MANCHESTER 1824
The University of Manchester

18th November 2025

Topological Data Analysis Ball Mapper for the Social Sciences

Dr Simon Rudkin
Senior Lecturer in Data Science

In this Session...

- Structure of the workshop
- Recap of the Topological Data Analysis Ball Mapper (TDABM) methodology
- Setting up the Python/R environment
- Exploring artificial data
- Initial TDABM plots

Materials are available through the accompanying GitHub site
https://github.com/srudkin12/mm2025

Structure of the Workshop

Session 1	• Getting started with TDABM • Artificial Data
Lunch	• 12:00-13:00
Session 2	• IMD Example • Inference from TDABM plots

Structured around two 2 hour sessions - there will be a chance to ask questions afterwards in the open spaces on 2nd floor

Role of Data Visualisation

Explore

- Structures
- Distributions
- Outliers
- Patterns and Behaviours

Model

- Relationships
- Dependencies
- Outcomes

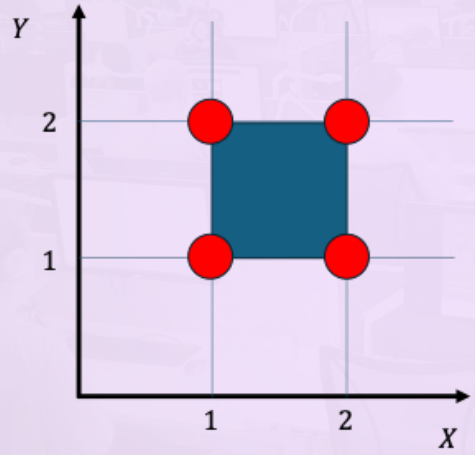
Many ways to classify - here consider exploratory and modelling uses of visualisations

Role of Data Visualisation 2

- Visualisation is about understanding - cognitive process
- Link from image presented and mental image formed
- Seeing data visualisations gives "understanding"
- Danger from preconceptions and biases
- Visualisations should provide insight
- Use visualisations in lectures to guide learning and convey intuition
- Power of visualisation also provides danger

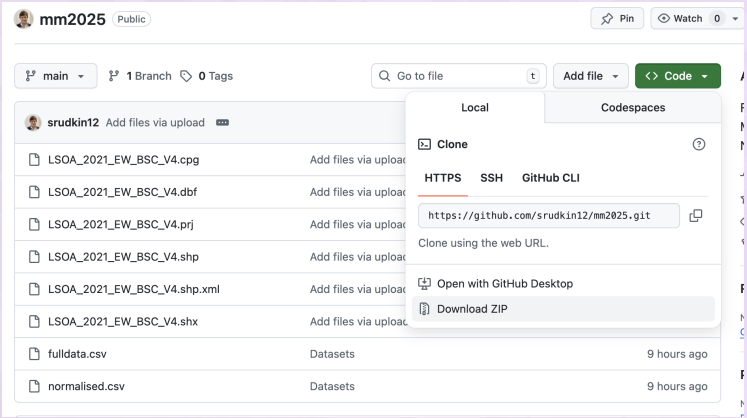
Scatter Plots

X	Y
1	1
1	2
2	1
2	2



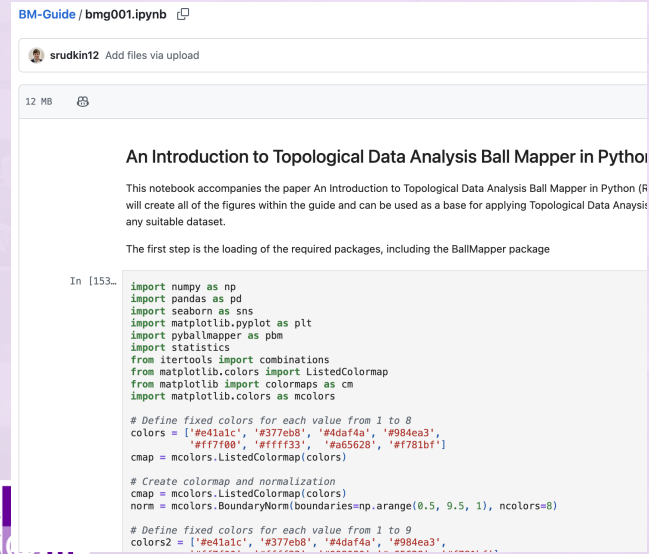
Example with four data points, and two variables, X and Y, creating a square

GitHub https://github.com/srudkin12/mm2025



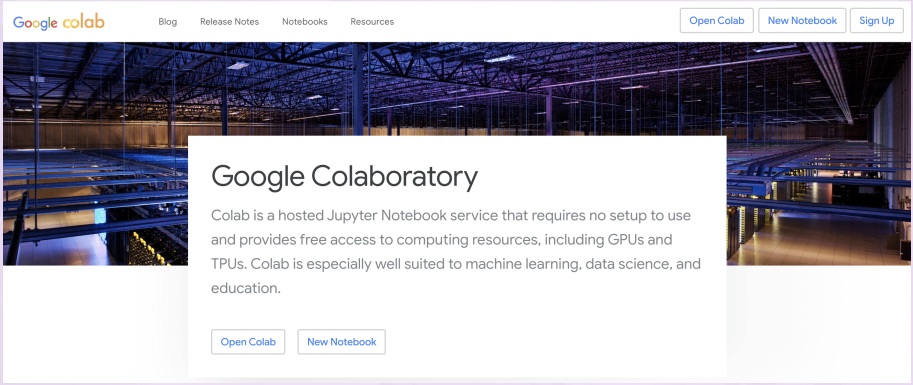
Download the GitHub folder from the link in the title - This folder contains all the files associated with today's workshop

Opening .ipynb on GitHub

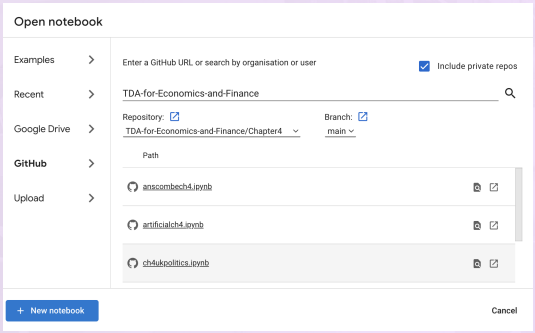


- Code may be opened directly from the GitHub, but not run
- Open `.ipynb` files in either Jupyter Notebook, Visual Studio or other coding editor.
- Upload files to Google Colab and run there
- Copy and paste into your own terminal

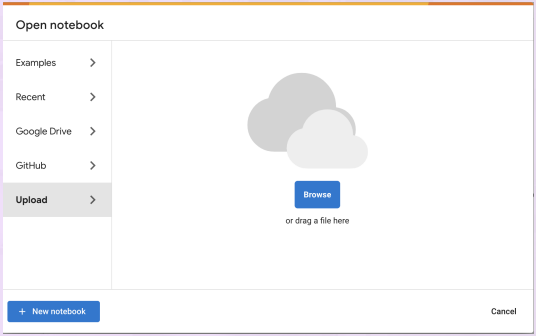
Google Colab https://colab.google/



Notes: You can access Google Colab at https://colab.google/.

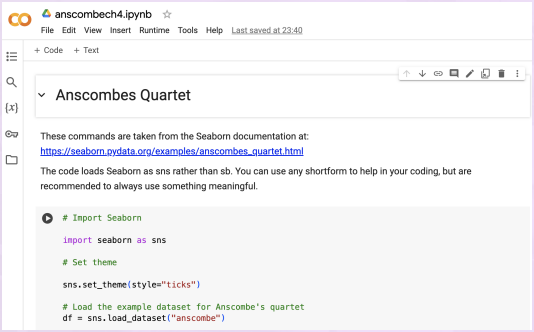


(a) Loading from GitHub

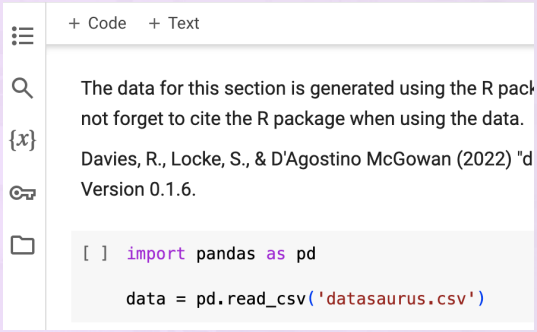


(b) Loading from a local file

Notes: The method to open a `.ipynb` file depends on whether your work is on your computer or whether you are connecting to the textbook GitHub. We show the GitHub connection to Chapter 4 in the example.



(a) Loaded Notebook



(b) Where files are required use folder

Notes: Panel (a) shows the Anscombe's quartet notebook which is from DATA70302. The top code block will run once the file has opened. Where a file is needed it must be uploaded to the colab. Upload files with the little folder icon seen in panel (b). Once uploaded the block with `data = pd.read_csv()` will work.

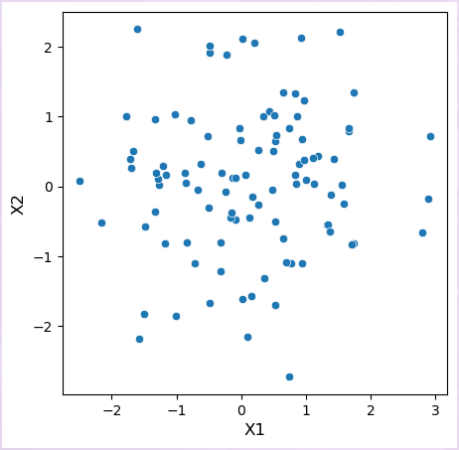
On All Platforms...

You should save your .ipynb files regularly. It is also important to download your files regularly to ensure that a lost internet connection does not delete your work.

