Regional Analysis with Topological Data Analysis Ball Mapper

Session 4: Further Use of Ball Mapper

Session 4 is focused on supporting individuals in their application of Ball Mapper to their own datasets. There is also a short additional section on the evaluation of regression models using Ball Mapper graphs. This document supplements the code file that is available on the GitHub repository as Session 3 Part A.txt and you are strongly encouraged not to try to copy and paste from this document. All code in this file appears in the courier new font

**Preliminaries**

If you have not previously used R then please make sure that you have a version of the software installed on your computer. There is a second alternative to use Google Colab for this session that does not require an installation of R. A separate document gives guidance on using Google Colab. The remainder of the instructions are the same however you are running the code.

Packages

R uses packages of user developed code to run many of its’ useful functions therefore you will need to install the packages for the session. This only needs to be done once, so if you already have the packages installed you may skip this step. Note that relative to Session 3 we are also now adding the car package.

install.packages("dplyr")

install.packages("BallMapper")

install.packages("car")

Working Directory (Mirrors Session 3)

R requires you to set a working directory in which the data is located. This will also be the folder where all output is saved. In the example code the working directory is set to setwd("D://oraf/") but you should set this with the right path for your computer. If you wish you can set the working directory through the menus. On Windows you would select File and then Change dir… , whilst on a Mac you would select Misc and then Change Working Directory.

Before starting you must make sure that the **region1.csv** file is in your working directory.

*For those using Google Colab the instructions are slightly different because Google creates a temporary working directory for you. There you will need to upload the data.*

**Setting Up the R Environment**

For this session we will need two packages and so will use the following code to bring them into the R environment

library(dplyr)

library(BallMapper)

library(car)

Next we should read in the data using the read.table function. The options are the filename, the fact that the file uses a comma to separate columns, and the fact that the first row is the variable names.

dtx<-read.table("region1.csv",sep=",",header=TRUE)

|  |  |
| --- | --- |
| Variable | Interpretation (All are percentages) |
| geog | Name of the Local Authority District |
| Geogcode | Office for National Statistics code for the Local Authority District |
| Armed | Respondents employed in the armed forces |
| 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 |
| 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 |
| Agriculture | Respondents working in the agriculture sector |
| Manufacturing | Respondents working in the manufacturing sector |
| Accommodation | Respondents working in the accommodation and travel sector |
| 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 |
| 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 |
| 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 |

Table 1: Variables used in this session

Table 1 is a useful reference for all of the variables used in this session. However, there are obviously large correlations between the variables. BallMapper can work with highly correlated variables, but to avoid any problems of correlation in the statistical analysis of the data we will restrict focus to a subset of variables. Selection is made using the following code (note each block is a single line):

dty<-cbind(dtx[,1:2],dtx$QualLevel4,dtx$Deprivation0,dtx$Accommodation,dtx$Married,dtx$HealthVeryGood,dtx$OwnedMortgage)

names(dty)<-c("geog","geogcode","QualLevel4","Deprivation0","Accommodation","Married","HealthVeryGood","OwnedMortgage")

These variables may be edited if you wish to try different combinations of the data.

Finally, as in Session 3, we create a dummy variable which takes the value 1 if the percentage of households where the highest qualified resident has a university degree or higher. The intuition of this code is that the value 1 is assigned wherever the condition is true.

dty$QL4<-as.numeric(dty$QualLevel4>33)

Our data is now ready for analysis. As a final verification the head of dty should appear as in Figure 2

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1: Head of the reduced dataset dty

**Regression Models**

R uses the function lm() to produce ordinary least squares (OLS) regression models. There are many other modelling options open but we are going to use OLS as an illustration. The basic model fitted is going to be:

Where is a constant term and is an iid error term with mean 0 and constant variance . In the fitted model we have:

With being the estimate of and representing the estimate of . In the fitted model the error term is 0.

To estimate the OLS model in R we must specify the equation and tell R which data to use:

lm1<-lm(QualLevel4~Deprivation0+Accommodation+Married+HealthVeryGood+OwnedMortgage,data=dty)

Before we may actually use this model in our analysis we must check that there are no issues of multicollinearity. There are two ways to do this. One is to look at the correlation matrix (Figure 2) and the second is to calculate the variance inflation factor (VIF) for each variable (Figure 3). There are potential concerns if there is a correlation between two independent variables of more than 0.7 in absolute value, or a VIF of 10 or more.

Text, application

Description automatically generated

Figure 2: Correlation matrix for dty

Text

Description automatically generated with medium confidence

Figure 3: Variance inflation factors

Consequently we do not need to be overly concerned about multicollinearity in the data.

To produce a summary of the regression model R offers the function summary() and so we will apply that here. We create an object and then view it. By doing this we are also able to save the coefficient matrix as a .csv file.

sm1<-summary(lm1)

write.table(sm1$coef,"Model1Coefficients.csv",sep=",",row.names=FALSE)

sm1

There are better ways to get summaries of models outputted into a format useful for incorporation into papers and documents. However, as the focus of this session is on Ball Mapper, we leave that to your own research.

