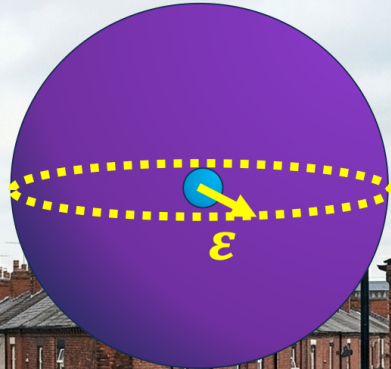


ballmapper: Topological Data Analysis Ball Mapper for Stata

Dr Simon Rudkin and
Dr Wanling Rudkin

MANCHESTER
1824



In this Presentation...

- Why multivariate visualisation creates an empirical problem
- Topological Data Analysis Ball Mapper (TDABM)
- The Stata implementation: `ballmapper`
- Artificial X-shaped data example
- Built-in Stata auto data example
- Reading outputs, summaries, and radius sensitivity

QR Code links to my GitHub with the ballmapper package.



Motivation: Seeing More Than Two Variables

- Empirical datasets usually have many variables.
- Standard scatterplots are easy to interpret, but only in 2D or 3D.
- PCA, t-SNE and UMAP give useful pictures, but after transforming the data.
- TDABM instead covers the original multivariate point cloud with balls.

Question

How can we produce a readable map of a K -dimensional dataset without making the plot itself the estimand?

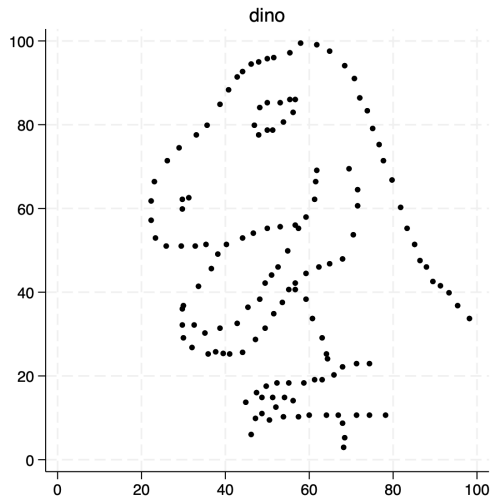


- Papers report means, standard deviations and correlations
- Is this sufficient to say anything about the data?

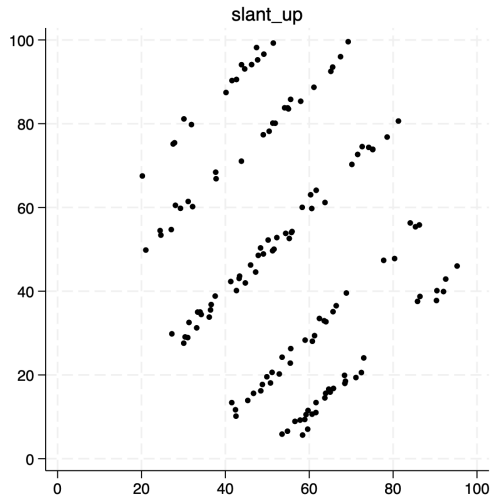
Question

Two variables with means 54.26 and 47.83, standard deviations 16.77 and 26.94, correlation -0.065. What does the scatter plot of these variables look like?

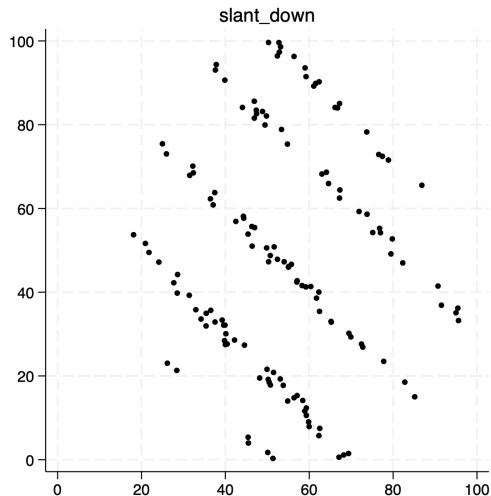
Datasaurus Dozen Plots 1



Datasaurus Dozen Plots 2



Datasaurus Dozen Plots 3



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	

Matejka and Fitzmaurice (2017) extend Anscombe (1973) on importance of visualisation. BallMapper allows visualisation.