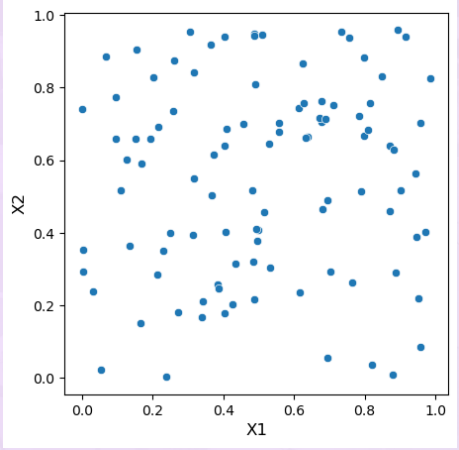
Artificial Data Sets

- Dataset A: 100 observations random normal, $X_1, X_2 \sim N(0, 1)$
- Dataset B: 100 observations random uniform $X_1^B, X_2^B \sim U[0, 1]$
- Both bivariate independent
- There are more for the questions that follow...
- Codes for the artificial dataset in the accompanying `.ipynb` files

Same Distribution? 1

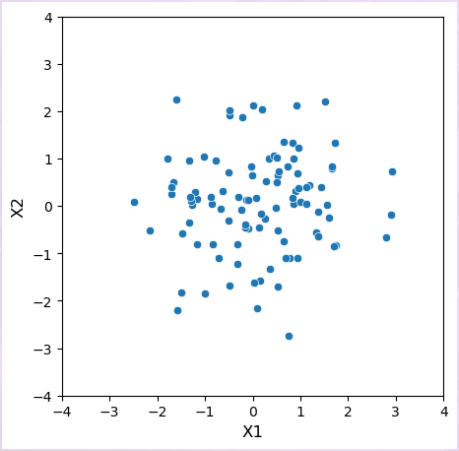


(a) Data set A

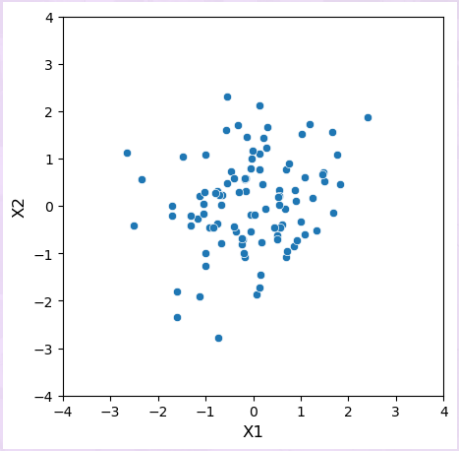


(b) Data set B

Same Distribution? 2

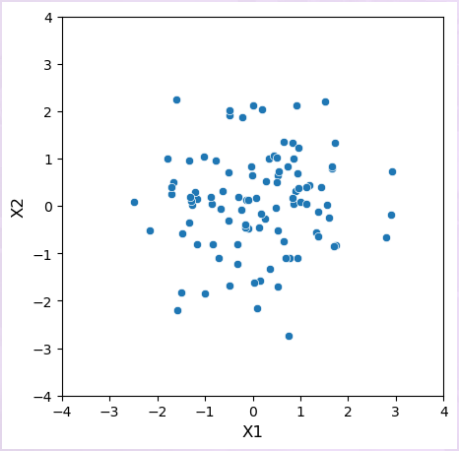


(a) Data set A

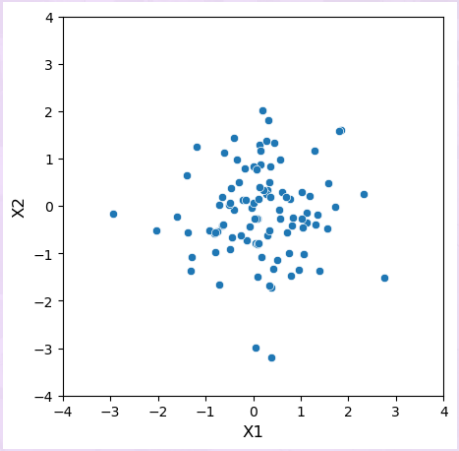


(b) Alternative 1

Same Distribution? 3

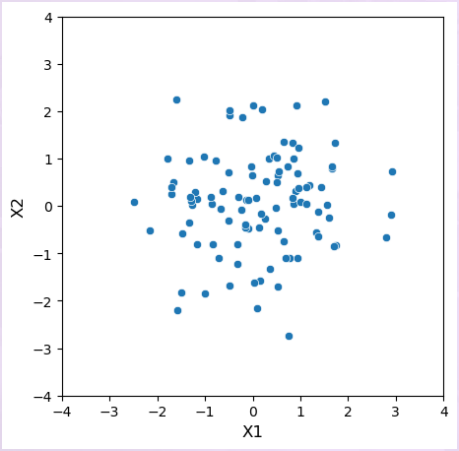


(a) Data set A

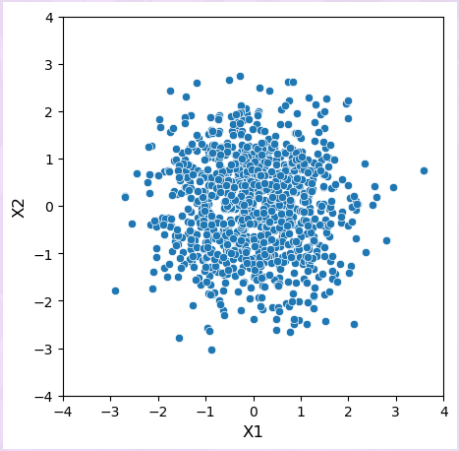


(c) Alternative 2

Same Distribution? 4



(a) Data set A



(d) Alternative 3

Correlation

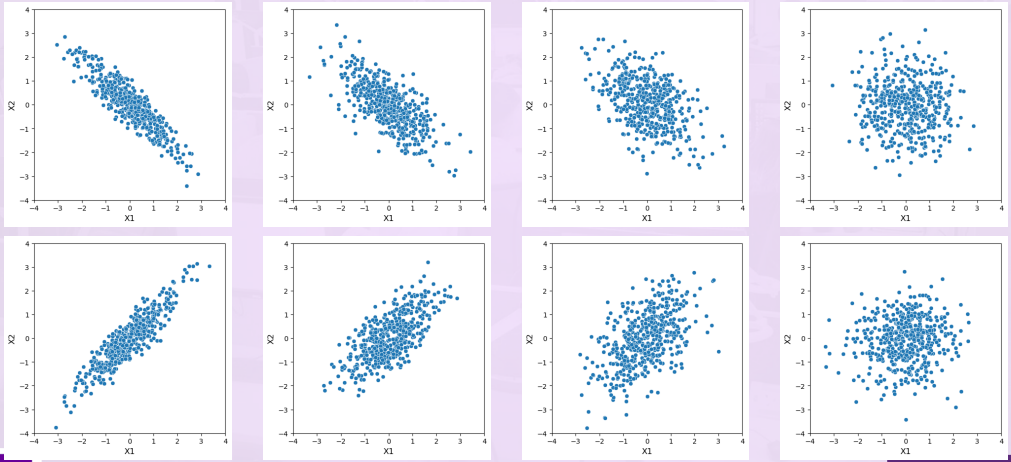
Definition 1

Pearson's Correlation Coefficient: For N observations on X_1 and X_2 with mean $\bar{X}_1$ and $\bar{X}_2$, standard deviations σ_1 and σ_2, and covariance σ_{12}, the Pearson's Correlation Coefficient, ρ_{12} is given by:

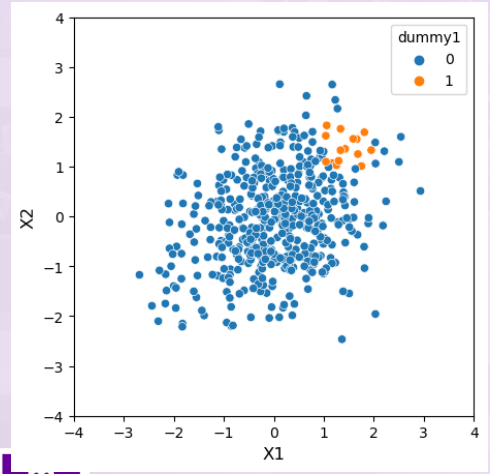
$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sum_{i=1}^{N} \left(X_{i1} - \bar{X}_1\right)\left(X_{i2} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^{N} \left(X_{i1} - \bar{X}_1\right)^2 \sum_{i=1}^{N} \left(X_{i2} - \bar{X}_2\right)^2}} \tag{1}$$

- Unless directly stated, Pearson's correlation is referred to as correlation

Correlation 2

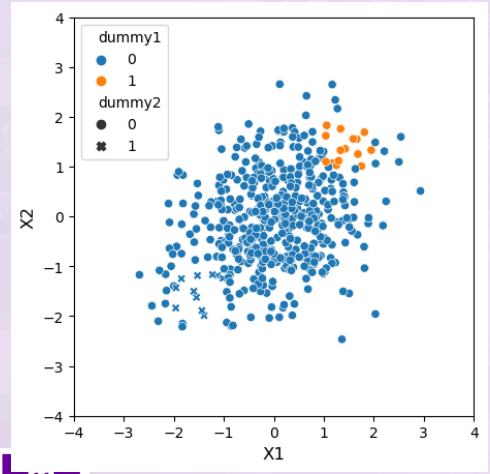


Augmenting Scatter Plots



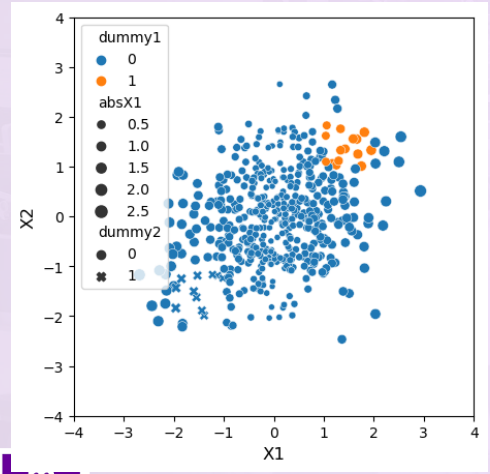
- Artificial dataset with $\rho_{12} = 0.30$
- Highlight points with $X_1 \in [1, 2)$ AND $X_2 \in [1, 2)$ in orange
- Colours stand out provided they do not overlap
- Pay attention to contrasts for visibility

