Regional Analysis with Topological Data Analysis Ball Mapper

Session 3 Part B: Producing a Ball Mapper Plot

Session 3 gives you the chance to use the Ball Mapper (BM) algorithm of Dlotko (2019) on an example dataset. In this second part we construct an example BM plot and look at some basic analysis there of. This document supplements the code file that is available on the GitHub repository as Session 3 Part B.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**

It is assumed that you have already completed the first part of the session. If not the necessary code appears at the top of the Session 3 Part B.txt file. You need to remove the # symbol from before the lines of code to get them to run. You are also encouraged to look at the Session 3 Part A Commentary file if you are in any doubt about what any of the lines are doing.

The variables referred to in this session are as follows:

|  |  |
| --- | --- |
| Variable | Interpretation (All are percentages) |
| geog | Name of 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

BM 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. IF you did not create dty earlier then 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 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 1.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1: Head of the reduced dataset dty

**Preparing the Data**

The functions within the BallMapper package all require that the data be in the data.frame format. Hence before we do anything else we will create data.frame for the two outcome variables in which we have interest.

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

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

We also need a data.frame of the variables that are going to be used as our axis variables for the BM plot. These variables are in columns 4 to 8 of the dty data

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

In Part A we noted that the different axes are on different scales, values for Accommodation being much smaller than other characteristics. To overcome this issue in almost all data science applications the first step is to normalise all variables onto the interval [0,1]. The formula for normalisation of observation of variable is:

In the BallMapper package there is a function normalize\_to\_min\_0\_max\_1 which performs the normalisation for us. Therefore we create a new data frame with all values in x1 normalised.

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

We are now ready to produce our first BM plot.

**A First BM Plot**

The command for producing a BallMapper object has three arguments. The data.frame that will be the axes, the data frame that will be the outcome and the radius of the balls.

bm1<-BallMapper(x2,y1,0.3)

Here we are creating a BM plot from the normalised characteristics and colouring according to the percentage of residents in the local authority district who have a degree or higher.

To plot the BM we have to use a further command

ColorIgraphPlot(bm1,seed\_for\_plotting=123)

Note here that the seed\_for\_plotting is set to 123. You can change this as it is like a lens through which to view the BM graph. Choosing other values will alter the appearance as if looking at the plot from a different angle. Changing the seed\_for\_plotting can help when the balls are being placed on top of each other in a projection. Note that changing the seed\_for\_plotting does not change any of the sizes, colours and edges.

Chart, scatter chart

Description automatically generated

In Figure 2 we see the first BM plot that has been created. Colouring is according to the percentage of households where the highest qualification is a university degree or higher with the values indicated on the colour scale. Remember that the colouration is the average value for the local authority districts contained within the ball.

We see that there are two areas where qualifications are high and that these sit at either end of the main shape in the centre and lower right. On the left the levels of qualification are lower.

Figure 2: Initial Ball Mapper Plot

**The Role of Radius**

BM has a single parameter, the radius of the balls. We can easily see the effects of the radius by creating new BM plots with different radii. We will first create the BM graphs using the following code:

bm125<-BallMapper(x2,y1,0.25)

bm130<-BallMapper(x2,y1,0.30)

bm135<-BallMapper(x2,y1,0.35)

bm140<-BallMapper(x2,y1,0.40)

bm145<-BallMapper(x2,y1,0.45)

bm150<-BallMapper(x2,y1,0.50)

We may then plot these objects. It is recommended to run all of the plots and then use the up and down arrows to plot again to allow you to compare the different graphs.

ColorIgraphPlot(bm125,seed\_for\_plotting=123)

ColorIgraphPlot(bm130,seed\_for\_plotting=123)

ColorIgraphPlot(bm135,seed\_for\_plotting=123)

ColorIgraphPlot(bm140,seed\_for\_plotting=123)

ColorIgraphPlot(bm145,seed\_for\_plotting=123)

ColorIgraphPlot(bm150,seed\_for\_plotting=123)

By going through the plots we can see that the inference is very consistent. Local authority districts with the highest levels of qualifications do not sit next to each other in the characteristic space considered in our example. The inference is particularly clear from the radius 0.3 case. We therefore use the bm130 plot. The following code outputs the BM graph to a .png file in exactly the way that we did for the scatter plots.

png("Level430.png")

ColorIgraphPlot(bm130,seed\_for\_plotting=123)

dev.off()

Q8. Describe what happens as the radius of the balls increases. Is this what you would expect?

**Colouring by a Dummy Variable**

We now repeat our BM analysis but use the dummy variable QL4 instead of the continuous variable QualLevel4 used so far. When we calculate the colouration of a ball we are now computing the average of a dummy. The colouration of a ball represents the proportion of points within the ball that have the value 1 for the dummy variable.

bm225<-BallMapper(x2,y2,0.25)

bm230<-BallMapper(x2,y2,0.30)

bm235<-BallMapper(x2,y2,0.35)

bm240<-BallMapper(x2,y2,0.40)

bm245<-BallMapper(x2,y2,0.45)

bm250<-BallMapper(x2,y2,0.50)

Notice now that the ball mappers are labelled as bm2 followed by the radius, with the second argument of the BallMapper() function now being y2 instead of y1. As before we should run the different radii plots and then move through them to carry out a comparison.