Main Contribution of the Paper

- Introduces TDABM for Stata users.
- Provides the `ballmapper` command and supporting summary tools.
- Demonstrates the package using controlled artificial data and Stata's auto dataset.
- Shows how ball membership supports interpretation after the plot has been produced.

Core message: TDABM gives a model-free visualisation of multivariate structure, while retaining a link between the abstract graph and the original observations.

What TDABM Takes as Inputs

- A dataset X with K variables.
- A distance metric on the scaled variables.
- A radius ϵ .
- A colouring variable Y .

Practical requirement

Variables used in the distance calculation need to be non-missing and placed on comparable scales.

Parameter discipline

The radius is not a nuisance detail; it is the scale at which the data are being viewed.

Ball Mapper Algorithm: Intuition

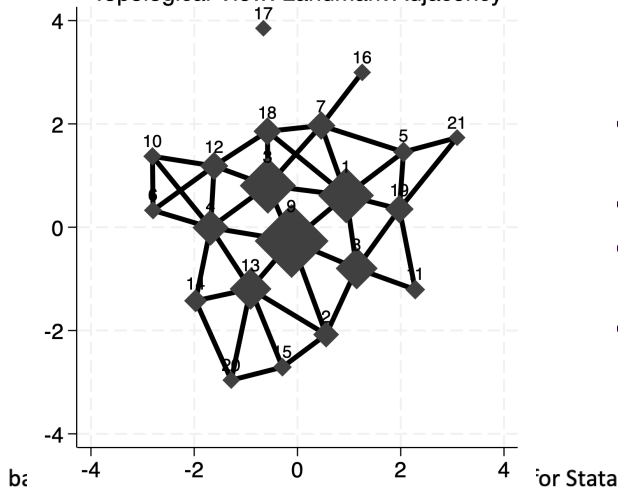
- 1 Select an uncovered point as a landmark.
- 2 Draw a ball of radius ϵ around it.
- 3 Mark all observations inside that ball as covered.
- 4 Repeat until every observation is covered.
- 5 Draw edges between balls with non-empty overlap.

Balls summarise local neighbourhoods. Edges summarise overlap. Ball sizes summarise local density. Colour summarises Y within each neighbourhood.

From Point Cloud to Abstract Graph

Ball Mapper: Abstract Skeleton

Topological View: Landmark Adjacency



- A ball is drawn in the original K -dimensional space.
- The final graph is plotted in 2D.
- Axes in the final TDABM plot do not carry substantive interpretation.
- Interpretation comes from membership, size, edge structure and colour.

Installing the Stata Package

```
net install ballmapper, ///  
    from("https://raw.githubusercontent.com/srudkin12/statabm/main") ///  
    replace
```

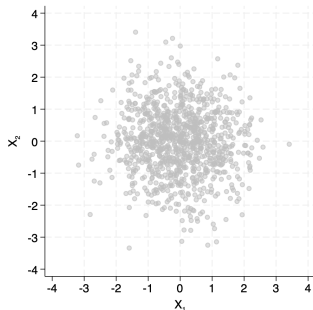
- The GitHub repository contains the Stata package files.
- The main example scripts are `statabi1.do`, `autodata.do`, and the appendix script.
- Produce bivariate example, the auto data case and the appendix respectively.