Augmenting Scatter Plots 2



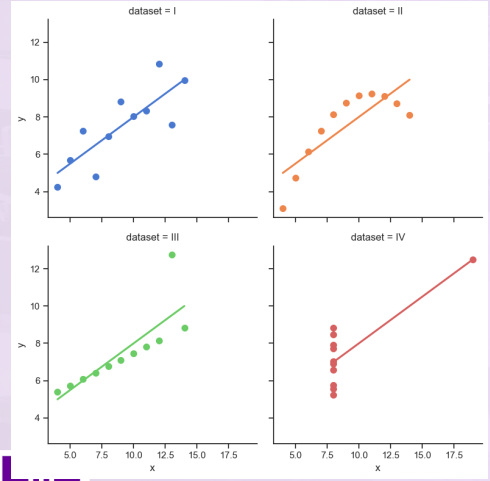
- Artificial dataset with $\rho_{12} = 0.30$
- Highlight points with $X_1 \in [1, 2)$ AND $X_2 \in [1, 2)$ in orange
- Show points with $X_1 \in [-2, -1)$ AND $X_2 \in [-2, -1)$ as crosses
- Shapes need to be distinct to stand out
- Here all crosses are blue

Augmenting Scatter Plots 3



- Artificial dataset with $\rho_{12} = 0.30$
- Highlight points with $X_1 \in [1, 2)$ AND $X_2 \in [1, 2)$ in orange
- Show points with $X_1 \in [-2, -1)$ AND $X_2 \in [-2, -1)$ as crosses
- Size of points based on absolute value of X_1
- Small differences in size hard to see

Anscombe's Quartet (Anscombe, 1973)

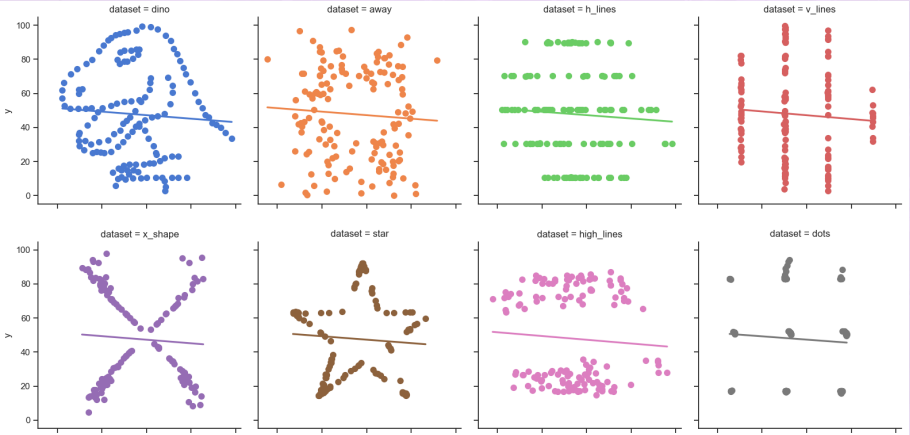


- All datasets have identical mean and standard deviation for each variable
- Correlation between two variables is identical
- Fitted regression line is identical
- R^2 for both regressions is identical
- Regression lines do not reflect reality of data

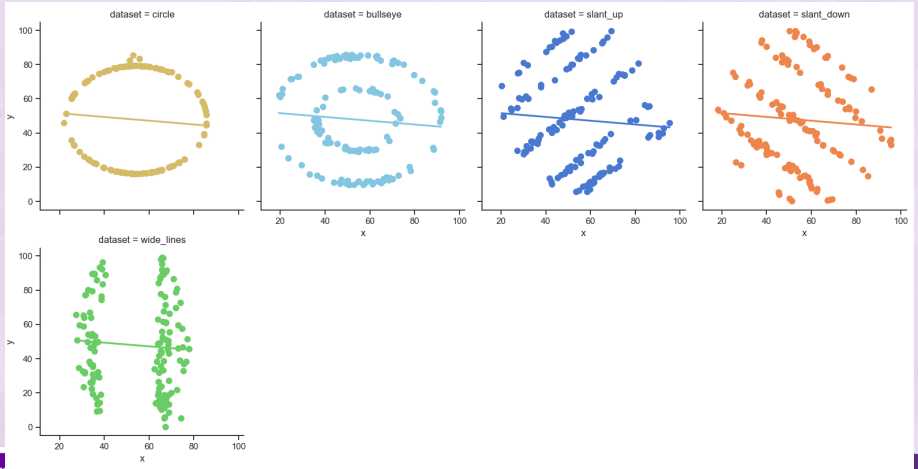
Datasaurus

- Papers report means, standard deviations and correlations
- Is this sufficient to say anything about the data?
- Matejka and Fitzmaurice (2017) extend Anscombe (1973) to show summary statistics insufficient to comment on distribution
- Davies et al. (2022) is an R package with the data
- Use output from the R file in Python

Datasaurus Dozen Plots 1



Datasaurus Dozen Plots 2



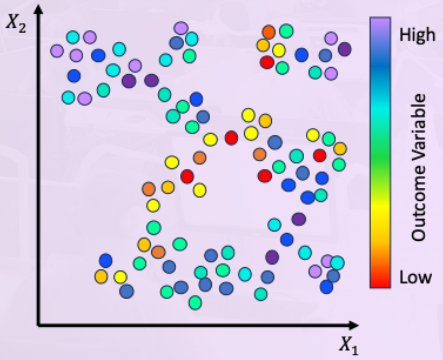
Datasaurus Summary Statistics

Dataset	Var	Mean	s.d.	Min	q25	q50	q75	Max	ρ_{12}
Dino	X_1	54.26	16.77	22.31	44.10	53.33	64.74	98.21	-0.064
	X_2	47.83	26.94	2.95	25.29	46.03	68.53	99.49	
Normal	X_1	54.26	16.77	9.18	42.42	57.43	66.91	100.05	-0.064
	X_2	47.83	26.93	3.64	24.45	46.54	67.97	106.38	
Away	X_1	54.27	16.77	15.56	39.72	53.34	69.15	91.64	-0.064
	X_2	47.83	26.94	0.02	24.63	47.54	71.80	97.48	
Bullseye	X_1	54.27	16.77	19.29	41.63	53.84	64.80	91.74	-0.069
	X_2	47.83	26.94	9.69	26.24	47.38	72.53	85.88	
Circle	X_1	54.27	16.76	21.86	43.38	54.02	64.97	85.66	-0.068
	X_2	47.84	26.93	16.33	18.35	51.03	77.78	85.58	

Topological Data Analysis Ball Mapper (TDABM)

- A ball centered on $x \in X$ is composed of points $y \in X$ such that the distance between x and y is less or equal the radius of the ball.
- All balls have fixed radius ϵ.
- A cover of the space X is obtained by constructing a set of balls, $B(X)$ having the property that $X \subset B(X)$.
- Construct using greedy algorithm based on ϵ-net construction.
- Start from an empty cover, $B(X) = \emptyset$.

Bivariate Construction

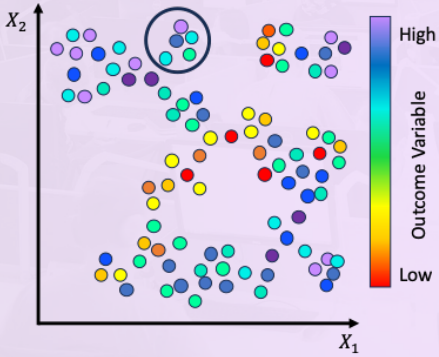


- Artificial dataset selected with holes
- Two variables X_1 and X_2
- Colouration is by third variable Y
- Lower values in red and orange (near leading diagonal)
- Higher values in blue and purple (around edges)

TDABM Algorithm 2

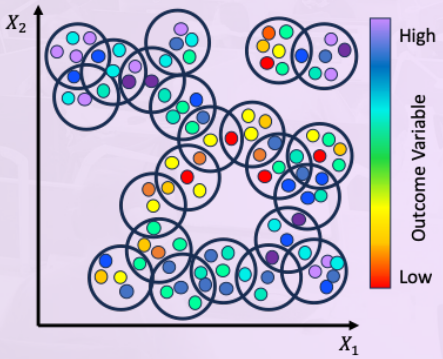
- Iteratively a point $I \in X$ that is not covered by ball in $B(X)$ selected at random.
- A ball, $B(I, \epsilon)$ is drawn and the points $x \in B(I, \epsilon)$ are now considered to be covered
- Ball $B(I, \epsilon)$ is added to $B(X)$.
- Loop ends when there are no longer any uncovered points.
- Number of balls in $B(X, \epsilon)$ is defined as N^B.
- Points at centre of each ball are referred to as landmarks, such that l_b is the landmark of ball b, $b = \{1, ..., N^B\}$.
- N^B is determined by the cover $B(X)$
- $B(X, \epsilon)$ corresponds to abstract vertices V of the constructed TDABM graph.

