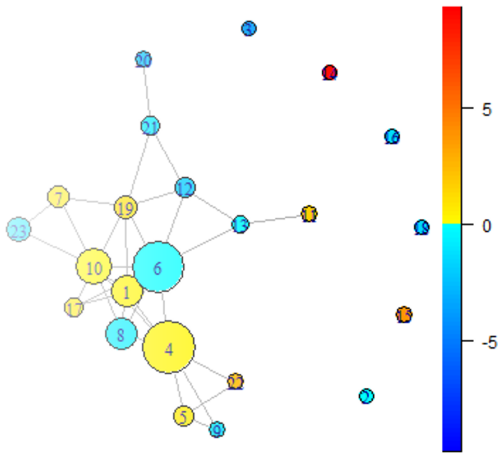


Regional Analysis with Topological Data Analysis Ball Mapper

Session 3: Introduction to Ball
Mapper in R

Dr Simon Rudkin
University of Manchester

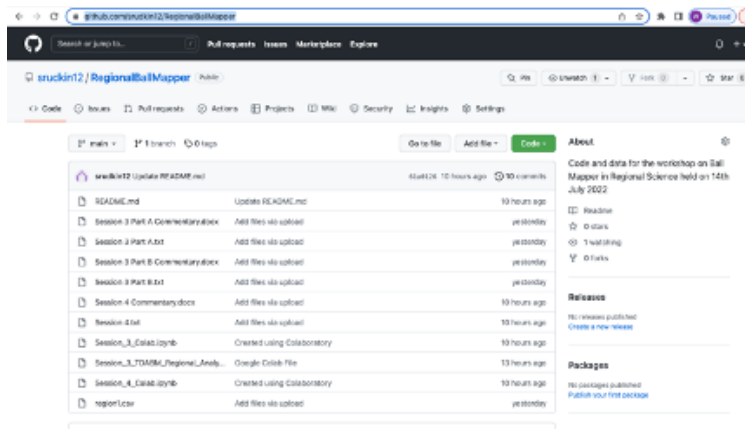


In this Session...

- Introduction to the use of R
- Part A: Summarising with statistics
- Part B: Ball Mapper plots
- Review

This session serves as an introduction to the R package BallMapper (Dlotko, 2019) which enables the use of Topological Data Analysis Ball Mapper (TDABM) as based upon the original working paper of Dłotko (2019).



GitHub: <https://github.com/srudkin12/RegionalBallMapper>



All of the material for this workshop is available on the GitHub site Link in Email

Regional Analysis with TDA BM: Session 3
Dr Simon Rudkin

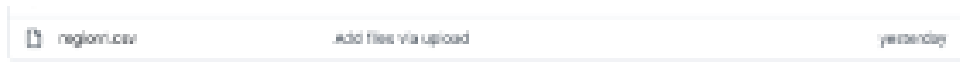
Files on GitHub

File	Commit Message	Commit Date
 Session 3 Part A: Commentary.docx	Add files via upload	yesterday
 Session 3 Part A.txt	Add files via upload	yesterday

Each half of the session has:

- Commentary file as a Word document
- Code file as a .txt file
- Google Colab .ipynb file - These allow you to run the code without installing R

Files on GitHub 2



The dataset for this session is contained in the file `region1.txt`

- Download the file and place it into a new folder
- Ensure that the folder is easy to navigate to
- The folder will be your working directory

Useful R Terminology

Working directory	Folder in which R finds data and saves any output
Command line	Prefaced with a ">" symbol. For entering commands into R
Function	For converting stated inputs into outputs. BallMapper() is an example converting axis variables, outcome variable and the ball radius into a BallMapper object
Object	For storing content in R. Defined by code with a <-
Package	Set of codes produced by a contributor for performing particular tasks (e.g. the BallMapper package) - Must be installed once* and then read into R using the library() function
data.frame	Format used by R to store data tables. Required as the format for data provided to the BallMapper function in Part B

Variables used in this Session

Group	Variable	Interpretation (All are percentages)
Geo	geog	Name of the Local Authority District
Depn	Deprivation0	Households with no deprivation as assessed against Income, health, Overcrowding and Education
	Deprivation1	Households defined as deprived on one of the four measures
	Deprivation2Plus	Households defined as deprived on two or more of the four measures
Health	HealthVeryGood	Respondents who self-identify as having very good health
	HealthGood	Respondents who self-identify as having good health
	HealthLow	Respondents who self-identify as having fair, bad or very bad health

