire dataset. If W is row-standardized, y_{lag-i} becomes an average of y weighted by the spatial criterium defined in W.

Given that spatial weights matrices are not the focus of this tutorial, we will stick to a very simple case. Since we are dealing with points, we will use K-nn weights, which take the k nearest neighbors of each observation as neighbors and assign a value of one, assigning everyone else a zero. We will use $k = 150$ to get a good degree of variation and sensible results. If your computer is struggles to compute the following lines of code, you can replace 50 by a lowed number. Technically speaking is the same thing, but the probability that you will pick up only houses in the same LSOA (and hence with exactly the same IMD score) will be higher.

```
# Because some rows are different units on the same house, slightly
# jitter the locations to break ties
xy.jit <- jitter(db@coords)
# Create knn list of each house
hnn <- knearneigh(xy.jit, k=50)
# Create nb object
hnb <- knn2nb(hnn)
# Create spatial weights matrix (note it row-standardizes by default)
hknn <- nb2listw(hnb)
```

We can inspect the weights created by simply typing the name of the object:

```
hknn
```

```
## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 6324
## Number of nonzero links: 316200
## Percentage nonzero weights: 0.7906388
## Average number of links: 50
## Non-symmetric neighbours list
##
## Weights style: W
## Weights constants summary:
##      n       nn   S0       S1       S2
## W 6324 39992976 6324 230.4304 25816.01
```

Exogenous spatial effects

Let us come back to the house price example we have been working with. So far, we have hypothesized that the price of a house sold in Liverpool can be explained

using information about whether it is newly built, the level of deprivation of the area where it is located, and its postcode. However, it is also reasonable to think that prospective house owners care about the larger area around a house, not only about its immediate surroundings, and would be willing to pay more for a house that was close to nicer areas, everything else being equal. How could we test this idea?

The most straightforward way to introduce spatial dependence in a regression is by considering not only a given explanatory variable, but also its spatial lag. In our example case, in addition to including the level of deprivation in the area of the house, we will include its spatial lag. In other words, we will be saying that it is not only the level of deprivation of the area where a house is located but also that of the surrounding locations that helps explain the final price at which a house is sold. Mathematically, this implies estimating the following model:

$$\log P_i = \alpha + \beta_1 NEW_i + \beta_2 IMD_i + \beta_3 IMD_{lag-i} + \epsilon_i$$

Let us first compute the spatial lag of imd_score:

```
db@data$w_imd_score <- lag.listw(hknn, db@data$imd_score)
```

And then we can include it in our previous specification. Note that we apply the log to the lag, not the reverse:

```
# `:` notation implies interaction variables
m6 <- lm('log(price) ~ new + imd_score + w_imd_score', db)
summary(m6)
```

```
##
## Call:
## lm(formula = "log(price) ~ new + imd_score + w_imd_score", data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2906 -0.3011 -0.0156  0.2822  5.2631
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.2814407  0.0145258 845.492  < 2e-16 ***
## newY         0.2474773  0.0195164  12.680  < 2e-16 ***
## imd_score   -0.0041823  0.0008915  -4.691 2.77e-06 ***
## w_imd_score -0.0148440  0.0009683 -15.330  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.52 on 6320 degrees of freedom
## Multiple R-squared:  0.3349, Adjusted R-squared:  0.3346
```

```
## F-statistic:  1061 on 3 and 6320 DF,  p-value: < 2.2e-16
```

As we can see, the lag is not only significative and negative (as expected), but its effect seems to be even larger that that of the house itself. Taken literally, this would imply that prospective owners value more the area of the surrounding houses than that of the actual house they buy. However, it is important to remember how these variables have been constructed and what they really represent. Because the IMD score is not exactly calculated at the house level, but at the area level, many of the surrounding houses will share that so, to some extent, the IMD of neighboring houses is that of the house itself[12]. This is likely to be affecting the final parameter, and it is a reminder and an illustration that we cannot take model results as universal truth but we need to use them as tools to inform analysis, couple with theory and what we know about the particular question of analysis. Nevertheless, the example does illustrate how to introduce spatial dependence in a regression framework in a fairly straight forward way.