Bivariate Construction 2



- First ball constructed around landmark l_1 selected at random from points in X
- Show ball as a circle of radius ϵ
- Points in this ball covered by ball
- Example ball contains 5 points

Bivariate Construction 3

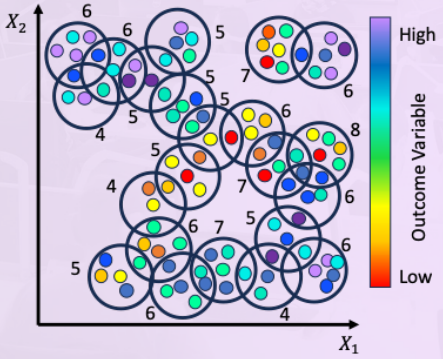


- Continuing to select landmarks from uncovered set produces $B(X)$
- Balls may overlap - points within the intersection
- $B(X)$ satisfies the definitions of a cover from Lecture 1

TDABM Algorithm 3

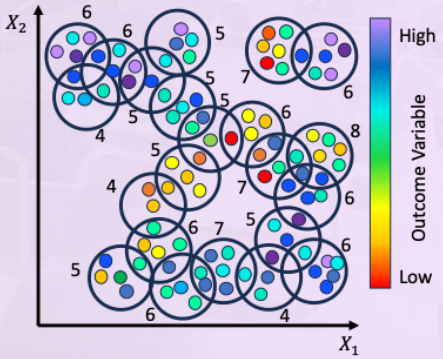
- Taking the cover, $B(X)$, the cloud X, and the radius ϵ, we may identify edges E of the graph.
- Edge between two vertices of the cover, corresponding to balls centered on l_1 and l_2, is drawn if and only if there is a point p_p which sits in $B(l_1, \epsilon) \cap B(l_2, \epsilon)$
- A distance between l_1 and l_2 is bounded by 2ϵ is a necessary, but not a sufficient condition for the existence of an edge
- Requirement for a single point in the intersection may be strengthened to a higher cardinality, such strengthening may be desirable in larger datasets.
- Through consideration of all possible pairs of vertices the set of edges, E is formed. A TDABM graph, $G(V, E)$, may then be presented.
- Total number of edges in the TDABM graph is denoted by N^E.

Bivariate Construction 4



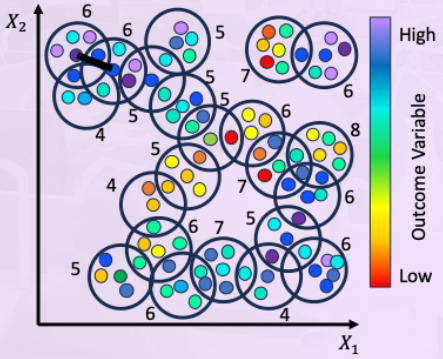
- Record the number of points in each ball
- Computer will do this in practice
- Aim to preserve information on density of joint distribution

Bivariate Construction 5



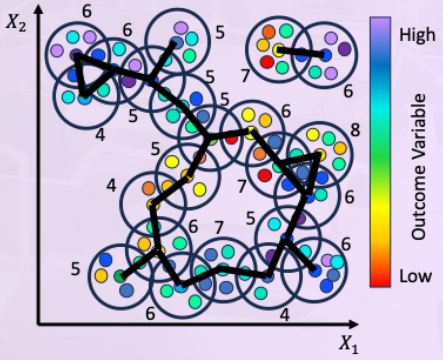
- Recolour the landmarks based on function of points
- Default function is average value of colouring variable
- Implementations in Python and R both use average

Bivariate Construction 6



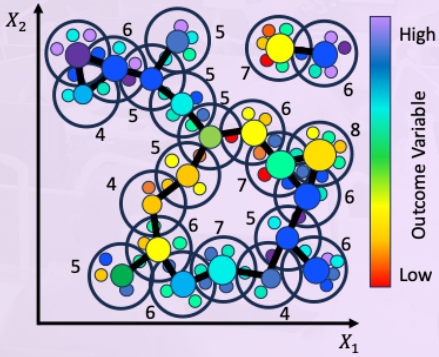
- Draw edges where there are points in the intersection
- Now know which points are close to each other in the space
- Result is the TDABM graph $G(V, E)$

Bivariate Construction 6



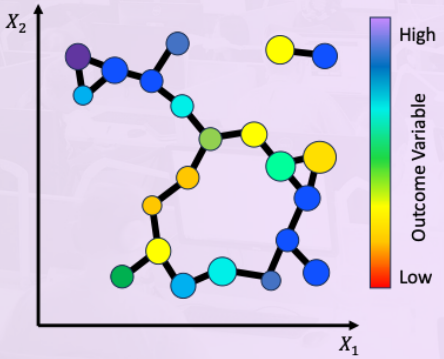
- Draw edges where there are points in the intersection
- Now know which points are close to each other in the space
- Result is the TDABM graph $G(V, E)$

Bivariate Construction 7



- Resize the landmarks based on number of points in ball
- The numbers added to the illustration convey the density
- Colouration provides information on local parts of space

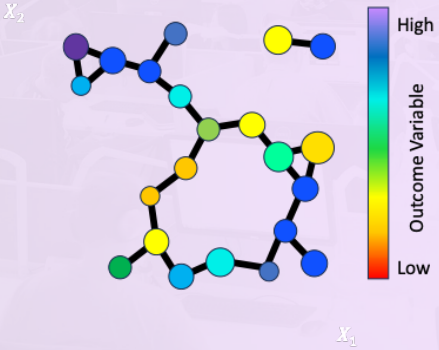
Bivariate Construction 7



- Remove redundant data points

There is no reason to make a two-dimensional plot abstract

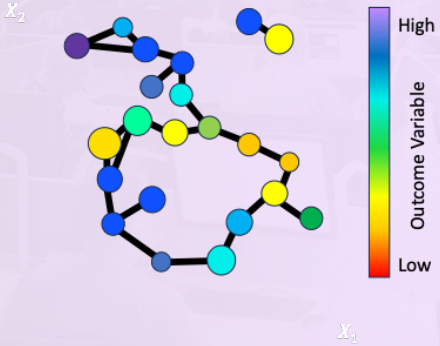
Bivariate Construction 7



- Remove redundant data points
- Remove the axes as TDABM is abstract

There is no reason to make a two-dimensional plot abstract

Bivariate Construction 7



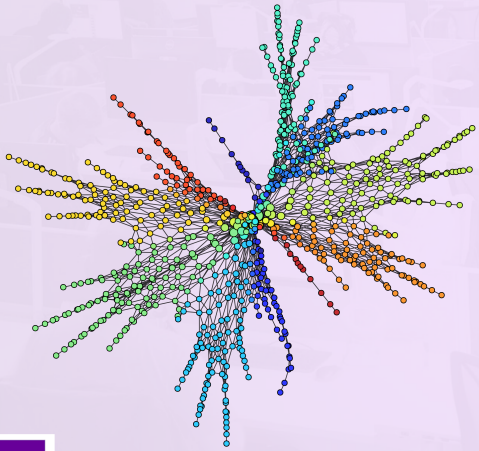
- Remove redundant data points
- Remove the axes as TDABM is abstract
- Make abstract

There is no reason to make a two-dimensional plot abstract

R BallMapper

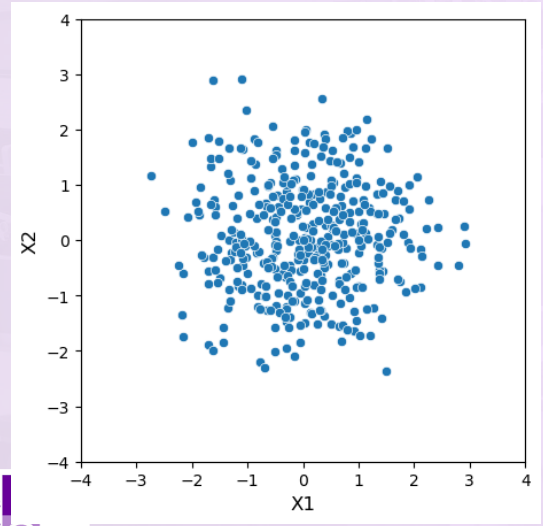
- `BallMapper` Dlotko (2019) ,,,,provides the TDABM algorithm (Dłotko, 2019)
- Documentation available at: https://github.com/srudkin12/BM-Guide
- R package is not actively maintained by the Dioscuri Centre team, but has all necessary functionality

pyballmapper



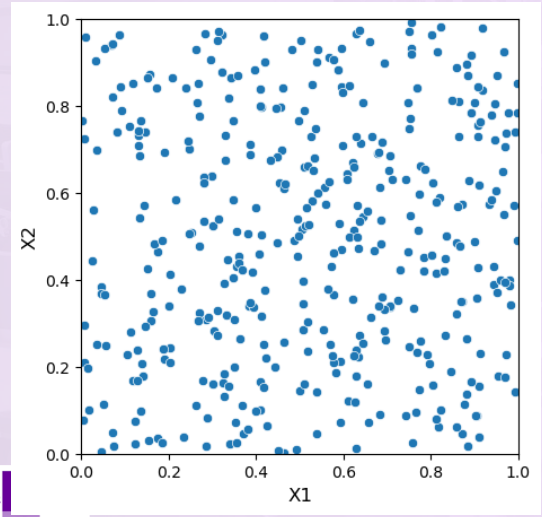
- `pyballmapper` provides the TDABM algorithm (Dłotko, 2019)
- Documentation available at: https://github.com/dioscuri-tda/pyBallMapper
- May need to be installed using `pip install pyballmapper`

Artificial Data A



- $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$
- No clustering within data set
- $N = 400$ to give greater density than previous
- $Y = X_1 + X_2$ as outcome

