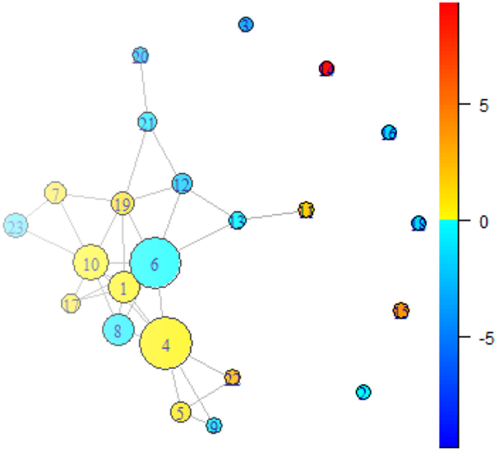
Regional Analysis with Topological Data Analysis Ball Mapper

Session 4: Further use of Ball Mapper in R

Dr Simon Rudkin
University of Manchester



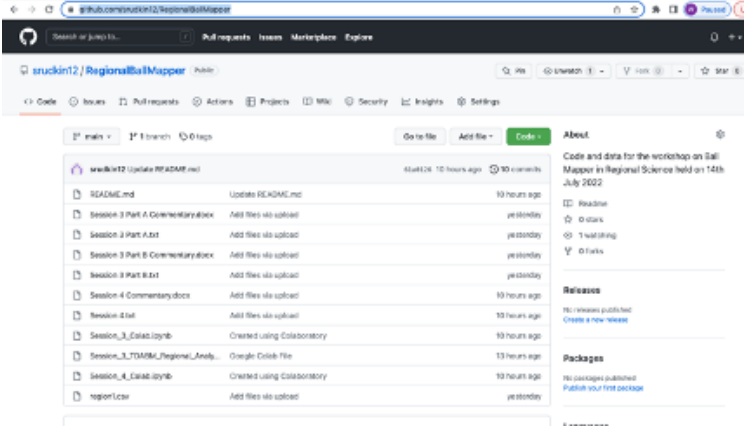
MANCHESTER 1824

In this Session...

- Regression models and Ball Mapper
- Ball Mapper on your data

This session gives further coverage to the R package BallMapper (Dlotko, 2019) which enables the use of Toplogical 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

Files on GitHub

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 qualificiations

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
15:15 - 15:45	Regression / Work on Own Data	No
15:45 - 15:55	Review of Session 4	Yes
15:55 - 16:30	Discussions	No

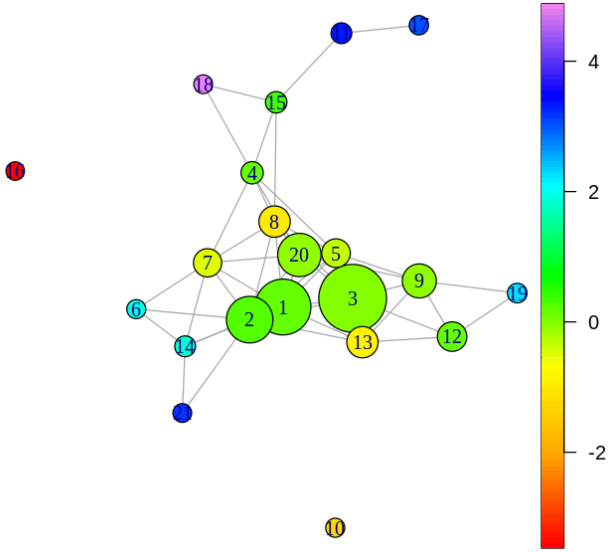
- A full commentary is available on the GitHub site
- Participants are strongly encouraged to share their results - please indicate willingness

Review of Session 4

```
Call:
lm(formula = QualLevel4 ~ Deprivation0 + Accommodation + Married +
    HealthVeryGood + OwnedMortgage, data = dty)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8954 -1.7347 -0.0764  1.7694 15.3443

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -13.67329    3.81260  -3.586 0.000384 ***
Deprivation0     0.69207    0.05928  11.674  < 2e-16 ***
Accommodation   -0.38520    0.11170  -3.448 0.000634 ***
Married         -0.22090    0.04409  -5.010 8.75e-07 ***
HealthVeryGood   1.00547    0.08335  12.063  < 2e-16 ***
OwnedMortgage   -0.72279    0.04655 -15.528  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.924 on 342 degrees of freedom
Multiple R-squared:  0.8621,    Adjusted R-squared:  0.8601
F-statistic: 427.7 on 5 and 342 DF,  p-value: < 2.2e-16
```

- Highest VIF is 5.73
- OLS output on the left hand side
- All variables are highly significant
- Coefficient on HealthVeryGood is close to 1
- R squared is high at 0.87

Review of Session 4



- Colour the BM plot by the residuals
- Values are close to 0 in the centre
- Positive residuals in all arms
- Outlier with very negative residual...

MANCHESTER 1824

Ball 16

A data.frame: 1 × 13

	pt	geog	geogcode	QualLevel4	DeprivationR	Accommodation	Married	HealthVeryGood	OwnedMortgage	QLR	fit1	res1	ball
	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
325	153	Isles of Scilly	E06000053	33.01023	38.42255	23.0749	53.03942	51.52965	16.31311	1	36.43936	-3.429131	16

- The outlier in this case is the Isles of Scilly
- The model thinks the Qualification Level 4 should be much higher
- Physical geography here is the important factor...

Summary of Session 4

- Ball Mapper also speaks to the modelling process
- Seeing where models fit well can inform model development
- Alternatively seeing the residuals allows us to understand why
- Long research agenda building on these observations...
- In all cases we can benefit from the visualisation enabled by Ball Mapper

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

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