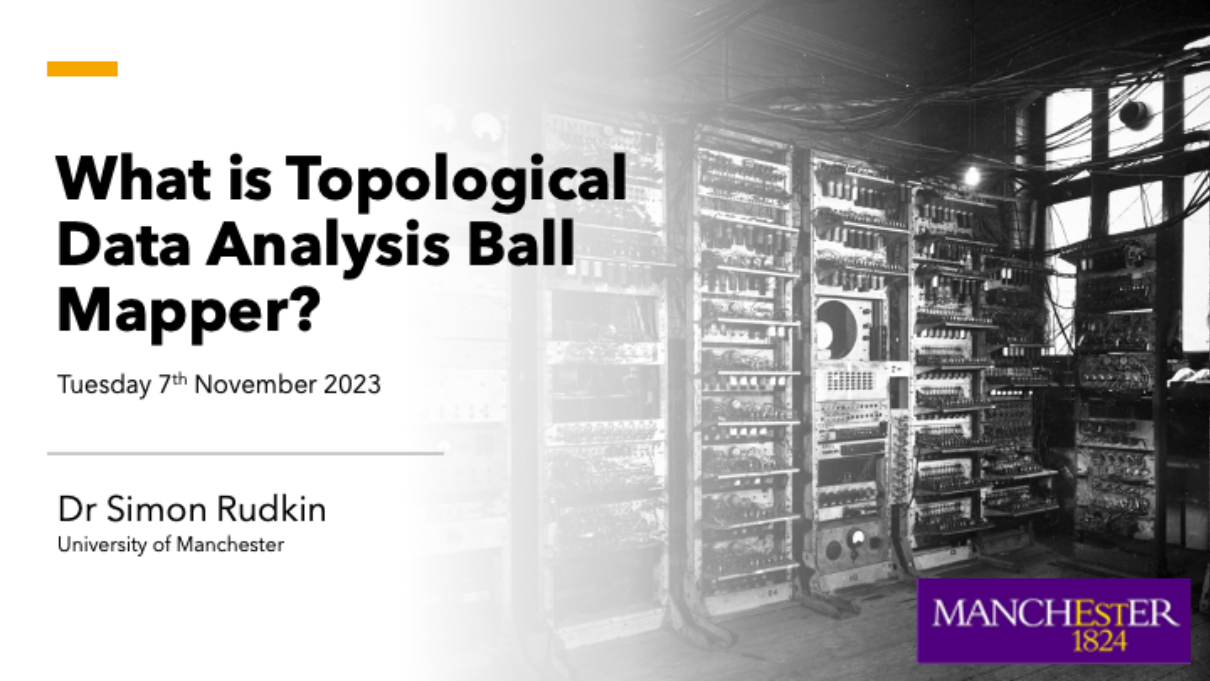
What is Topological Data Analysis Ball Mapper?

Tuesday 7th November 2023

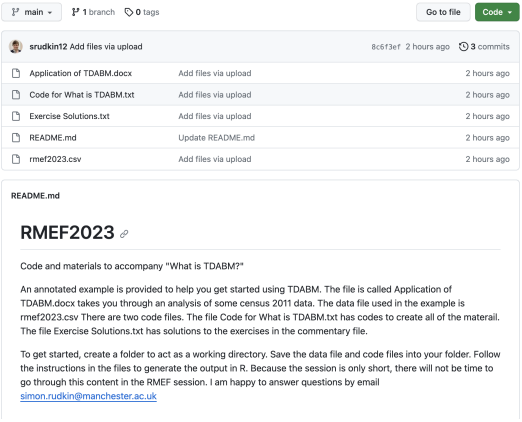
Dr Simon Rudkin
University of Manchester



In this Presentation...

- Motivation for TDABM
- Data properties and TDABM plots (Rudkin et al., 2023a)
- An Economic Topology of Brexit (Rudkin et al., 2023b)
- Regional Trajectories and Resilience (Rudkin and Webber, 2023)

GitHub and Code



GitHub site contains:

- Slides from this talk
- "Computer Lab" for applying TDABM to census data
- Code to recreate the computer lab
- Solutions code
- Data for the example

Key Motivating Arguments

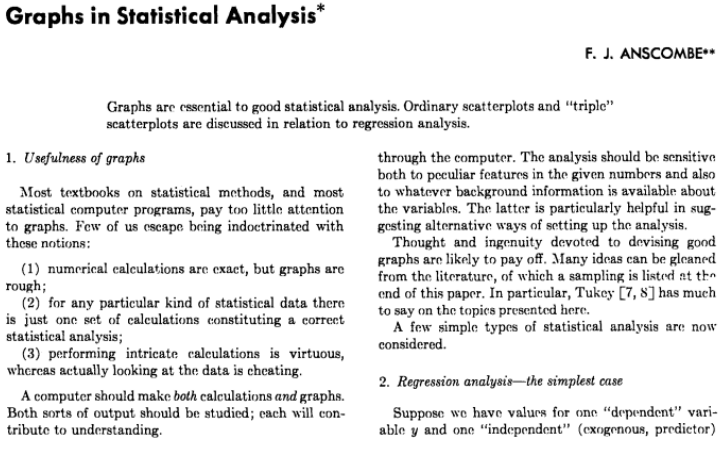
Visualising data is an essential phase of the modelling process (Anscombe, 1973)