Artificial Data B

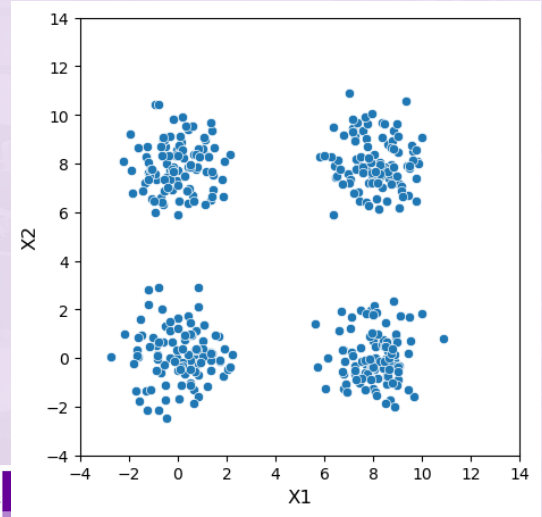


- $X_1 \sim U[0,1]$ and $X_2 \sim U[0,1]$
- No clustering within data set
- $N = 400$ to give greater density than previous
- Note range of values is smaller than other cases
- $Y = X_1 + X_2$ as outcome

Artificial Datasets A and B: Summary Statistics

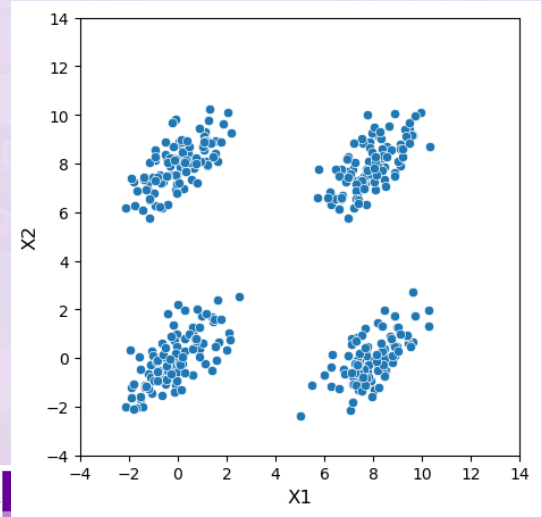
Data Set	Variable	Mean	SD	Min	q25	q50	q75	Max
A	X1	0.067	0.992	-2.365	-0.606	-0.001	0.734	2.904
	X2	0.032	1.034	-2.73	-0.647	0.04	0.739	2.913
	Y	0.1	1.439	-3.91	-0.997	0.118	1.158	3.317
B	X1	0.491	0.281	0.001	0.25	0.489	0.731	0.992
	X2	0.523	0.281	0.002	0.303	0.527	0.762	0.998
	Y	1.015	0.41	0.052	0.741	1.02	1.299	1.898

Artificial Data C



- Comprises 4 subsets each with 100 points and $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$
- Groups are then translated to be centered on $(0,0)$, $(0,8)$, $(8,0)$ and $(8,8)$
- Natural clustering created by groups
- $Y = X_1 + X_2$ as outcome

Artificial Data D



- Comprises 4 subsets each with 100 points and $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$
- Adjusted to have correlation of 0.3 between X_1 and X_2
- Groups are then translated to be centered on $(0,0)$, $(0,8)$, $(8,0)$ and $(8,8)$
- Natural clustering created by groups
- $Y = X_1 + X_2$ as outcome

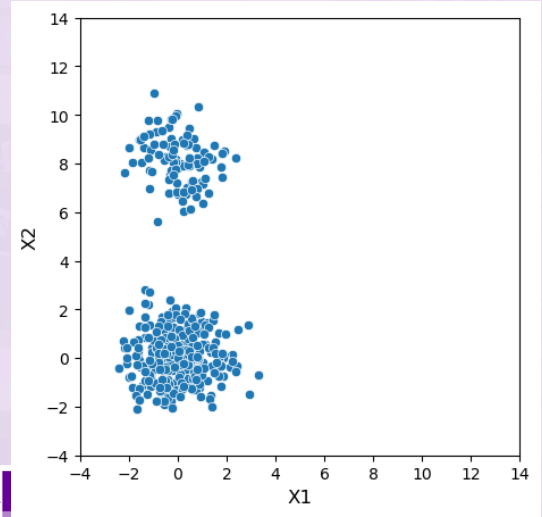
Artificial Datasets C and D: Summary Statistics

Data Set	Variable	Mean	SD	Min	q25	q50	q75	Max
C	X1	4.067	4.155	-2.73	0.051	3.945	8.129	10.895
	X2	4.033	4.09	-2.492	0.011	4.413	7.873	10.904
	Y	8.1	5.823	-3.424	4.565	8.067	11.913	19.883
D	X1	3.926	4.097	-2.141	-0.07	3.761	7.795	10.325
	X2	3.951	4.121	-2.358	-0.122	4.243	7.94	10.241
	Y	7.877	5.928	-4.156	4.044	7.944	12.336	20.055

Artificial Datasets A to D: Correlation

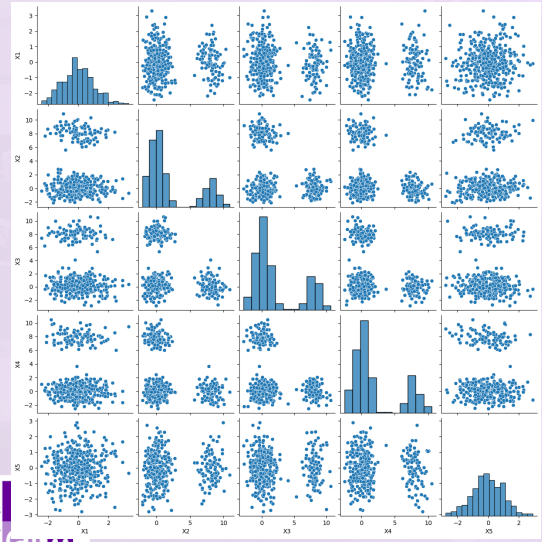
		X_1	X_2	Y			X_1	X_2	Y
A	X_1	1.0	0.008	0.724	B	X_1	1.0	0.065	0.73
	X_2	0.008	1.0	0.695		X_2	0.065	1.0	0.73
	Y	0.724	0.695	1.0		Y	0.73	0.73	1.0
C	X_1	1.0	-0.003	0.712	D	X_1	1.0	0.041	0.719
	X_2	-0.003	1.0	0.701		X_2	0.041	1.0	0.723
	Y	0.712	0.701	1.0		Y	0.719	0.723	1.0

Artificial Data 5



- Comprises 4 subsets each with 100 points and $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, $X_3 \sim N(0,1)$, $X_4 \sim N(0,1)$ and $X_5 \sim N(0,1)$
- Translated to have centres at $(0,0,0,0,0)$, $(0,8,0,0,0)$,$(0,0,8,0,0)$ and $(0,0,0,8,0)$
- $Y = X_1 + X_2 + X_3 + X_4 + X_5$

Artificial Data 6



- Seaborn pairplot
- Still cannot see the shape of the data in these plots
- Data shape not visible in any of these plots

Summary Statistics for Multivariate Case

Data Set	Variable	Mean	SD	Min	q25	q50	q75	Max
E	X1	-0.017	1.012	-2.423	-0.705	-0.071	0.608	3.307
	X2	2.05	3.654	-2.086	-0.375	0.367	3.511	10.904
	X3	2.082	3.609	-2.846	-0.371	0.438	4.402	10.652
	X4	1.975	3.556	-2.511	-0.408	0.491	4.246	10.486
	X5	-0.029	1.008	-2.787	-0.673	-0.024	0.687	2.895
	Y	6.06	3.984	-5.812	3.763	6.996	8.925	12.915

Correlation Matrix for Multivariate Case

	X_1	X_2	X_3	X_4	X_5	Y
X_1	1.0	0.003	-0.112	0.01	0.076	0.183
X_2	0.003	1.0	-0.295	-0.324	0.02	0.367
X_3	-0.112	-0.295	1.0	-0.326	-0.009	0.314
X_4	0.01	-0.324	-0.326	1.0	-0.019	0.297
X_5	0.076	0.02	-0.009	-0.019	1.0	0.266
Y	0.183	0.367	0.314	0.297	0.266	1.0

- Some correlation between pairs (e.g. X_2 and X_3)

TDABM in Python

```
bm1=pbm.BallMapper(X=bmx1, eps=14, coloring_df=cdf)
```

- Axis variables input as X - here `bmx1`
- Radius is set with second argument
- Colour variable set with third argument - here `cdf`

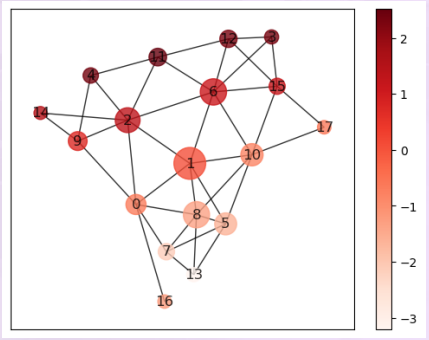
```
bmx1 = df2[['hc1','oc1','ca1','ns1','he1','qu1','de1']]
cdf = pd.DataFrame(df2['Con19'])
```

TDABM in Python 2

```
bm1.draw_networkx(coloring_variable='Con19', colorbar=True)
```

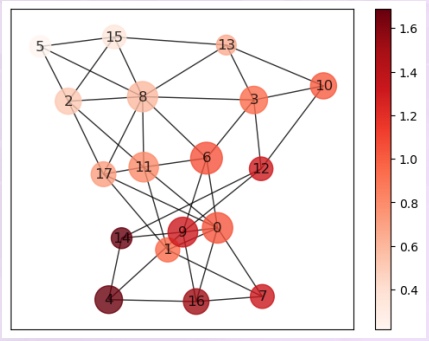
- Use the `networkx` package to plot the graph
- Specify the coloring variable. This should be the same as the one specified in `cdf`
- Specify that we want the coloration bar on the right