A note on more advanced spatial regression

Introducing a spatial lag of an explanatory variable, as we have just seen, is the most straightforward way of incorporating the notion of spatial dependence in a linear regression framework. It does not require additional changes, it can be estimated with OLS, and the interpretation is rather similar to interpreting non-spatial variables. The field of spatial econometrics however is a much broader one and has produced over the last decades many techniques to deal with spatial effects and spatial dependence in different ways. Although this might be an over simplification, one can say that most of such efforts for the case of a single cross-section are focused on two main variations: the spatial lag and the spatial error model. Both are similar to the case we have seen in that they are based on the introduction of a spatial lag, but they differ in the component of the model they modify and affect.

The spatial lag model introduces a spatial lag of the *dependent* variable. In the example we have covered, this would translate into:

$$\log P_i = \alpha + \rho \log P_{lag-i} + \beta_1 NEW_i + \beta_2 IMD_i + \epsilon_i$$

Although it might not seem very different from the previous equation, this model violates the exogeneity assumption, crucial for OLS to work.

Equally, the spatial error model includes a spatial lag in the *error* term of the equation:

$$\log P_i = \alpha + \beta_{1r} NEW_i + \beta_{2r} IMD_i + u_i$$

[12] **EXERCISE** *How do results change if you modify the number of neighbors included to compute the K-nn spatial weight matrix?* Replace the originak k used and re-run the regressions. Try to interpret the results and the (potential) differences with the original ones.

$$u_i = u_{lag-i} + \epsilon_i$$

Again, although similar, one can show this specification violates the assumptions about the error term in a classical OLS model.

Both the spatial lag and error model violate some of the assumptions on which OLS relies and thus render the technique unusable. Much of the efforts have thus focused on coming up with alternative methodologies that allow unbiased, robust, and efficient estimation of such models. A survey of those is beyond the scope of this note, but the interested reader is referred to Anselin (1988), Anselin (2003), and Anselin and Rey (2014) for further reference.

5.7 Predicting house prices

So far, we have seen how exploit the output of a regression model to evaluate the role different variables play in explaining another one of interest. However, once fit, a model can also be used to obtain predictions of the dependent variable given a new set of values for the explanatory variables. We will finish this session by dipping our toes in predicting with linear models.

The core idea is that once you have estimates for the way in which the explanatory variables can be combined to explain the dependent one, you can plug new values on the explanatory side of the model and combine them following the model estimates to obtain predictions. In the example we have worked with, you can imagine this application would be useful to obtain valuations of a house, given we know the IMD of the area where the house is located and whether it is a newly built house or not.

Conceptually, predicting in linear regression models involves using the estimates of the parameters to obtain a value for the dependent variable:

$$\bar{\log P_i} = \bar{\alpha} + \bar{\beta_{1r}} NEW_i^* + \bar{\beta_{2r}} IMD_i^*$$

where $\bar{\log P_i}$ is our predicted value, and we include the $\bar{}$ sign to note that it is our estimate obtained from fitting the model. We use the * sign to note that those can be new values for the explanatory variables, not necessarily those used to fit the model.

Technically speaking, prediction in linear models is fairly streamlined in R. Suppose we are given data for a new house which is to be put in the market. We know it is been newly built on an area with an IMD score of 75, but surrounded by areas that, on average, have a score of 50. Let us record the data first:

```
new.house <- data.frame(new='Y', imd_score=75, w_imd_score=50)
```

To obtain the prediction for its price, we can use the `predict` method:

```
new.price <- predict(m6, new.house)
new.price
```

```
##        1
## 11.47304
```

Now remember we were using the log of the price as dependent variable. If we want to recover the actual price of the house, we need to take its exponent:

```
exp(new.price)
```

```
##        1
## 96090.29
```

According to our model, the house would be worth GBP96,060.29[13].

5.8 References

[13] **EXERCISE** *How would the price change if the surrounding houses did not have an average of 50 but of 80?* Obtain a new prediction and compare it with the original one.

Chapter 6

Multilevel Modelling - Part 1

This chapter[1] provides an introduction to multi-level data structures and multi-level modelling.