Humans cannot see in multiple dimensions - dimension reduction

Mapping helps rationalise space - look to map our data

Summary statistics (1st and 2nd moments) are insufficient (Matejka and Fitzmaurice, 2017)

Motivation



Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;
(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

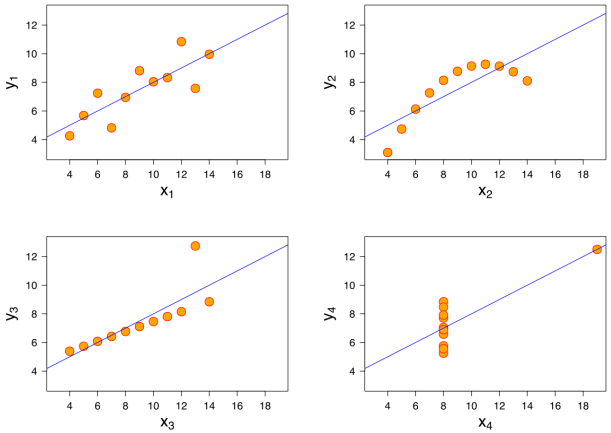
A few simple types of statistical analysis are now considered.

2. *Regression analysis—the simplest case*

Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor)

Seminal work from Anscombe (1973). Datasets with identical means, standard deviations and correlations.

Motivation 2



- Plot 1: Noisy data on line
- Plot 2: Quadratic relationship
- Plot 3: Outlier with high Y value
- Plot 4: High leverage point

Regression line has identical coefficients and R-squared:

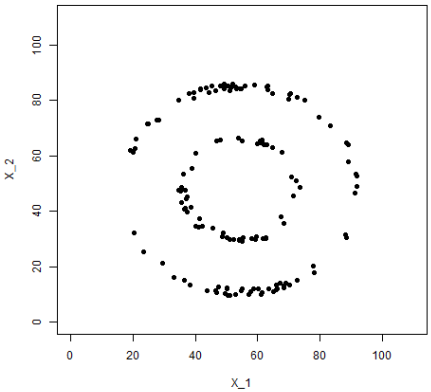
$$Y = \alpha + \beta X + \epsilon \qquad (1)$$

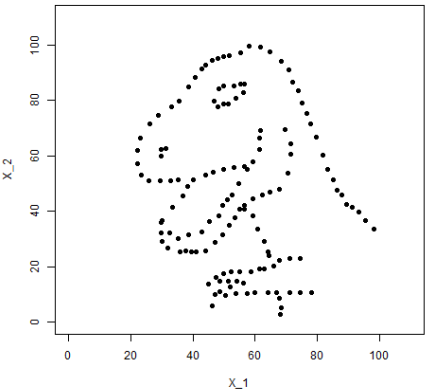
Importance of Topology

Sketch a scatterplot of two variables X and Y with the following information:

- The mean of X is 54 and mean of Y is 47.
- The standard deviation of X is 17 and the standard deviation of Y is 27.
- The correlation between X and Y is -0.06

Importance of Topology 2

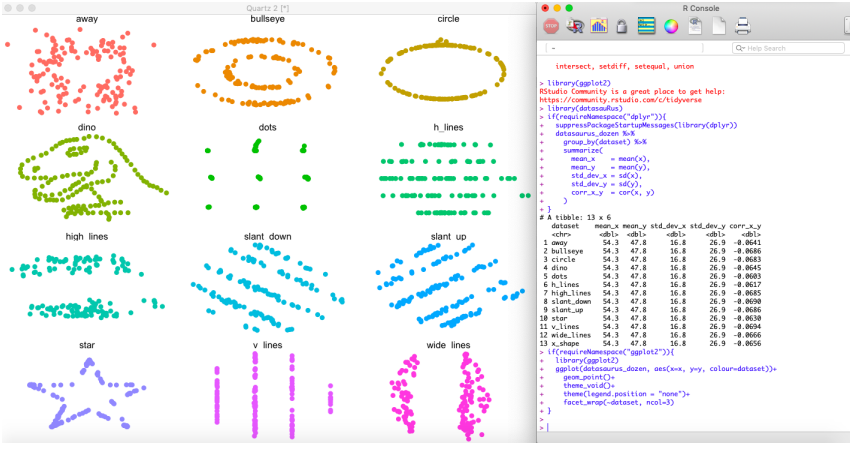




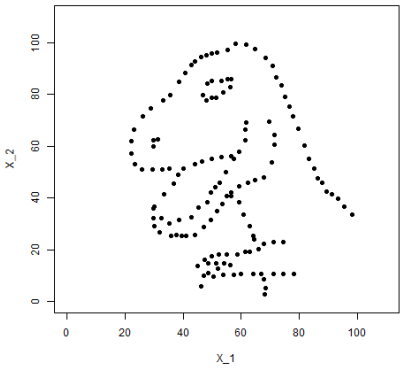
- Both datasets have correlation between horizontal and vertical of -0.06
- Neither are the plot you would expect
- Examples are from Matejka and Fitzmaurice (2017)

Importance of Topology 3

- See Matejka and Fitzmaurice (2017)
- Our eyes tell us these are not the same dataset even though the summary statistics are identical



Point Clouds



Point Cloud - Set of points plotted with 1 axis per dimension

- Simplest point cloud is scatter plot
- Essential to see data - Examples all have same first and second moments for both axes