TDABM Examples A



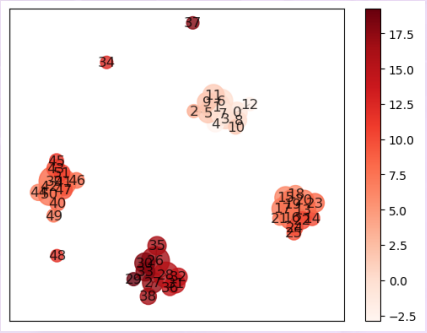
- $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$
- No clustering within data set
- $N = 400$ to give greater density than previous
- Density in centre as larger balls

TDABM Examples B



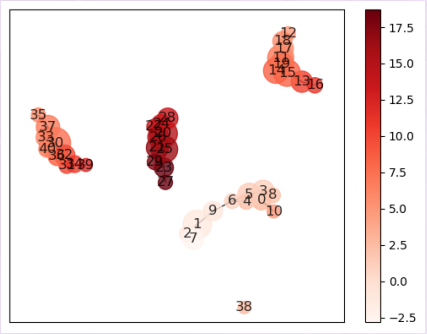
- $X_1 \sim U[0,1]$ and $X_2 \sim U[0,1]$
- No clustering within data set
- $N = 400$ to give greater density than previous
- Ball sizes are similar across the graph

TDABM Examples C



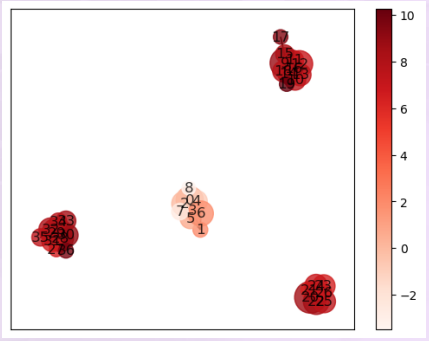
- Data process has 4 sub groups each with 100 points
- Four main groups emerge but with some disconnected balls
- Can play with parameters to obtain "correct" plot

TDABM Examples D



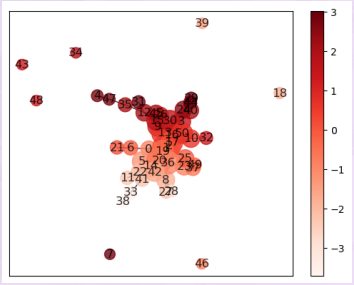
- Data process has 4 sub groups each with 100 points
- Four main groups emerge but with some disconnected balls
- Can play with parameters to obtain "correct" plot
- Correlation within group creates longer and thinner sub graphs

TDABM Examples E

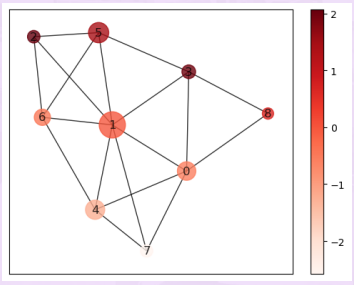


- Data process has 4 sub groups each with 100 points
- Four groups ARE identified in the 5 variable case
- Closeness of points within a sub group draws balls together and leaves big gaps
- Adjustment needed to tease out sub group structure

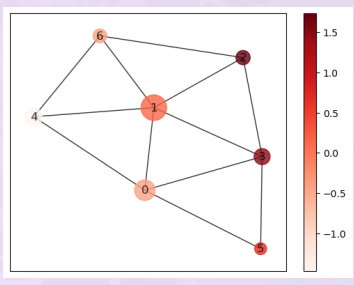
Role of Radius A



$\epsilon = 0.5$

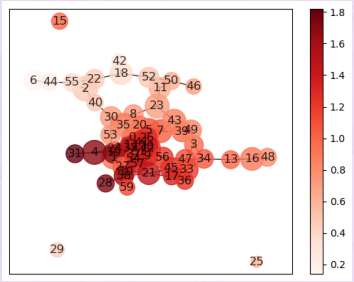


$\epsilon = 1.5$

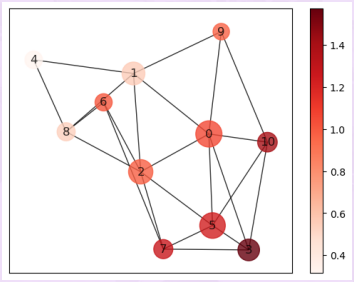


$\epsilon = 2$

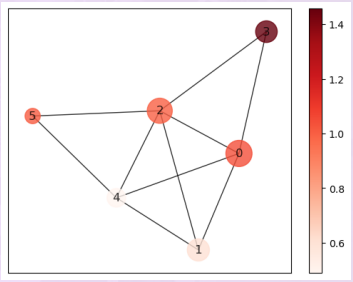
Role of Radius B



$\epsilon = 0.1$

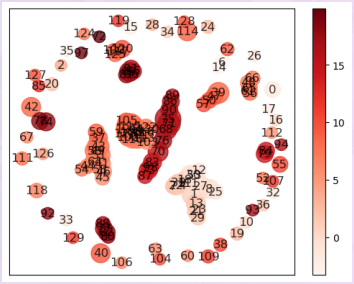


$\epsilon = 0.3$

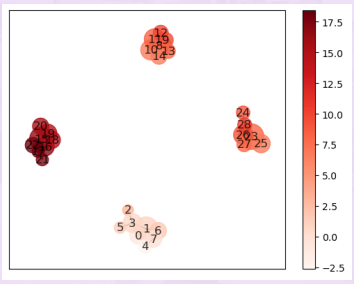


$\epsilon = 0.4$

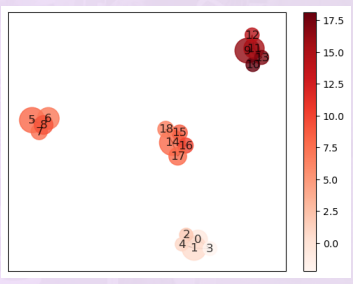
Role of Radius C



$\epsilon = 0.5$

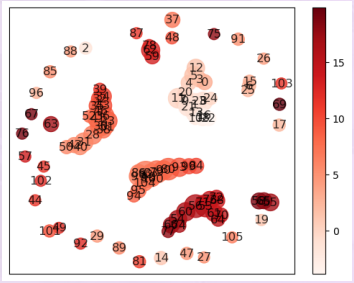


$\epsilon = 1.5$

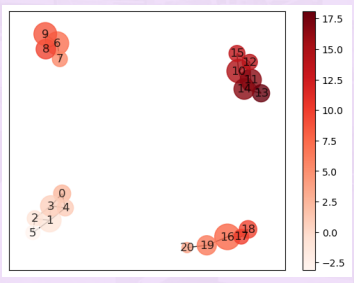


$\epsilon = 2$

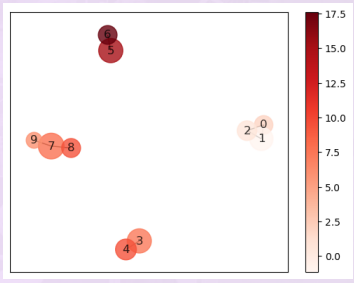
Role of Radius D



$\epsilon = 0.5$

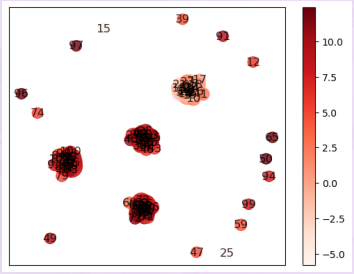


$\epsilon = 1.5$

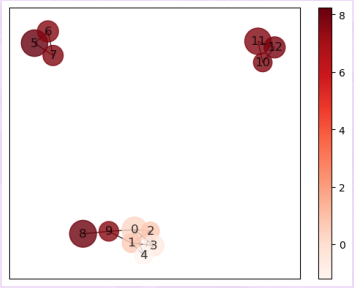


$\epsilon = 2$

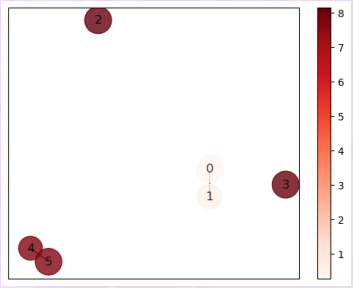
Role of Radius E



$\epsilon = 1$



$\epsilon = 3$

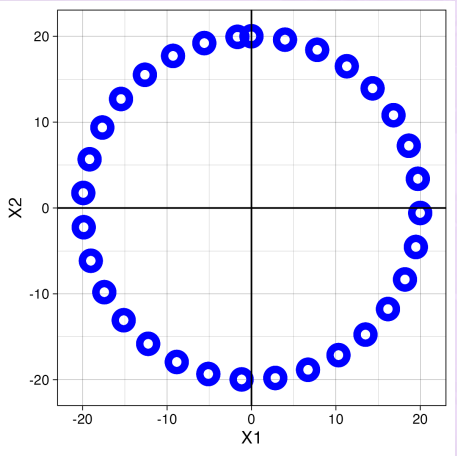


$\epsilon = 4$

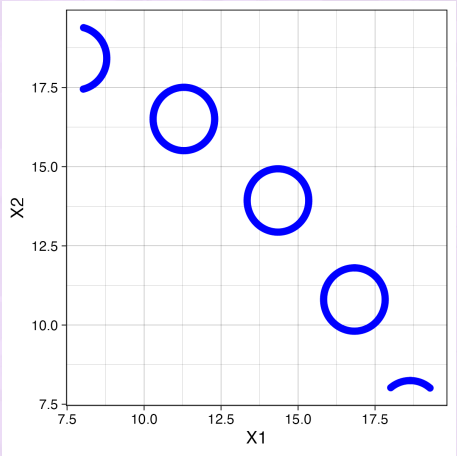
Role of Radius: Summary

- Smaller radii pick up local structure but leave disconnect in global picture
- Larger radii show global structure but merge local features
- Persistence across radii of densest regions
- Discussion on the extent of persistence observed by changing radius
- Consider the optimal radius using R...