Variables used in this Session 2

Group	Variable	Interpretation (All are percentages)
Employment	Armed	Respondents employed in the armed forces
	Agriculture	Respondents working in the agriculture sector
	Manufacturing	Respondents working in the manufacturing sector
	Accommodation	Respondents working in the accommodation and travel sector
Household	Married	Households where the owners are married
	Cohabit	Households where the owners cohabit
	Single	Households with one adult resident who is single
	Other	Households with one adult resident in a relationship, widowed or divorced

Variables used in this Session 3

Group	Variable	Interpretation (All are percentages)
Qualifications	QualNone	Highest level of qualification in household is below secondary school
	QualLevel1	1-4 GCSEs at grade A-C
	QualLevel2	5+ GCSEs at grade A-C
	QualApprentice	Apprenticeships
	QualLevel3	Two or more A-Levels
	QualLevel4	University degree or higher – includes professional qualifications
	QualOther	Includes vocational qualifications

Variables used in this Session 4

Group	Variable	Interpretation (All are percentages)
Ownership	OwnedOutright	Household is owned outright
	OwnedMortgage	Household is owned with support from a mortgage
	SocialRental	Household is rented from a social housing agency (e.g council)
	PrivateRental	Household is rented from a private individual or company

- The full table can be found the the Session 3 Part A commentary

Outline of the Session

Time	Activity	Recorded
13:40 - 14:10	Part A	No
14:10 - 14:15	Review of Part A	Yes
14:15 - 14:45	Part B	No
14:45 - 15:00	Review of Part B	Yes

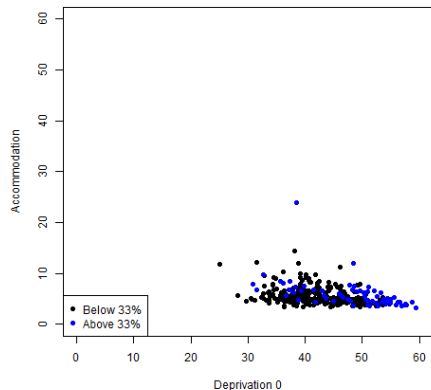
- A full commentary is available on the GitHub site
- Results for the questions in the Review slides

Review of Part A

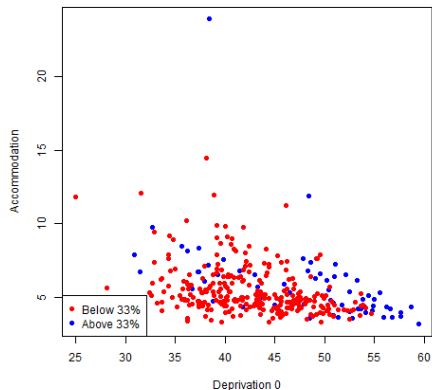
```
> head(dty)
      geog  geogcode QualLevel1 Deprivation0 Accommodation  Married HealthVeryGood OwnedMortgage QL1
1      Adur E07000333  31.98936  40.46508  4.040060  47.30034  42.67165  35.95356  0
2 Allerdale E07000026  22.77838  42.97886  8.481689  51.91272  44.47844  29.59269  0
3 Arder Valley E07000632  23.12130  43.42916  4.329668  50.16805  44.00821  35.87345  0
4      Arun E07000224  22.77353  40.41765  6.433581  49.77865  41.98424  30.99122  0
5      Ashfield E07000170  15.09338  37.31715  4.367862  47.80116  42.05378  35.94864  0
6      Ashford E07000105  24.62986  46.76586  5.135022  50.75990  47.58384  36.58945  0
```

- Check the data with the `head()` function
- Useful to ensure that the data has been read in as expected