Image shows datasaurus example from Matejka and Fitzmaurice (2017) - See R package `datasauRus` and yesterday's lab

Topological Data Analysis (TDA)

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 46, Number 2, April 2009, Pages 255–308
S 0273-0979(09)01249-X
Article electronically published on January 29, 2009

TOPOLOGY AND DATA

GUNNAR CARLSSON

1. Introduction

An important feature of modern science and engineering is that data of various kinds is being produced at an unprecedented rate. This is so in part because of new experimental methods, and in part because of the increase in the availability of high powered computing technology. It is also clear that the *nature* of the data we are obtaining is significantly different. For example, it is now often the case that we are given data in the form of very long vectors, where all but a few of the coordinates turn out to be irrelevant to the questions of interest, and further that we don't necessarily know which coordinates are the interesting ones. A related fact is that the data is often very high-dimensional, which severely restricts our ability to visualize it. The data obtained is also often much noisier than in the

Data has Shape and Shape has Meaning

- Seminal work Carlsson (2009)
- Mapper algorithms after Singh et al. (2007) - Stability?
- Potential fraud and money laundering (Singh and Best, 2016, 2019; Chang and Luo, 2019; Lokanan, 2022).

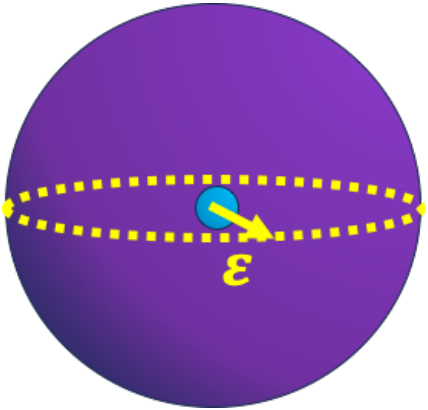
Core Elements of TDABM

Axis Variables - Must be continuous with sufficiently many observations. Would be plotted using a scatterplot

Outcome Variable - Must be available for each data point

Colouration Function - How should outcome values for each point be combined within the ball? (default is geometric mean)

Topological Data Analysis Ball Mapper (Dłotko, 2019)

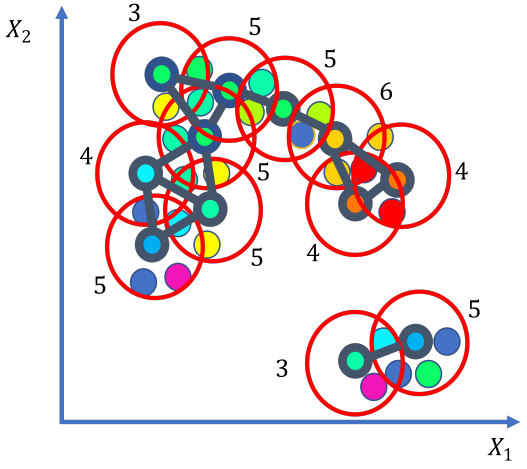


- TDABM "covers" data with "balls" of fixed radius ϵ (single parameter)
- TDABM has advantage of stability over mapper (Dłotko, 2019; Carriere and Oudot, 2018)
- Applied in finance by (Qiu et al., 2020), Dłotko et al. (2021)
- Benefits of visualisation to understand Brexit Rudkin et al. (2023)
- Capturing trajectories in regional development time series Rudkin et al (2023)
- Points in ball are "similar" in all dimensions

TDABM Setup

- Dataset X - N observations on K variables, x_1, x_2, ..., x_k, ..., x_K
- Point p_i, $i \in [1, N]$ has co-ordinates x_{1i}, x_{2i}, ..., x_{ki}, ..., x_{Ki}
- Cover $B(X)$ - p_i is assigned to at least one ball $B_b(l_{kb}, \epsilon)$.
- l_k is co-ordinate of point at centre of ball on axis k
- l iteratively selected randomly from uncovered points
- ϵ - ball radius - only choice parameter
- V shown as discs to represent balls - E connect pairs where $B(p_1, \epsilon) \cap B(p_2, \epsilon) \neq \emptyset$
- Produce abstract two-dimensional visualization of $B(X, \epsilon)$, $G(V, E)$

TDA Ball Mapper Overview



- Cover with balls of fixed radius
- Colour is a function on the data - colour each point according to outcome or characteristic of interest
- Ball colour is function on the members - colour according to average outcome of members, standard deviation of outcomes etc.
- Edges show points in overlap of balls
- Resize balls according to number of points contained within - indicative of density of space

Artificial Data Sets

	Noise Cloud ($N = 500,\ k = 5$)	Five-Part Cloud ($N = 500,\ k = 5$)
Step 1	Draw 500 x_{ik} from $N \sim (0, 1)$	
Step 2		Add 2 to each x_{ik} for $i \in [101, 200]$ - repeat for groups of 100 adding 4, 6 and 8
Step 3	Construct outcome variable $Y_k = \sum_{i=1}^{N} x_{ik}$	
	No clustering	5 distinct subclouds