Core Command Pattern

```
ballmapper x1 x2, epsilon(1.2) color(y1) layout ///  
repulsion(0.05) attraction(0.01) filename("xy1bm12")
```

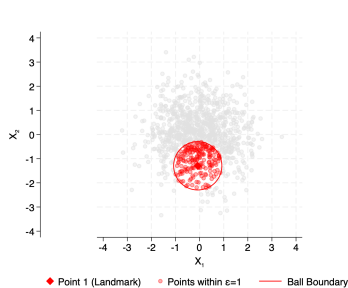
- Variables before the comma define the axis space X .
- `epsilon()` fixes the ball radius.
- `color()` selects the variable summarised inside balls.
- `filename()` writes the graph to a PNG file.

Bivariate Gaussian Example

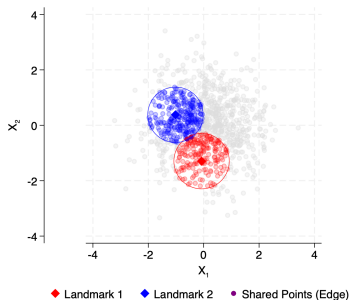


- $N = 1000$ observations.
- $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$.
- The dense centre and sparse tails provide a clean construction example.
- The plot can be understood directly because it is still bivariate.

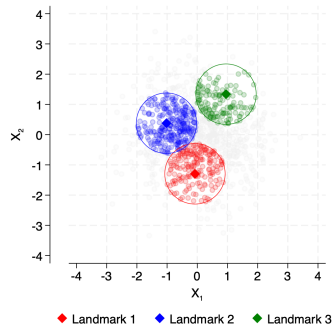
Drawing Balls and Finding Overlap



First ball



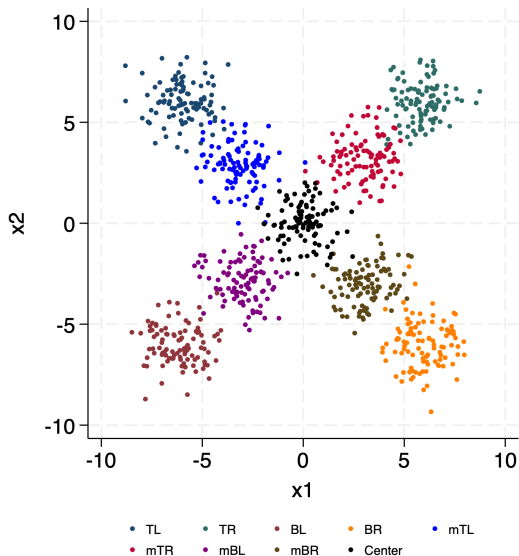
Overlap example



Next landmark

Overlap between balls becomes an edge in the TDABM graph.

Artificial X Dataset: Why Use It?

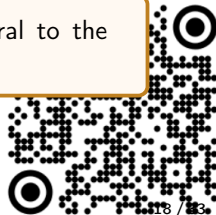


- Nine Gaussian sub-clusters are translated into an X shape.
- Summary statistics do not identify the shape.
- Local and global structure are both deliberately present.
- The example tests whether TDABM recovers known structure.

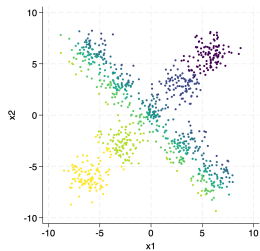
Five Colouring Variables

Outcome	Meaning	Pattern expected
$Y_1 = X_1 + X_2 + \theta$	linear	diagonal gradient
Y_2	group label	sub-cloud identity
$Y_3 = X_1^2 + X_2^2 + \theta$	quadratic	high at extremes
$Y_4 = \phi$	noise	no spatial pattern
$Y_5 = 1[0 < X_1 < 3, 0 < X_2 < 3]$	restricted range	local hot spot

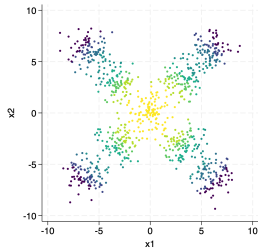
The same X can be coloured by many different Y variables. This is central to the inferential use of TDABM.