Review of Part A 2 - Deprivation 0 and Accommodation

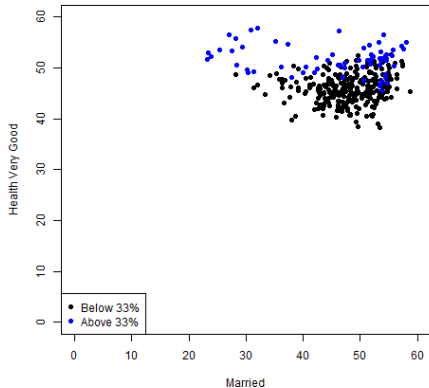


(a) Fixed axis size

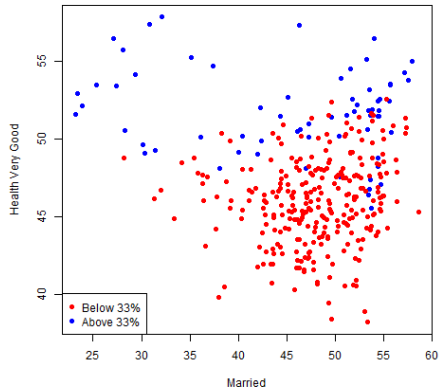


(b) Determined by two variables

Review of Part A 3 - Married and Health Very Good



(a) Fixed axis size

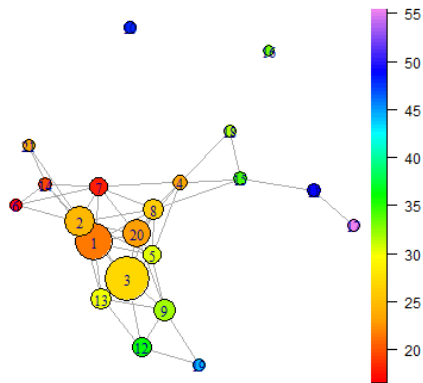


(b) Determined by two variables

Outline of the Session

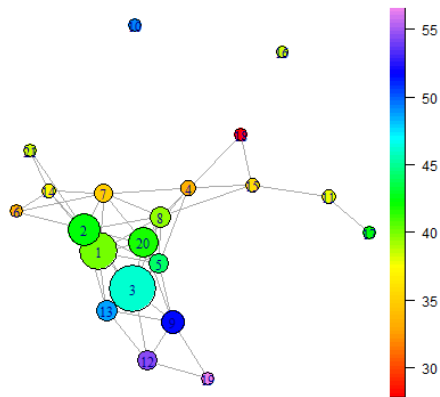
Time	Activity	Recorded
13:40 - 14:10	Part A	No
14:10 - 14:15	Review of Part A	Yes
14:15 - 14:45	Part B	No
14:45 - 15:00	Review of Part B	Yes

Review of Part B



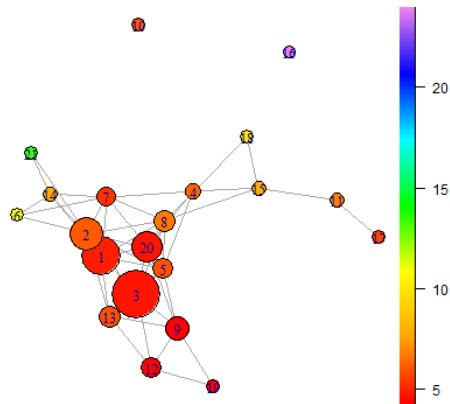
- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Coloured by % of households where highest qualified resident has a university degree or higher
- Radius is 0.30
- Highest blocks in arms on the right
- Two outliers with lower levels of qualifications

Review of Part B 2



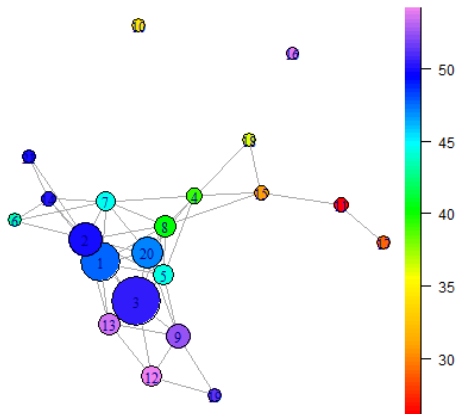
- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Deprivation0 highest bottom centre

Review of Part B 2



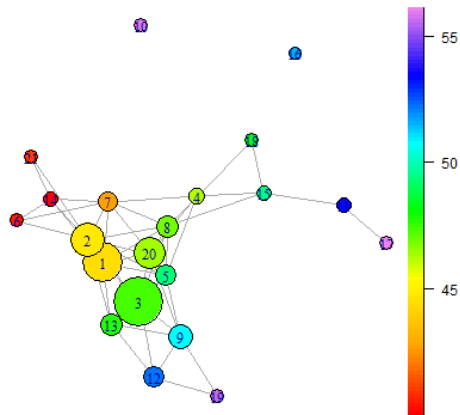
- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Deprivation0 highest bottom centre
- Highest Accommodation creates outlier