The content of this chapter is based on:

- Snijders and Bosker (2012)
- for Multilevel Modelling (nd)

6.1 Dependencies

This tutorial uses the libraries below. Ensure they are installed on your machine[2] before loading them executing the following code chunk:

```
# Data manipulation, transformation and visualisation
library(tidyverse)
# Nice tables
library(kableExtra)
# Simple features (a standardised way to encode vector data ie. points, lines, polygons)
library(sf)
# Spatial objects conversion
library(sp)
```

[1]This note is part of Spatial Analysis Notes Multilevel Modelling – Random Intercept Multilevel Model by Francisco Rowe is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

[2]You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

```
# Thematic maps
library(tmap)
# Colour palettes
library(RColorBrewer)
# More colour palettes
library(viridis) # nice colour schemes

# Set working directory
setwd(".")
```

> Flow of the argument: explanation of key components of linear regressions which matter to introduce multilevel models

Relevant assumptions of the linear regression model to recall * independence of residuals * constant variability * mean of residuals = 0 Not sure if all of this is relevant - not to be written down.

Chapter 7

Multilevel Data Structures

Geographically hierarchical data structure: * Stric hierarchy eg. People nested within households, and households nested within neighbourhoods; states nested within countries * Cross-classified hierarchy eg. People within neighbouhoods and workplaces

Temporally hierarchical data structures eg. * Repeated measures design or longitudinal data: Individuals followed over time at a particular location

Why should we care?

Conceptually, individual observations within a particular grouping can be expected to be more alike than observations from a random set.

- Spatial hetorogeneity
- Spatial dependency
- Contextual dependencies: micro and macro level relationships

Technically, we have incorrect estimates of precision, standard errors, confidence intervals and tests, increased risk of finding relationships and differences where none exists.

7.1 Data

7.2 Long vs wide format

Chapter 8

Single-level Regression Models

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

Example: estimate regression model

8.1 Limitations

- Captures averages
- Does not account for spatial and temporal dependencies
- Only captures on macro *vs* micro processes
  - A macro approach: enables understanding of higher level unit analysis, not individuals within these units > Ecological fallacy
  - A micro approach: enables understanding of individuals without acknoledgement of dependencies between individuals within the same group or spatial unit > Underestimation of standard errors
- May lead to drawn erroneous conclusions frome the identification of misleading relationships

8.2 Adding Group Fixed Effects

Introduce this as adding dummy variables

Example: estimate regression model

- The number of groups could be very large, eg households?
- No single parameter assesses between group differences
- Cannot include group-level predictors

Chapter 9

Multilevel Modelling

9.1 The Intuition

Get a more general and simple explanation: * 2 levels * capture the hierarchical structure of the data * enable capturing for spatial heterogenity, spatial dependence and macro & micro factors in a same model

Micro level or level 1:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}$$

Macro level or level 2:

$$\beta_{0j} = \beta_0 + u_{0j}$$

Embedding level 2 into level 1:

$$y_{ij} = [\beta_0 + u_{0j}] + \beta_{1j}x_{1ij} + e_{ij}$$

where: u_{0j} is the level-2 random component and represents variability between groups / areas; e_{ij} is the level-1 random component and represents variability within groups / areas ie. between level-1 units (eg. houses)

By dividing the variability of the model into different levels, we can estimate the proportion of the variance at each level and assess their extent. This can be achieved by using the variance partitioning coefficient.

Estimating the residuals is more complex! For linear regressions $e_i = y_i - \hat{y}_i$ For MLMs, a two step procedure is required: 1. Estimate the variance 2. Estimate individual random effects

Introduce mathematical notation of multilevel models for a two-level model Explain technically what it does

> Too technical * Partition residual variance into between- and within-group (level 2 and level 1) components. * Allows for un-observables

at each level, corrects standard errors: measures of uncertainty * Micro and macro factors * richer set of research questions

Various labels: * Random-effects Models * Hierarchical Models * Variance-components Models * Random-coefficient Models * Linear / Generalised Linear Mixed Models

9.2 Practice

In R, there are several Packages implementing multilevel models. This tutorial will show how to use the `lme4` Package (short for linear mixed effect models) in R to fit MLM and use the companion `merTools` Package to neatly analyse the model output from `lme4`. This tutorial will cover getting set up and running a few random interpet multilevel models and interpret model estiamtion results.