Optimum Radius? Example Datasets

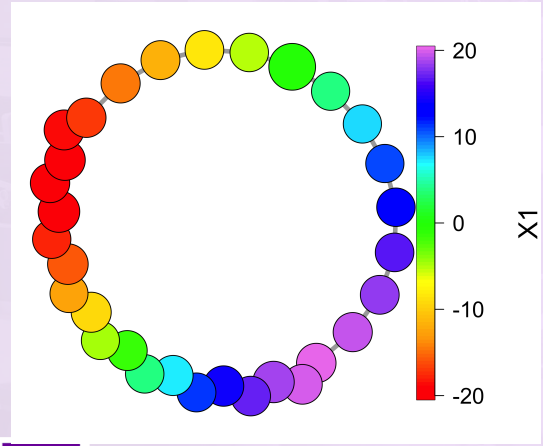


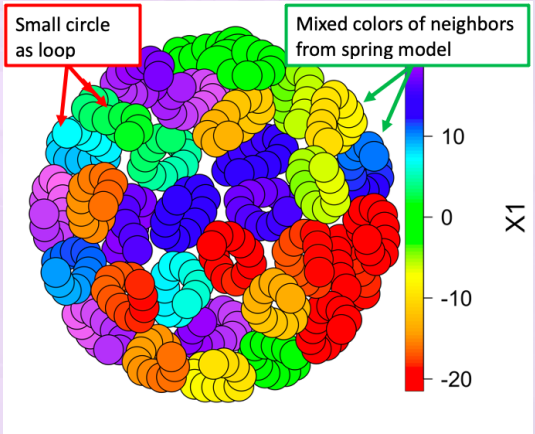
(a) Full Dataset



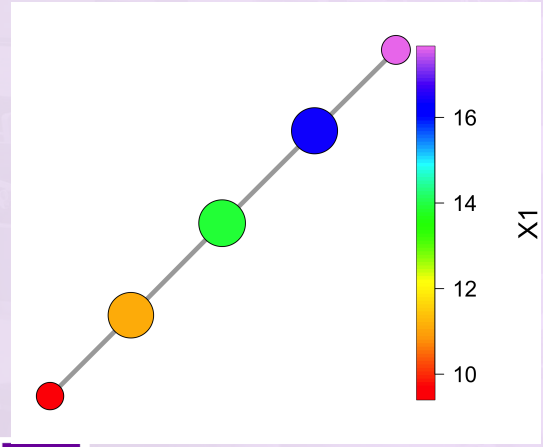
(b) Zoomed Section

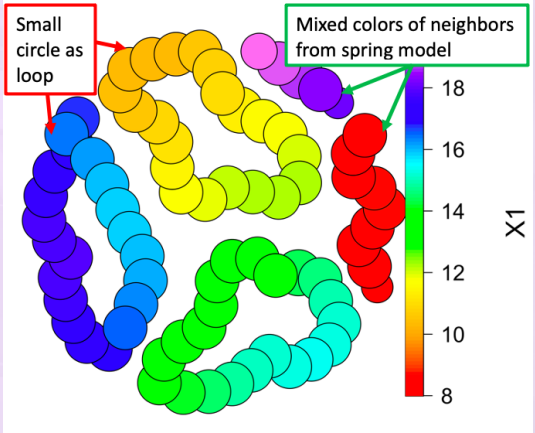
Optimum Radius ? (Full Dataset)





Optimum Radius ? (Zoomed Section)





Index of Multiple Deprivation

- The most deprived neighbourhood in England according to the IMD25 is to the east of the Jaywick & St Osyth area of Clacton-on-Sea in Tendring
- Seven neighbourhoods in Blackpool rank amongst the top 10 most deprived in England according to the IMD25.
- Overall, 82% of neighbourhoods that are in the most deprived decile according to the IMD25 were also the most deprived according to the IMD19.
- Deprivation is dispersed across England - 65% of Local Authority Districts contain at least one of the most deprived neighbourhoods in England.
- Middlesbrough, Birmingham, Hartlepool, Kingston upon Hull and Manchester are the Local Authority Districts with the highest proportions of neighbourhoods among the most deprived in England.

Index of Multiple Deprivation 2

There are 7 domains of deprivation, which combine to create the Index of Multiple Deprivation (IMD25):

- Income (22.5%): Measures the proportion of the population experiencing deprivation relating to low income.
- Income Deprivation Affecting Children Index: Measures the proportion of all children aged 0 to 15 living in income deprived families
- Income Deprivation Affecting Older People Index: Measures the proportion of those aged 60+ who experience income deprivation
- Employment (22.5%): Measures the proportion of working age population involuntarily excluded from the labour market.
- Education (13.5%): Measures the lack of attainment and skills in the local population.

Index of Multiple Deprivation 3

There are 7 domains of deprivation, which combine to create the Index of Multiple Deprivation (IMD25):

- Health (13.5%): Measures the risk of premature death and impairment of life quality through poor physical/mental health.
- Crime (9.3%): Measures the risk of personal and material victimisation at local level.
- Barriers to housing and services (9.3%): Measures the physical and financial accessibility of housing and local services.
- Living environment (9.3%): Measures the quality of both the 'indoor' and 'outdoor' local environment.

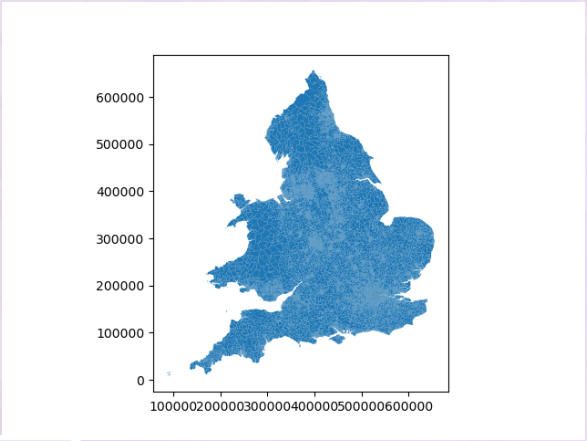
Index of Multiple Deprivation Usage: CAN

- compare small areas across England
- identify the most deprived small areas
- explore the domains (or types) of deprivation
- compare larger administrative areas e.g. local authorities
- look at changes in relative deprivation between iterations (i.e. changes in ranks)

Index of Multiple Deprivation Usage: CANNOT

- quantify how deprived a small area is
- identify deprived people
- say how affluent a place is
- compare with small areas in other UK countries
- measure absolute change in deprivation over time

Loading the Shapefiles



- Shapefile is called `LOSA_2021_EW_BSC_V4.shp`
- Map is created using geopandas following the code in the Python codebook
- R version is similar, showing how many LSOA there are
- This workshop will focus on Greater Manchester...

Loading the Data

- Main data is provided as `fulldata.csv`
- Normalised data is provided as `normalised.csv`

```
df1 = pd.read_csv('fulldata.csv')
nd1 = pd.read_csv('normalised.csv')
```

Summary Statistics

Var	Mean	SD	Min	q25	q50	q75	Max
IMD	28.45	18.89	0.604	11.86	24.96	43.08	82.63
INC	0.283	0.187	0.011	0.111	0.254	0.435	0.812
EMP	0.170	0.098	0.007	0.085	0.154	0.239	0.555
EST	24.576	18.99	0.093	7.611	20.95	38.15	84.58
HDD	0.561	0.802	-1.970	-0.032	0.616	1.149	2.962
CRM	0.490	0.781	-1.950	-0.033	0.563	1.051	2.575
BHS	19.58	9.732	3.044	12.35	17.95	24.48	60.95
LIV	25.29	14.94	0.460	13.50	22.75	34.32	80.33
CYP	-0.012	0.925	-2.612	-0.646	0.083	0.651	2.752
ADS	0.264	0.114	0.022	0.177	0.253	0.344	0.601

Correlation Matrix

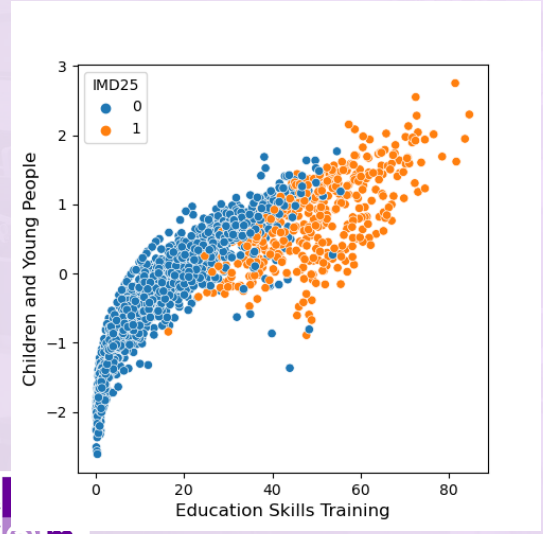
	IMD	INC	EMP	EST	HDD	CRM	BHS	LIV	CYP	ADS
IMD	1.000	0.975	0.962	0.927	0.888	0.841	0.564	0.210	0.771	0.895
INC	0.975	1.000	0.951	0.912	0.830	0.776	0.537	0.164	0.733	0.916
EMP	0.962	0.951	1.000	0.896	0.865	0.791	0.433	0.055	0.768	0.859
EST	0.927	0.912	0.896	1.000	0.810	0.745	0.446	0.125	0.856	0.918
HDD	0.888	0.830	0.865	0.810	1.000	0.832	0.441	0.131	0.812	0.751
CRM	0.841	0.776	0.791	0.745	0.832	1.000	0.404	0.250	0.750	0.682
BHS	0.564	0.537	0.433	0.446	0.441	0.404	1.000	0.116	0.301	0.525
LIV	0.210	0.164	0.055	0.125	0.131	0.250	0.116	1.000	0.066	0.196
CYP	0.771	0.733	0.768	0.856	0.812	0.750	0.301	0.066	1.000	0.694
ADS	0.895	0.916	0.859	0.918	0.751	0.682	0.525	0.196	0.694	1.000