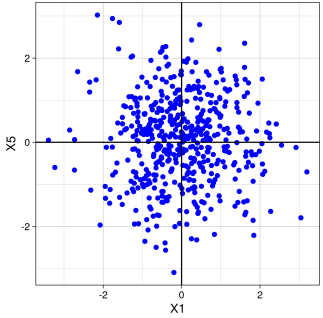
Artificial Data Summary Statistics

		Mean	sd	Min	q25	q50	q75	Max	Skew	Kurtosis
Noise	X_1	0.002	1.035	-3.396	-0.672	-0.021	0.661	3.196	0.046	3.243
	X_2	-0.055	0.959	-2.907	-0.673	-0.063	0.589	2.706	-0.084	3.180
	X_3	0.032	0.938	-2.930	-0.592	0.067	0.670	2.691	-0.176	2.960
	X_4	-0.003	1.024	-3.122	-0.670	-0.052	0.680	3.168	0.022	2.867
	X_5	0.051	1.042	-3.095	-0.664	0.102	0.750	3.022	-0.028	2.868
FP	X_1	3.978	2.980	-3.206	1.465	4.099	6.310	11.34	-0.028	2.040
	X_2	4.030	2.959	-2.455	1.529	4.047	6.338	11.42	0.064	2.022
	X_3	3.910	3.017	-2.526	1.433	3.912	6.292	10.17	-0.034	1.944
	X_4	3.941	2.989	-2.200	1.361	3.995	6.474	10.96	0.019	1.923
	X_5	3.988	3.030	-2.971	1.482	3.944	6.571	10.45	-0.046	1.988

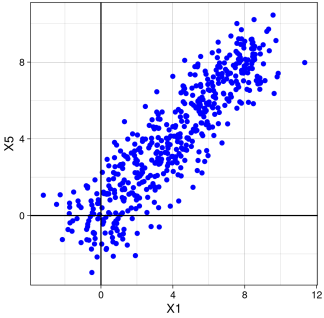
Artificial Data Correlations

Panel (a): Noise Cloud						Panel (b): Five-Part Cloud				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
X_1	1	0.082	0.082	0.004	0.014	1	0.894	0.894	0.893	0.899
X_2	0.082	1	-0.041	0.037	0.010	0.886	1	0.901	0.896	0.911
X_3	0.081	-0.027	1	-0.005	-0.011	0.886	0.889	1	0.896	0.905
X_4	-0.005	0.032	-0.009	1	0.018	0.885	0.887	0.890	1	0.899
X_5	-0.011	0.030	0.006	0.026	1	0.888	0.899	0.898	0.890	1