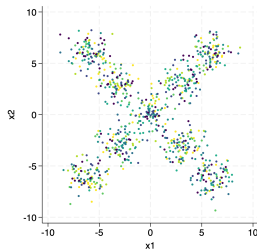
Scatterplots: Outcomes on the X Dataset



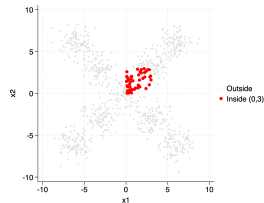
Linear



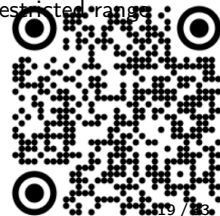
Quadratic



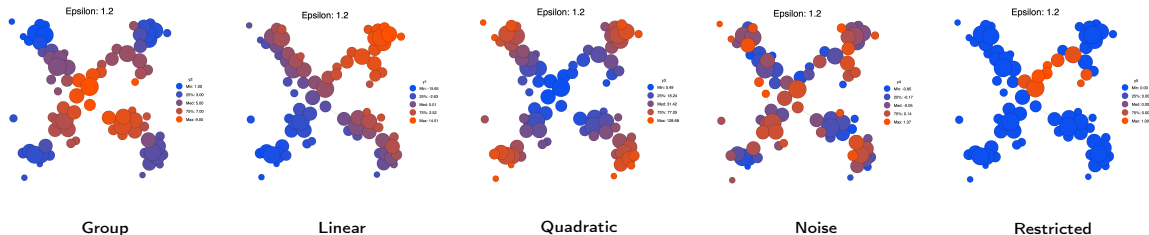
Noise



Restricted range

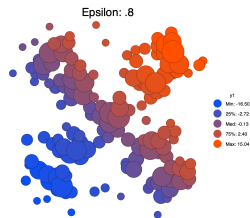


TDABM Recovers the Expected Patterns

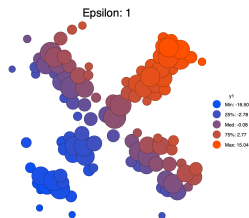


In the artificial example, the correct visual pattern is known before applying TDABM. The graph acts as a proof of concept.

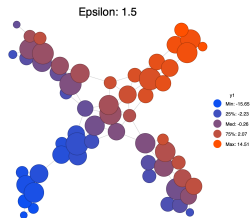
Role of Radius



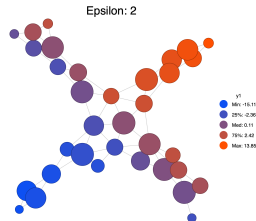
$$\epsilon = 0.8$$



$$\epsilon = 1.0$$



$$\epsilon = 1.5$$

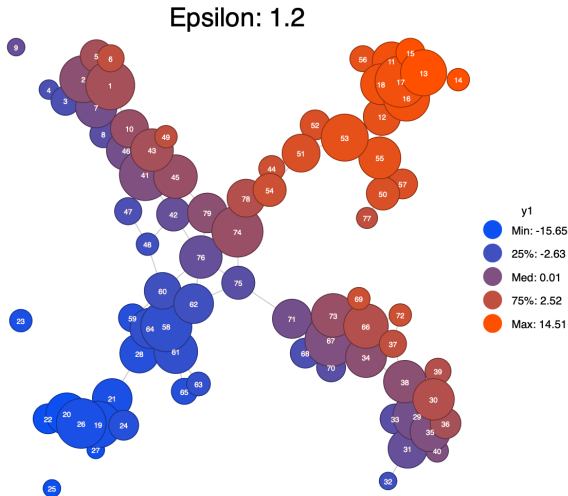


$$\epsilon = 2.0$$

- Smaller radii show local structure but create more disconnected balls.
- Larger radii reveal global structure but can merge local features.



Labels Turn the Graph into an Object of Analysis



- Labels allow ball-by-ball discussion.
- Ball membership links graph locations back to observations.
- Labels support explanation of extremes, bridges, and mixed balls.