Graphical user interface, text, application, email

Description automatically generated

The resulting summary of the regression is shown in Figure 4 on the left.

We see that all of the variables are highly significant, with Health Very Good having a coefficient of close to 1. Some of the other coefficients warrant more discussion.

Figure 4: OLS model summary

Q1. What inference can we gather from the OLS model coefficients?

Q2. What other factors may help explain the variation in QualLevel4 which is not explained by the suggested model?

To complete the discussion of regression we will create the fitted values for each of the local authority districts. The fitted value is added as a new column in dty.

dty$fit1<-fitted(lm1)

We then find the residual by subtracting the fitted value from the true value. Again the residuals are stored as a new column in dty

dty$res1<-dty$QualLevel4-dty$fit1

**Linking Regression to Ball Mapper**

This workshop is focused on Ball Mapper and so we will now consider how to use Ball Mapper with regression output to gain inference on the quality of our modelling.

To use Ball Mapper it is necessary to have the axis variables and outcomes as individual data.frame objects. Using code from Session 3, but now adding an outcome for the fitted value and residual we run the following code block:

y1<-as.data.frame(dty$QualLevel4)

y2<-as.data.frame(dty$QL4)

y3<-as.data.frame(dty$fit1)

y4<-as.data.frame(dty$res1)

x1<-as.data.frame(dty[,4:8])

x2<-normalize\_to\_min\_0\_max\_1(x1)

Notice that we are again normalising the characteristics because of the difference between the range of values in Accommodation and many of the other variables.

Let us now produce three BallMapper plots. One of the true value (t), one of the fitted values (f) and one of the residuals (r). Here we continue with the 0.3 radius identified in Session 3.

bm130t<-BallMapper(x2,y1,0.3)

bm130f<-BallMapper(x2,y3,0.3)

bm130r<-BallMapper(x2,y4,0.3)

We may plot these Ball Mapper graphs using the ColorIgraphPlot() function as before. In the block of code below you will see first three lines that produce the graphs in the R window. By using the arrows you can go back to see any one of the three.

ColorIgraphPlot(bm130t,seed\_for\_plotting=123)

ColorIgraphPlot(bm130f,seed\_for\_plotting=123)

ColorIgraphPlot(bm130r,seed\_for\_plotting=123)

As previously these Ball Mapper graphs can be output to .png files for inclusion in projects or papers. The resulting graphs are displayed in Figure 5

png("bm130t.png")

ColorIgraphPlot(bm130t,seed\_for\_plotting=123)

dev.off()

png("bm130f.png")

ColorIgraphPlot(bm130f,seed\_for\_plotting=123)

dev.off()

png("bm130r.png")

ColorIgraphPlot(bm130r,seed\_for\_plotting=123)

dev.off()

In Figure 5 we see that the lowest residuals are in the main mass of the data. OLS is a method which fits best at the average and therefore we should not be surprised that the fit is best in the bigger balls and around the centre of the plot. An interesting observation from Figure 5 panel (c) is that the only large negative residual is in the outlier ball 16.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

1. True values (b) Residuals

Figure 5: True values and residuals from the fitting of the OLS model to the data in dty. We do not include the fitted values here for brevity

Our next task is to identify the points in ball 16 to find out which Local Authority District(s) are contained within ball 16. To do so we follow code from Session 3 Part B to first add a column of point numbers to the dty dataframe

dty$pt<-seq(1:nrow(dty))

Secondly, we define the function points\_to\_balls and apply it onto the BallMapper object

points\_to\_balls<-function(l){

a001<-length(l$landmarks)

a1<-matrix(0,nrow=a001,ncol=2)

a1<-as.data.frame(a1)

names(a1)<-c("pt","ball")

for(i in 1:a001){

a<-as.data.frame(l$points\_covered\_by\_landmarks[i])

names(a)<-"pt"

a$ball<-i

a1<-rbind.data.frame(a1,a)

}

a1<-a1[2:nrow(a1),]

return(a1)

}

bmp1<-points\_to\_balls(bm130r)

bmp1<-as.data.frame(bmp1)

names(bmp1)<-c("pt","ball")

Because some points are in more than one ball we see that the length of the table bmp1 (802 rows) is much longer than just the number of points in the data set (348).

Now let us merge bmp with dty using the point number

dtz<-merge(dty,bmp1,by="pt")

Subsetting to look at ball 16 is then straightforwards.

dtz16<-subset(dtz,dtz$ball==16)

Graphical user interface, application

Description automatically generated

Figure 6: Details of the membership of Ball 16.

Q3. Does the overprediction of the model for ball 16 make sense?

Q4. What else can we say about ball 16?

Q5. What further inference can be taken from the Ball Mapper plots of the residuals?

**Summary**

This final practical session presents an introduction to the use of Topological Data Analysis Ball Mapper in statistical modelling. We considered a simple OLS model for the data that had been introduced in Session 3. There are many extensions which can be explored and it is an ongoing research agenda to formalise those extensions. For example, even the examination of residuals has great potential in informing the analyst as to where in the space models are fitting well and where they are fitting badly. Again the ability of Ball Mapper to give a visualisation of multi-dimensional data is an important asset.