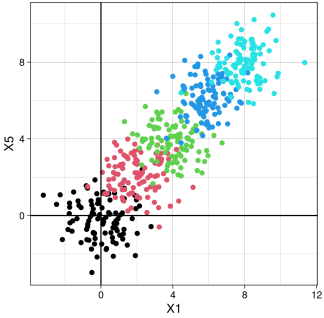
Artificial Point Clouds



(a) Noise Cloud

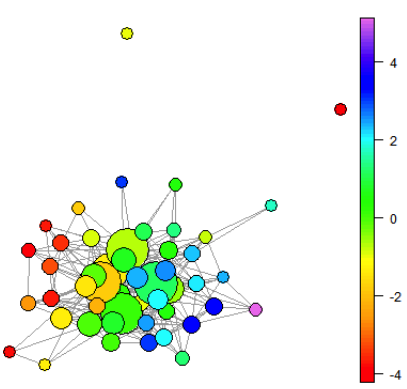


(b) Five-Part Cloud

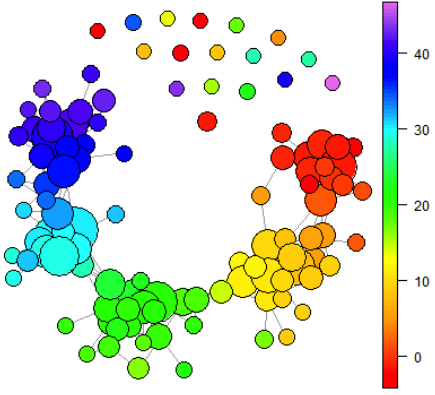


(c) Groups

Example TDABM Plots

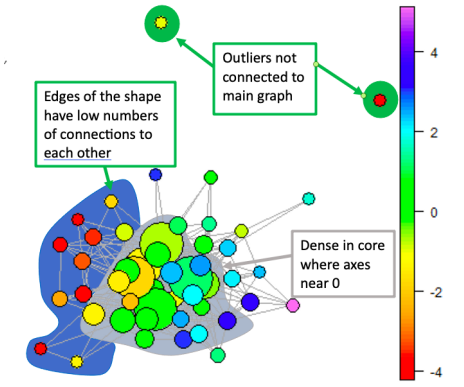


(a) Noise Cloud

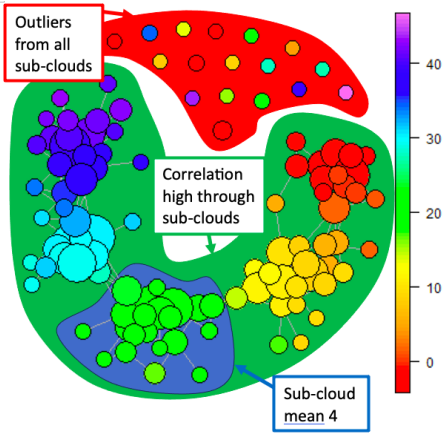


(b) Five-Part Cloud

Example TDABM Plots

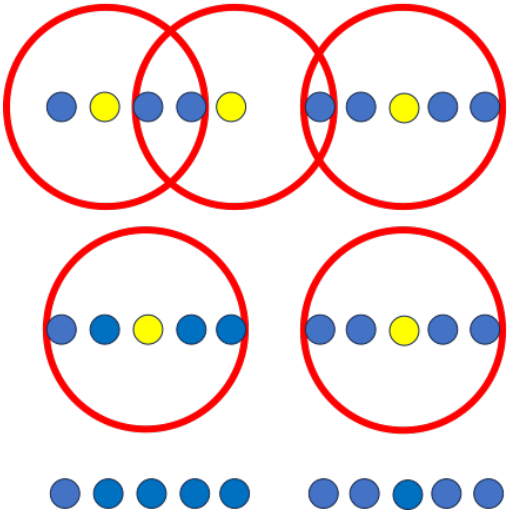


(a) Noise Cloud



(b) Five-Part Cloud

Repetitions and Bootstrapping



- 10 data points in 1 dimension
- Split into two groups
- Landmarks are yellow, balls red outline
- Top example three balls with connection between groups
- Second example two balls with no connection
- Repetition makes sure we have TDABM diagrams with both possibilities

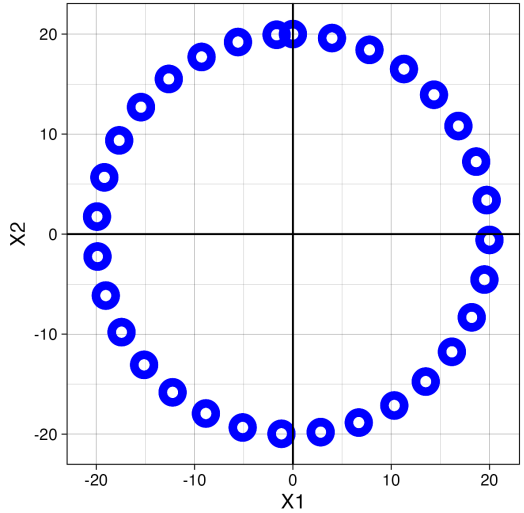
Is there an Optimum Radius?

Data Driven: Identify the radius which optimises an objective function (e.g out of sample prediction of credit risk)

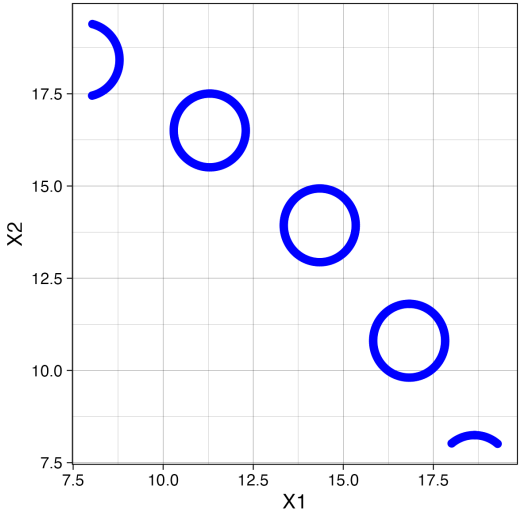
Small Radius: Capture detail within the data and isolate special cases

Large Radius: Obtain an overall picture of the data and connectivity betwene regions of data space