What the Package Returns

BM_RESULTS

Ball number, plotting coordinates, size, colour bin, node rows and edge rows.

BM_MERGED

Original data expanded with ball membership, allowing post-plot analysis.

- The graph is the exploratory object.
- The frames allow the researcher to verify, summarise and narrate it.

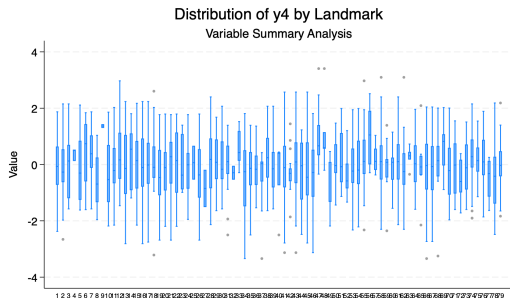
Summary Commands

```
ballsummary y1 y2 y3 y4 y5, csvfile("ymmeans12")
```

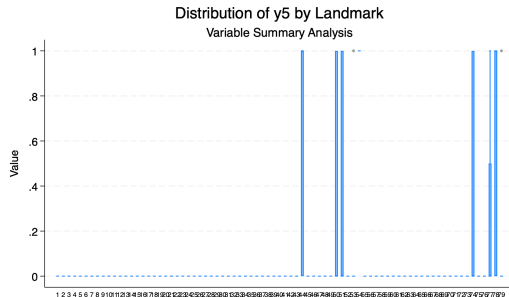
```
variablessummary y5, boxplot boxfile("y5_12_box") ///  
    csvfile("y5_12_stats")
```

- `ballsummary`: means of many variables by ball.
- `variablessummary`: detailed distribution of one variable by ball.

Within-Ball Variation



Noise variable



Restricted-range variable

Boxplots help diagnose whether a colouring variable varies systematically across the TD-ABM graph.

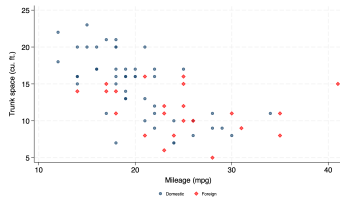
Auto Dataset Example

- Uses Stata's built-in 1978 car dataset.
- Axis variables include mpg, trunk, weight, length, turn, displacement and gear ratio.
- Axis variables are standardised before TDABM.
- Colouring variables: price and foreign manufacturer share.

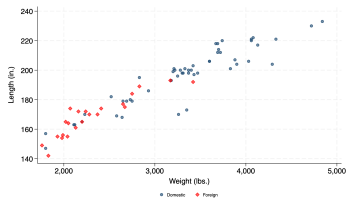
Substantive question

Do multivariate car characteristics separate domestic and foreign models, and do they help explain price variation?

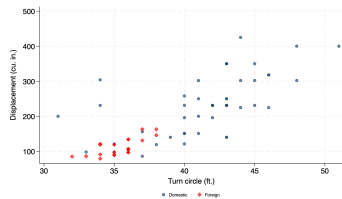
Pairwise Scatterplots Motivate the Multivariate View



MPG and trunk

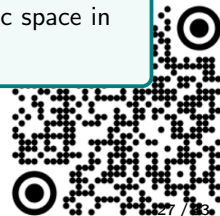


Weight and length

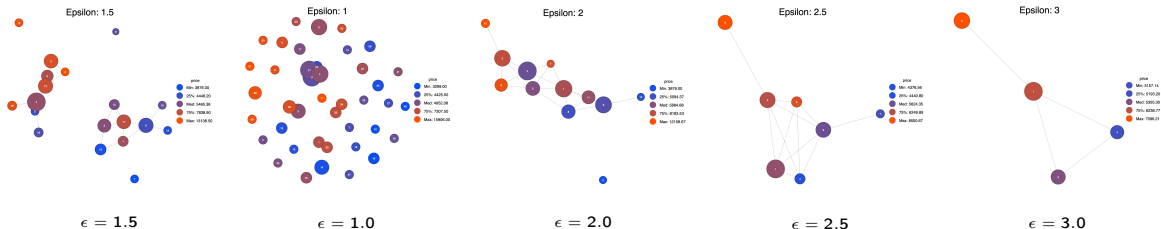


Turn and displacement

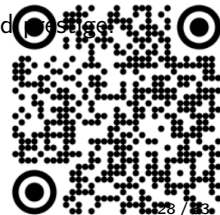
Pairwise views show associations, but they do not show the joint characteristic space in one object.