Review of Part B 2



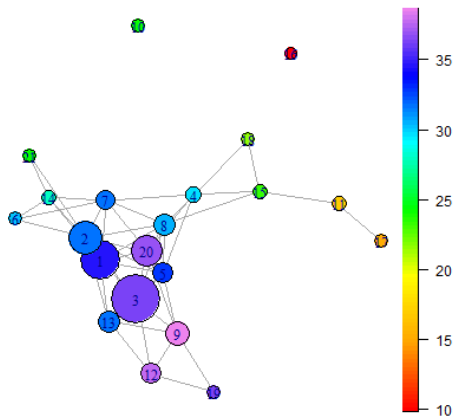
- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Deprivation0 highest bottom centre
- Highest Accomodation creates outlier
- Highest Married along the left

Review of Part B 2



- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Deprivation0 highest bottom centre
- Highest Accommodation creates outlier
- Highest Married along the left
- HealthVeryGood in the arms and outliers

Review of Part B 2



- Axes are Deprivation0, Accommodation, Married, HealthVeryGood, OwnedMortgage
- Deprivation0 highest bottom centre
- Highest Accommodation creates outlier
- Highest Married along the left
- HealthVeryGood in the arms and outliers
- OwnedMortgage similar to married

Review of Part B 3

[2]

A data.frame: 4 x 11

	pa	psg	psgcode	QualLevel1	Deprivation0	Accommodation	Married	RealVVeryGood	OwnedMortgage	QLE	ball
	<cat>	<cat>	<cat>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
129	67	City of London	E09000020	68.28117	-14.98430	-1.308112	28.11221	55.75883	17.37742	1	17
205	127	Hammerstein and Fulham	E09000013	48.56089	-11.43785	6.839125	27.89230	55.46025	18.14121	1	17
327	152	Hamington and Chelsea	E09000023	52.67936	-13.54523	5.030873	32.85896	57.84369	12.82724	1	17
737	331	Westminster	E09000033	52.28006	-18.24195	8.536105	29.35315	54.95233	12.88312	1	17

- Four Local Authority Districts in ball 17 with Deprivation0 and Married around 30%

Review of Part B 4

```
[22] ball19<-subset(data, $ball==19)
ball19
```

A data block 5 x 11										
pt	geog	geocode	gaisaen00	deprivation00	accommodation	married	mainlverywood	ownedmortgage	ql4	ball
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
215	102	Elmbridge	E87800007	43.85234	57.83010	3.704889	54.66508	55.45081	35.81373	1 19
265	102	Guildford	E87800008	35.52973	54.32932	4.801387	50.61437	53.51595	35.09080	1 19
587	229	Richmond upon Thames	E80800027	52.83761	57.82510	2.804718	46.33812	57.33836	33.81224	1 19
889	272	St Albans	E87800010	46.31025	57.83806	2.732438	53.31862	46.11387	36.58878	1 19
747	338	Windsor and Maidenhead	E80800010	38.39817	54.43482	6.133128	51.86304	54.51162	35.41277	1 19

- Five Local Authority Districts in ball 19 with OwnedMortgage and Married around 50%
- Deprivation0 is higher than ball 17 as well

Review of Part B 5

A Data Based BM 21											
pa	geog	geocode	qualLevel04	Deprivation00	Accommodation	married	residVeryPoor	OwnedMortgage	gls	ball	
code	area	code	code	code	code	code	code	code	code	code	
538	304	Southern	887000108	23.89016	38.84285	11.98329	47.81858	41.86448	28.88986	0	21
738	338	West Somerset	887000191	24.84587	38.07287	16.17737	51.80382	40.80862	37.74388	0	21

- Two Local Authority Districts in ball 21 with OwnedMortgage at just 20%
- Large values in accommodation (above 10%)
- Lower values on Deprivation0 also

Summary of Session 3

- Summary statistics are useful but we need to visualise data
- Plots can be coloured but hard to understand the joint distribution
- Ball Mapper requires only three inputs - axes, outcome and radius
- We may identify local authority districts from within the data
- This session gave a first look at what can be done, more in Session 4...

Dłotko, P. (2019). Ball mapper: a shape summary for topological data analysis. *arXiv preprint arXiv:1901.07410*.

Dłotko, P. (2019). *BallMapper: Create a Ball Mapper graph of the input data*. R package version 0.1.0.