Optimum Radius? Example Datasets

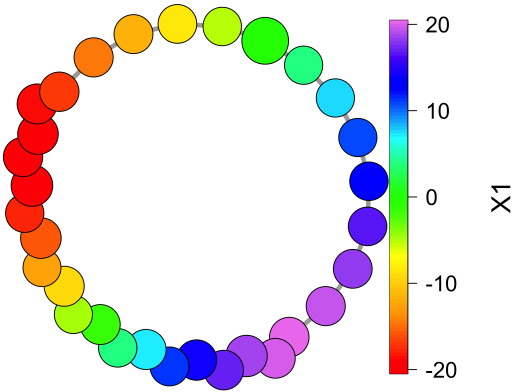


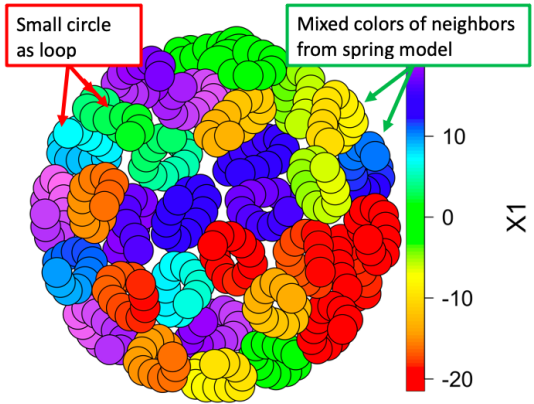
(a) Full Dataset



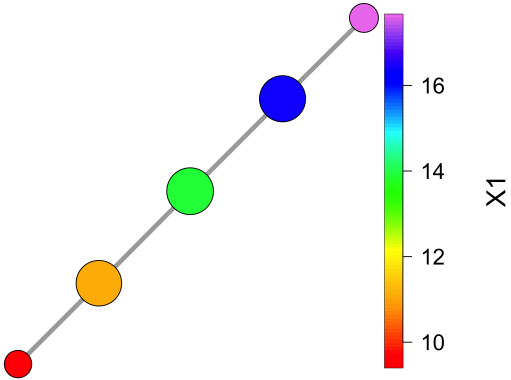
(b) Zoomed Section

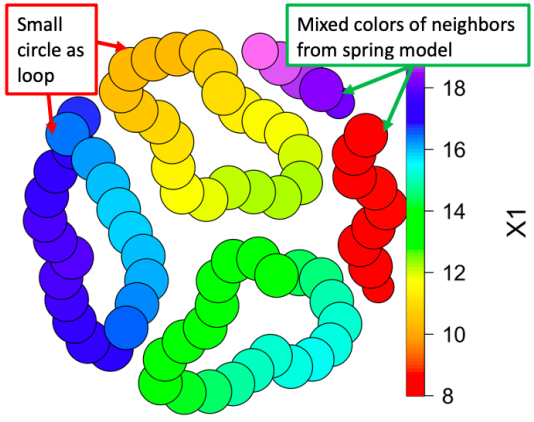
Optimum Radius ? (Full Dataset)





Optimum Radius ? (Zoomed Section)

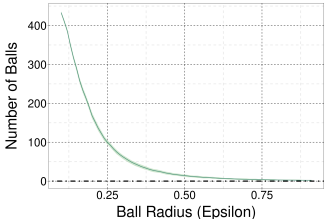




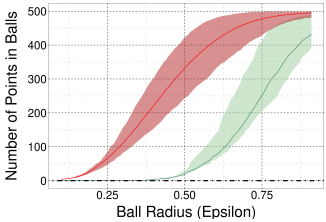
Measures of TDABM Graphs

Colouration	Minimum	h^{Min}	Within ball range	Average	Δh^{Avg}
	Maximum	h^{Max}		s.d.	Δh^{sd}
	s.d	h^{sd}		Minimum	Δh^{Min}
	Range $(h^{Max} - h^{Min})$	h^{Range}		Maximum	Δh^{Min}
Ball Size	Minimum Size	S^{Min}	Within ball s.d.	Average	hsd^{Avg}
	Maximum Size	S^{Max}		Minimum	hsd^{Min}
	Range $(S^{Max} - S^{Min})$	S^{Range}		Maximum	hsd^{Max}
Connections	Zero	N^{Z}	Number of Balls		N^{B}
	Average	N_{Con}^{Avg}			
	Total	N^{E}			

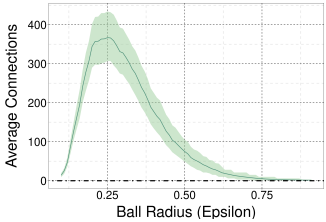
Role of Radius: Noise Cloud (Normalised on $[0, 1]$)



(a) N^B



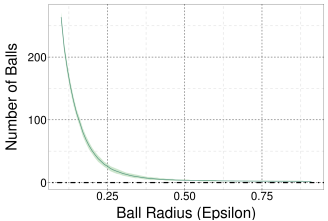
(b) S^{Min} and S^{Max}



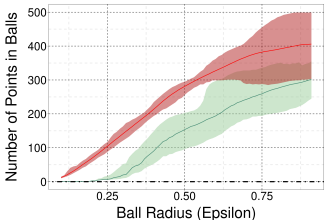
(c) N^E

- Estimate 10000 repetitions for ϵ in $[0.10, 0.90, 4]$ intervals 0.01

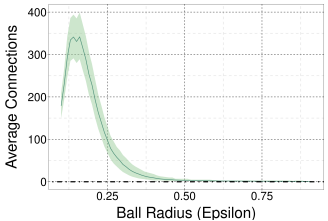
Role of Radius: Five-Part Cloud (Normalised on $[0, 1]$)



(a) N^B



(b) S^{Min} and S^{Max}



(c) N^E

- Estimate 10000 repetitions for ϵ in $[0.10, 0.90]$ intervals 0.01

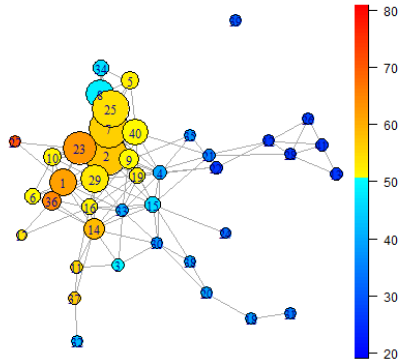
An Economic Topology of Brexit

- Joint work with Dr Pawel Dlotko (Discouri Centre in Topological Data Analysis), Dr Wanling Qiu (University of Exeter) and Dr Lucy Barros (Swansea University)
- Brexit referendum modelled in many ways - how did polls get it wrong?
- Subsequent discussion of the "red wall" and the subsequent collapse
- Create a dataset of constituencies - Publically available data from Census 2011
- Axes include home ownership, marital status, social classification, car ownership, education, health and deprivation

An Economic Topology of Brexit

- Housing Tenure (Owned, Social Rent, Private Rent, Other)
- Household Composition (Married, Cohabit, Other)
- Car Ownership (0 Cars, 1 Car, 2+ Cars)
- NSSEC (4 Levels)
- Qualifications (Less than 5 GCSES, 5-9 GCSES, A-Levels and Higher Education)
- Self-reported Health (Very Good, Good, Low)
- Deprivation (0, 1, 2+)
- Age (0-18, 18-24, 25-59, 60+)