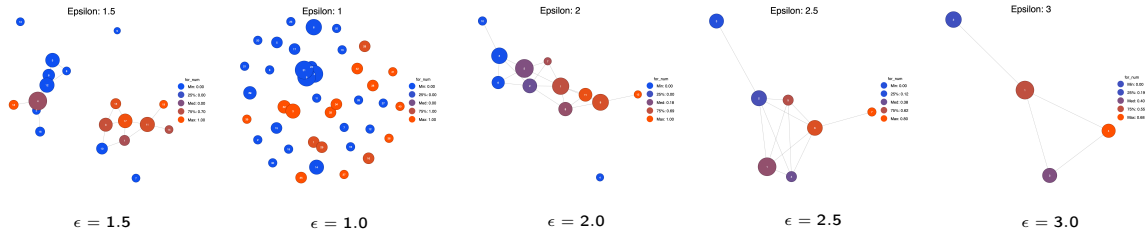
Auto Data: Price Colouring Across Radii



- Connected components merge as radius increases.
- Price varies within components, suggesting omitted dimensions such as brand prestige.

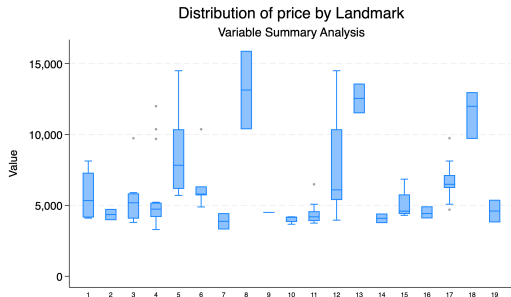


Auto Data: Foreign Manufacturer Share



Foreign and domestic cars are not perfectly separated, but the multivariate characteristic space makes the segregation easy to inspect.

Interpreting Auto Data Balls



Price variation within balls

- Some balls are almost entirely domestic or foreign.
- Mixed balls identify cars with similar characteristics across manufacturer groups.
- Price variation inside a ball flags variables not included in the graph.

Example Workflow

- ❶ Install `ballmapper` from GitHub.
- ❷ Run `statabi1.do` to reproduce the artificial X example.
- ❸ Show how changing `epsilon()` changes the graph.
- ❹ Run `autodata.do` to move from teaching data to built-in Stata data.
- ❺ Open `BM_RESULTS` and `BM_MERGED` after a labelled run.

Limitations

- TDABM is exploratory unless linked to a formal estimand or modelling task.
- Variables must be scaled and complete for distance calculations.
- Radius selection is a research judgement; multi-radius inspection is essential.
- The graph layout is a display object; horizontal and vertical directions are not substantive axes.
- The Stata package provides common workflows, while large-scale perturbation or radius sweeps may need Python or R.

Takeaway

- TDABM maps the shape of multivariate data through equal-radius balls.
- `ballmapper` brings the method into a familiar Stata workflow.
- Artificial data confirm that expected structures are recovered.
- Auto data show how the method supports substantive interpretation.
- The key analytical discipline is to treat the graph, the ball membership table and radius sensitivity together.

References

- Rudkin, S. and Rudkin, W. (2026). *ballmapper: Applying Topological Data Analysis Ball Mapper in Stata*. arXiv:2601.00508.
- Dlotko, P. (2019). Ball mapper: A shape summary for topological data analysis. arXiv:1901.07410.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*.
- Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs. CHI.
- GitHub repository: <https://github.com/srudkin12/statabm>