ColorIgraphPlot(bm225,seed\_for\_plotting=123)

ColorIgraphPlot(bm230,seed\_for\_plotting=123)

ColorIgraphPlot(bm235,seed\_for\_plotting=123)

ColorIgraphPlot(bm240,seed\_for\_plotting=123)

ColorIgraphPlot(bm245,seed\_for\_plotting=123)

ColorIgraphPlot(bm250,seed\_for\_plotting=123)

Again let us output the plot with radius 0.3, bm230 in this case, as a .png file.

Chart, scatter chart

Description automatically generatedThere are some balls in which every local authority district exceeds the 33% of residents with a degree level qualification or higher. The highest proportions appear at the ends of the main shape and in the outliers. On the left we notice that there are many balls in which the percentage of usual residents with a university degree does not make the threshold.

Q9. Does the colouration in Figure 3 make sense? How does this compare with Figure 2?

Q10. Is the message on proportions consistent across the radii?

Figure 3: Proportion of districts with more than 33% at Level 4

The code to produce Figure 3 is:

png("Proportion430.png")

ColorIgraphPlot(bm230,seed\_for\_plotting=123)

dev.off()

The use of proportion 430 this file name is to represent the use of proportions of respondents with qualifications of level 4 or higher. The 30 is the radius used multiplied by 100 to make it suitable for use as a file name.

**Axis Variables**

Because the BM graph is an abstract representation of a multi-dimensional dataset there are no axes on the plot. Instead we should colour the plot according to the values of each of the axes. Remember that we used normalised variables for the plot and these will not have meaning when viewed independently. Therefore we will colour by the values in x1 not x2.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

(c) Married

(b) Accommodation

(a) Deprivation0

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

(e) Owned Outright

(d) Health Very Good

Figure 4: BallMapper plots coloured by the average value of the five axis variables within each ball. Scales for each plot inform of the value ranges. These plots may be compared with the outcomes in Figures 2 and 3.

Figure 4 provides 5 panels and is generating by editing the output from the colorByAllVariables function:

colorByAllVariables(bm130,x1,"bm1axis")

Here the inputs to the function are a BallMapper object, in this case bm130, the variables that will be used for colouration, in this case x1, and the prefix that will be given to the output files. The code will produce one BM plot for each of the axes and give the file a name with the prefix stated followed by the axis number. Since the first axis variable is Deprivation0, the output bm1axis1.png is coloured by Deprivation0.

Q11. What inference may we draw from our BM graph using Figures 2, 3 and 4?

**Points in Balls**

One of the big advantages of BM is that we know exactly which points are in each ball. Therefore if we want to know more about a particular ball we only need to go back to the R output and find the points associated with that ball. In this final section of Part B we will see how to uncover which regions are in each of the balls.

The first step is to prepare the data with a point number. The BallMapper package treats the data it receives in order, with the output recording a point number and a ball number. Where a point appears in the overlap of two or more balls separate lines appear in the output associating that point with each of the balls. Our task is to extract the coverage information. For this we use a user generated function:

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)

}

Adding a column to the dty dataset to allow merging with the points\_to\_balls output is straightforward with the following code:

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

We may now apply the points\_to\_balls function on our BallMapper object bm130, storing the outcome as bmp1.

bmp1<-points\_to\_balls(bm130)

The resulting output is turned into a dataframe and the columns are named for easier merging with the dataset.

bmp1<-as.data.frame(bmp1)

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

Finally, we use the merge function from R to connect together the data with the points\_to\_balls list.

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

It is useful to have this list to query later. As such let us save dtz as a .csv file as follows:

write.table(dtz,"Ballmembers130.csv",sep=",",row.names=FALSE)

We may now choose to switch to a spreadsheet programme like Excel to query the .csv file that has been produced. However, R has the functionality to let us look at the members of each ball. We will do this using the subset function

ball17<-subset(dtz,dtz$ball==17)

The object ball17 is simply the dataset dtz but only containing observations that are in ball 17. Hence we may simply type ball17 to see the full details of the members. We may also use ball17$geog to just see the list of local authority names.

ball17$geog

ball17

Q12. Repeat the analysis with Ball 19. What can we say about the two ends of the main shape in Figures 2 to 4?

Q13. Which local authority districts are in Ball 21. What can we say about the differences between Ball 21 and Balls 17 and 19?

Q14. How would we construct summary statistics for a ball? (Hint: Refer to Part A)

The process of analysis can continue in many ways, but we leave the discussion here. Session 4 will offer a chance to talk about additional options.

**Summary**

This second half of the practical session has been targeted at giving a basic appreciation of how to use the BallMapper package in R to perform BM analysis. We have continued with the same five variables from Part A and gained further insight into the relationship between the characteristics of local authority districts and the educational attainment of the residents thereof. In many cases simply looking at the data produces new stories that would be of value to regional scientists and practitioners. We may then refer the evidence from this session back to the discussion of session 2.

**Sources:**

The BallMapper package should be cited as:

Pawel Dlotko (2019). BallMapper: The Ball Mapper Algorithm. R package version 0.2.0. <https://CRAN.R-project.org/package=BallMapper>

The original BallMapper algorithm may be found in the working paper of Dlotko (2019):

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

The data used in this session is taken from Nomis and can be extended and explored through [2011 Census - Census of Population - Data Sources - home - Nomis - Official Census and Labour Market Statistics (nomisweb.co.uk)](https://www.nomisweb.co.uk/sources/census_2011?release=5.1a)