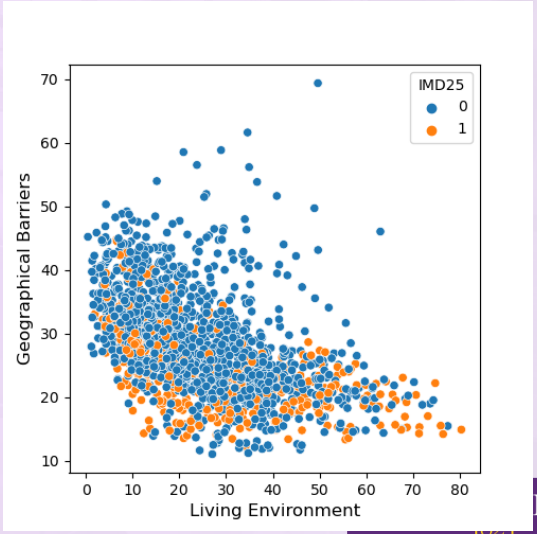
Dummy Variables

```
df1['IMD25'] = (df1['IMD']>df1['IMD'].quantile(0.75))*1
df1['IDACI25'] = (df1['IDACI']>df1['IDACI'].quantile(0.75))*1
df1['IDAOPI25'] = (df1['IDAOPI']>df1['IDAOPI'].quantile(0.75))*1
```

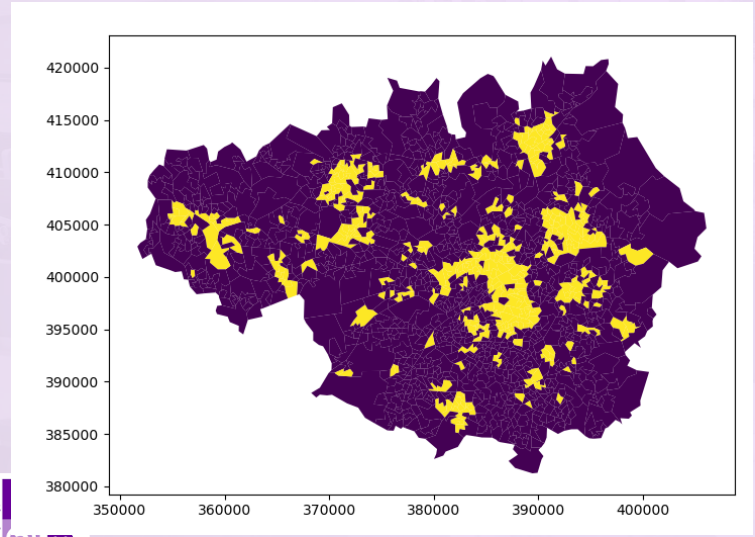
- Create dummies to identify LSOAs in the top 25% of the Manchester data.
- Use the dummies as colours in the scatterplots on next slide

Scatter Plots





Manchester Deprivation Map



Concentration of the high deprivation around urban areas and East Manchester

Preparing for TDABM

```
nd1 = nd1.sort_values(by="LSOA2021").reset_index(drop=True)
nd1["pt"] = nd1.index
nd1.head()
df1 = df1.sort_values(by="LSOA2021").reset_index(drop=True)
```

We sort according to the LSOA so that the row index is consistent across the data

Preparing for TDABM2

```
adf = nd1[['INCN','EMPN','ESTN','HDDN','CRMN','BHSN','LIVN','CYPN','ADSN
cdf1 = df1[['IMD']]
cdf2 = df1[['IMD25']]
cdf3 = df1[['IDACI25']]
cdf4 = df1[['IDAOPI25']]
```

Axis and colour data must be a data frame

TDABM Command

```
bm1=pbm.BallMapper(X=adf, coloring_df=cdf1, eps=0.5)
```

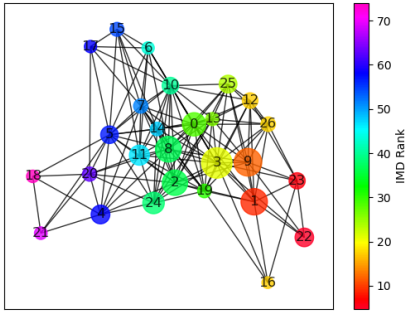
- Three inputs needed, axes, colouration and radius
- Note Python feature of setting out the arguments carefully

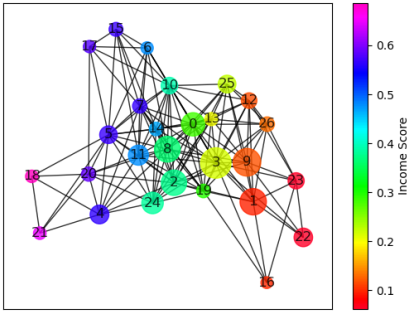
TDABM Plotting Command

```
hsvp = cm.get_cmap("gist_rainbow")
bm1.draw_networkx(coloring_variable='IMD', color_palette=hsvp,
  colorbar=True, colorbar_label="IMD Rank")
```

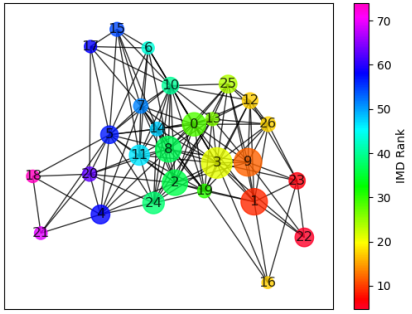
- Here we specify a colouring scheme, rainbow to match the published papers
- A label is added to the colour bar to help readability

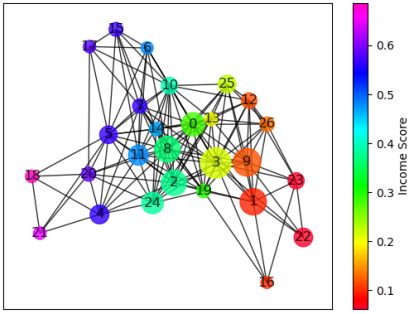
TDABM Plots + INC





TDABM Plots +





Reading from TDABM Plots - Ball Membership

```
pb1 = bm1.points_and_balls()
nd2 = nd1.merge(pb1, on='point')
df5 = df1.merge(nd2, on='LSOA2021')
```

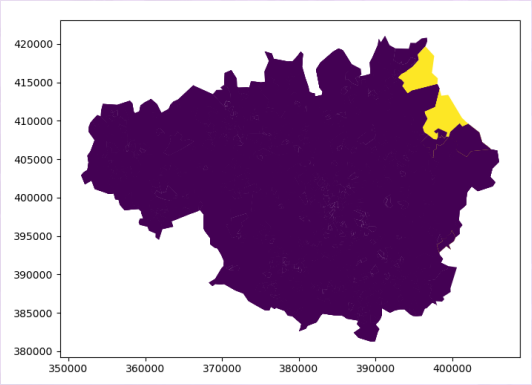
- Merge with the data frame using the point
- Merge with the larger data set using the LSOA code

Seeing Ball Membership

```
ballslist = pd.DataFrame(df5[["ball","LSOA2021","LSOA_NAME_2021",
  "LAD_NAME_2024","IMD_RANK","IMD"]])
ballslist.to_csv('ballslistr5.csv', index=False)
```

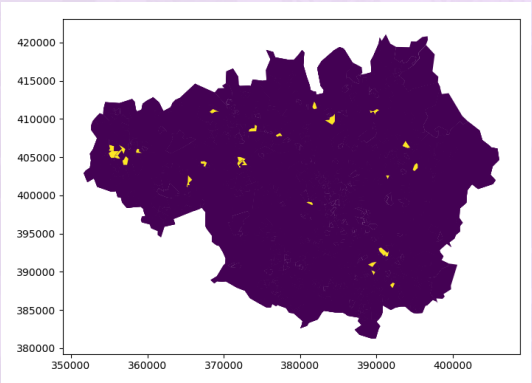
```
ball16 = ballslist[ballslist["ball"] == 16]
```

Mapping Ball Membership



- Ball 16 was down to the lower right and had differing characteristics to its neighbours despite also having low deprivation
- See mostly semi-rural areas of Oldham and Rochdale
- Ball 21 scatters across the suburbs. Here children were more exposed to deprivation than the overall index

Mapping Ball Membership



- Ball 16 was down to the lower right and had differing characteristics to its neighbours despite also having low deprivation
- See mostly semi-rural areas of Oldham and Rochdale
- Ball 21 scatters across the suburbs. Here children were more exposed to deprivation than the overall index

Summary

- This workshop has introduced TDABM as a means of visualising data
- An empirical example has been presented using the Index of Multiple Deprivation
- Using Greater Manchester data we have seen greater heterogeneity amongst deprived neighbourhoods
- Using ball membership allows identification of ball members
- Mapping brings a spatial understanding to the data
- Now to understand what is behind all of the patterns, add variables...

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Davies, R., Locke, S., and D'Agostino McGowan, L. (2022). *datasauRus: Datasets from the Datasaurus Dozen*. R package version 0.1.6.

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.

Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1290–1294.