An Economic Topology of Brexit



- Leave percentage by constituency
- Concentration of Leave versus Remain
- Labour majorities link to Remain
- Conservatives in the Leave
- Gains of conservatives in 2017 vs 2015
- More “red wall” falling in 2019 vs 2015

Regional Trajectories

Regional Growth Paths and Regional Resilience

52 Pages • Posted: 24 Jan 2023

Simon Rudkin
The University of Manchester - Social Statistics Department

Don J Webber
Sheffield University Management School; The University of Sheffield

Date Written: January 20, 2023

Abstract

This article explores whether regions following common development paths experience similar levels of resilience when faced with a shock. Analyses of three different indicators of resilience (value added, employment, productivity) across UK local authority districts between 1980 and 2015 reveal that regions following a common evolutionary path did not respond homogeneously to the global financial crisis, and they experienced different levels of resilience across resilience indicators. There does not appear to be a common development path which guarantees resilience to a future shock. Understanding connections between conflicting indicators of resilience is paramount to inform policies to enhance resilience.

Keywords: Resilience; Evolutionary Economic Geography; GVA; Employment; Productivity; Topological Data Analysis

JEL Classification: R11, R12, R23, 049

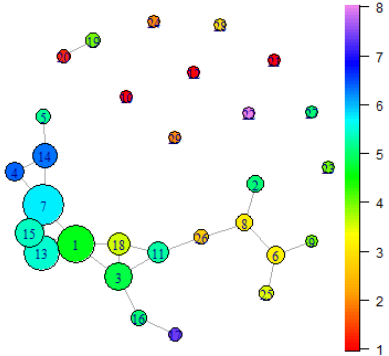
Working paper is available on SSRN

Regional Trajectories Data

- Measure resilience in Gross Value Added (GVA), Employment (EMP) and Productivity (PRO)
- Data from 1980 to 2006 inclusive
- NUTS3 level for United Kingdom

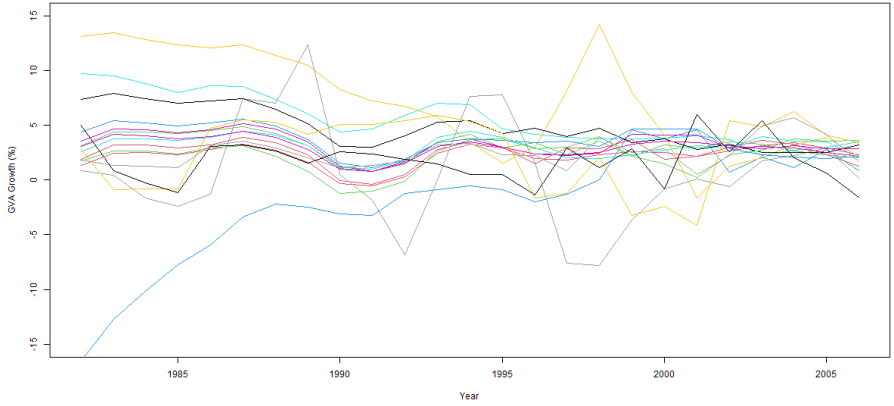
Year at 2006	2009	2010	2011	2012	2013	2014	2015	?
Resilience	1	2	3	4	5	6	7	8

GVA Trajectories



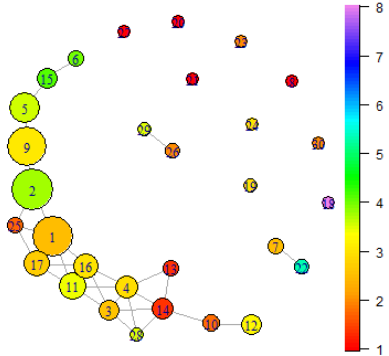
- Axes constructed from GVA levels each year from 1980 to 2006
- Colouration is number of years after 2008 to return to 2006 levels
- A lot of "average"
- What lies behind? What is challenging the assertion that trajectory determines resilience?

GVA Trajectories 2



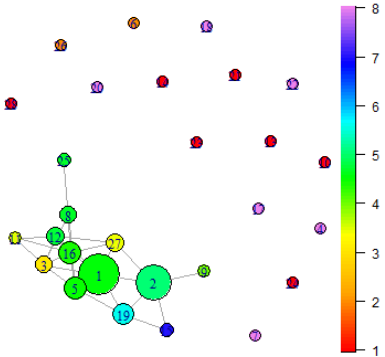
Trajectories of regions from 27 year of data (1980-2006 inclusive)

EMP Trajectories



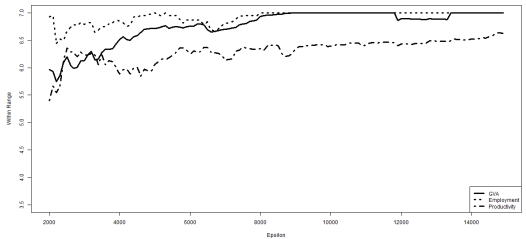
- Axes constructed from EMP levels each year from 1980 to 2006
- Ball 1 contains thirty regions that followed similar trajectories between 1980 and 2006
- Sixteen regions recovering in 1 year...
- 3 regions (Somerset, Walsall, and South Nottinghamshire) taking 8 years or more to recover