Recall set the working directory before reading the data.

```
# read data
oa_shp <- st_read("data/mlm/OA.shp")
```

```
## Reading layer `OA' from data source `/home/jovyan/work/data/mlm/OA.shp' using drive
## Simple feature collection with 1584 features and 19 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 332390.2 ymin: 379748.5 xmax: 345636 ymax: 397980.1
## epsg (SRID):    NA
## proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-1
```

Visualise the hierarchical structure of the data

```
# attach data frame
attach(oa_shp)

# sort data by oa
oa_shp <- oa_shp[order(OA_CD),]
head(oa_shp)
```

```
## Simple feature collection with 6 features and 19 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 335056 ymin: 389163 xmax: 336155 ymax: 389642
## epsg (SRID):    NA
## proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-1
##         OA_CD   LSOA_CD   MSOA_CD    LAD_CD pop H_Vbad H_bad H_fair H_good
## 558 E00032987 E01006515 E02001383 E08000012 198      3    13     23     70
## 441 E00032988 E01006514 E02001383 E08000012 348      5    36     65    116
## 736 E00032989 E01033768 E02001383 E08000012 333      7    22     49    115
```

```
## 41  E00032990 E01033768 E02001383 E08000012 330      8    17     65    113
## 419 E00032991 E01033768 E02001383 E08000012 320      7    28     46    101
## 907 E00032992 E01033768 E02001383 E08000012 240      7    16     42     69
##     H_Vgood  age_men age_med     age_60    S_Rent    Ethnic illness
## 558      89 32.19192    28.0 0.11616162 0.5862069 0.5050505     198
## 441     126 40.30460    39.5 0.16954023 0.6500000 0.2327586     348
## 736     140 35.35435    34.0 0.09009009 0.7586207 0.3423423     333
## 41      127 36.87879    32.0 0.15151515 0.5799087 0.3848485     330
## 419     138 34.73125    33.0 0.04687500 0.6073059 0.3406250     320
## 907     106 34.40000    32.0 0.05833333 0.8121212 0.4833333     240
##         unemp males   dis_ind                       geometry
## 558 0.1130435    92 0.2378571 MULTIPOLYGON (((335187 3894...
## 441 0.1458333   203 0.2378571 MULTIPOLYGON (((335834 3895...
## 736 0.1049724   214 0.2378571 MULTIPOLYGON (((335975.2 38...
## 41  0.1329787   197 0.2378571 MULTIPOLYGON (((336030.8 38...
## 419 0.1813725   194 0.2378571 MULTIPOLYGON (((335804.9 38...
## 907 0.2519685   139 0.2378571 MULTIPOLYGON (((335804.9 38...
```

What is the hierarchy of the data?

9.3 Intercept Only Model

Chapter 10

Multilevel Models (Pt. II)

FR to fill in

Chapter 11

GWR

FR

Chapter 12

Space-Time Analysis

FR

Bibliography

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, 26(2):153–166.

Anselin, L. (2007). Spatial regression analysis in r–a workbook.

Anselin, L. and Rey, S. J. (2014). *Modern spatial econometrics in practice: a guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Arribas-Bel, D. (2014). Spatial data, analysis, and regression-a mini course. *REGION*, 1(1):R1.

Arribas-Bel, D. (2019). Geographic data science'19.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Springer New York.

Brunsdon, C. and Comber, L. (2015). *An introduction to R for spatial analysis & mapping*. Sage.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

for Multilevel Modelling, C. (n.d.). Introduction to multilevel modelling.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gibbons, S., Overman, H. G., and Patacchini, E. (2014). Spatial methods.

Grolemund, G. and Wickham, H. (2019). *R for Data Science*. O'Reilly, US.

Lovelace, R. and Cheshire, J. (2014). Introduction to visualising spatial data in R. *National Centre for Research Methods Working Papers*, 14(03).

Lovelace, R., Nowosad, J., and Muenchow, J. (2020). *An introduction to statistical learning*, volume 112. CRC Press, R Series.

Singleton, A. (2017). Geographic data science for urban analytics. Online course.

Snijders, T. and Bosker, R. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage Publishing.

Williamson, P. (2018). Survey analysis.

Xie, Y., Allaire, J., and Grolemund, G. (2019). *R Markdown: The Definitive Guide*. CRC Press, Taylor & Francis, Chapman & Hall Book.