PRO Trajectories



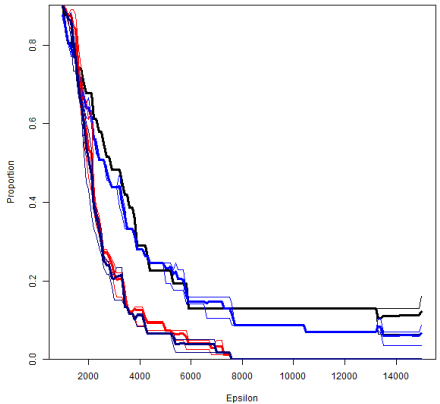
- Axes constructed from PRO levels each year from 1980 to 2006
- Balls 6 (West Cumbria), 26 (Torbay), 28 (Perth, Kinross, and Stirling), 10 (Chorley and West Lancashire), 24 (Isle of Wight) are outliers.
- Least resilient: 18 (Wolverhampton), 20 (Hackney and Newham), 22 (West Sussex – South West), 17 (Sandwell), 4 (Durham County Council), and 7 (Greater Manchester South East).

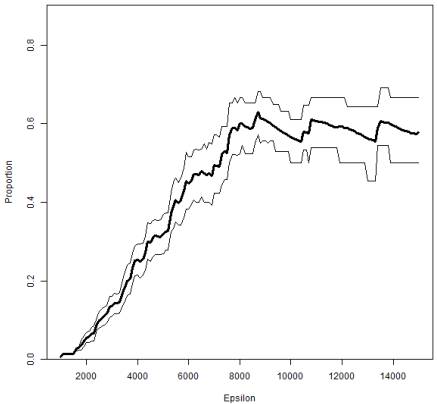
Trajectories to Resilience?



- Axes constructed from GVA levels each year from 1980 to 2006
- Resilience measured for GVA, employment and GVA per worker
- Plot average range of resiliencies for balls with 5+ regions
- Min is 1, Max is 8 so ranges go to 7
- Consistent regardless for all three axis variables

Further Analysis: GVA

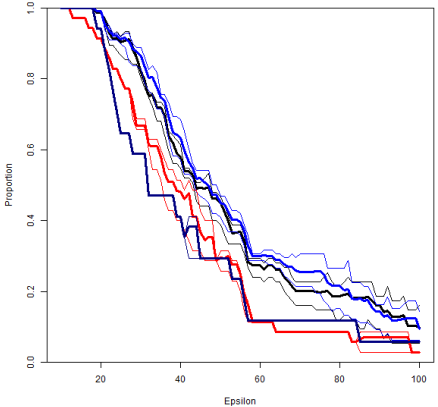


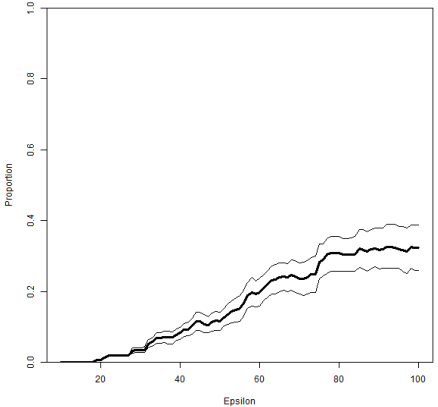


Left panel is 1 (Black), 2 (Blue), 6 (Red), 8 (Navy Blue)

Right panel proportion of balls with 1 and 8

Further Analysis: EMP





Left panel is 1 (Black), 2 (Blue), 6 (Red), 8 (Navy Blue)

Right panel proportion of balls with 1 and 8

Key Motivating Arguments

Visualising data is an essential phase of the modelling process (Anscombe, 1973)

Humans cannot see in multiple dimensions - dimension reduction

Mapping helps rationalise space - look to map our data

Summary statistics (1st and 2nd moments) are insufficient (Matejka and Fitzmaurice, 2017)

Summary

- Topological Data Analysis Ball Mapper offers new chance to see data
- Inference from the BM plots and summaries of the balls
- Each ball represents data - can take more characteristics from that data to discuss
- Interested in whole picture - BM shows the picture and enables understanding social phenomena and policy decisions
- Moving into analysis of modelling and model evaluation
- Build from the simple premise as in Anscombe's quartet

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

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Carriere, M. and Oudot, S. (2018). Structure and stability of the one-dimensional mapper. *Foundations of Computational Mathematics*, 18(6):1333–1396.

Chang, C. J. and Luo, Y. (2019). Data visualization and cognitive biases in audits. *Managerial Auditing Journal*.

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

Dłotko, P., Qiu, W., and Rudkin, S. (2021). Financial ratios and stock returns reappraised through a topological data analysis lens. *The European Journal of Finance*, pages 1–25.

Lokanan, M. E. (2022). Financial fraud detection: the use of visualization techniques in credit card fraud and money laundering domains. *Journal of Money Laundering Control*, (ahead-of-print).

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.

Qiu, W., Rudkin, S., and Dłotko, P. (2020). Refining understanding of corporate failure through a topological data analysis mapping of altman's z-score model. *Expert Systems with Applications*, page 113475.

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. *SPBG*, 91:100.

Singh, K. and Best, P. (2016). Interactive visual analysis of anomalous accounts payable transactions in sap enterprise systems. *Managerial Auditing Journal*, 31(1):35–63.

Singh, K. and Best, P. (2019). Anti-money laundering: using data visualization to identify suspicious activity. *International Journal of Accounting Information Systems*, 34:100418.