

Lost Villages, Bank Failure and Drugs: Multidimensional Scaling to the rescue

A talk by Professor Tony Coxon

AQMeN

14 March 2011

BACKGROUND & SOURCES

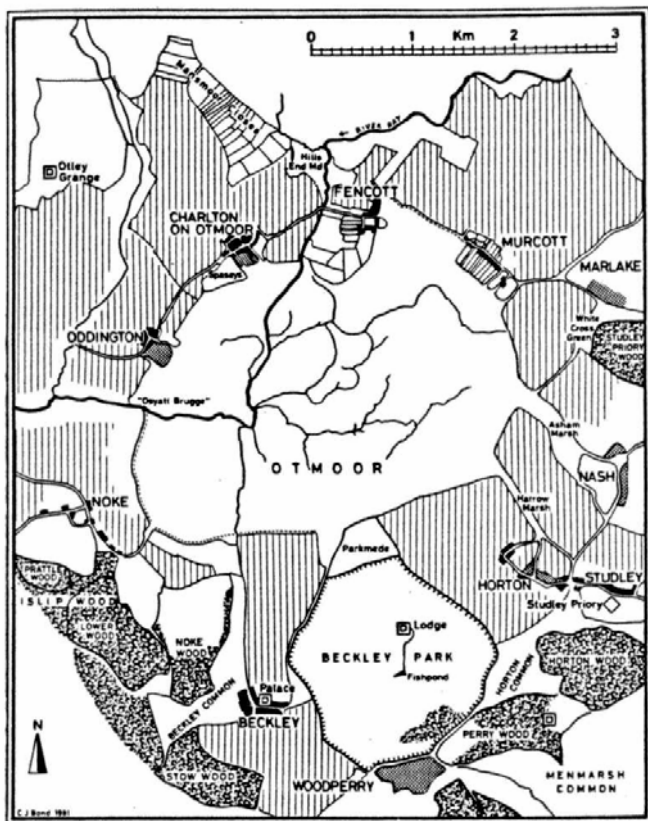
- D. Kendall, "Construction Of Maps From Odd Bits Of Information," *Nature*, No.231 (1971) pp 158-9
 - David Kendall FRS "Maps from Marriages: an application of non-metric MDS to parish registers" *, which used:
 - Bob Hiorns' material from a historico-genealogical study near Oxford
- The Kendall-Hiorns study relates to a set of eight parishes in the Otmoor region, close to Oxford .
- These are known locations, but can their positioning be reconstructed (using non-metric MDS) from centuries-old but available demographic data?

* *In* Hodson & Tautu Mathematics in the Historical and Archaeological Sciences (Anglo-Romanian conference at Edinburgh 1971), Edinburgh: University Press.

Lost Villages: KENDALL-HIORN'S OTMOOR PARISHES INTERMARRIAGE DATA

- Under what conditions – if any -- is it possible to [re] construct an “accurate” map from highly fallible , deficient, derivative data?
 - i.e. to recover and discover “lost” locations
- What methodological aspects of MDS are involved in making a good analysis?
- And .. to illustrate application of MDS to historical /archaeological & geographical data.

Otmoor itself

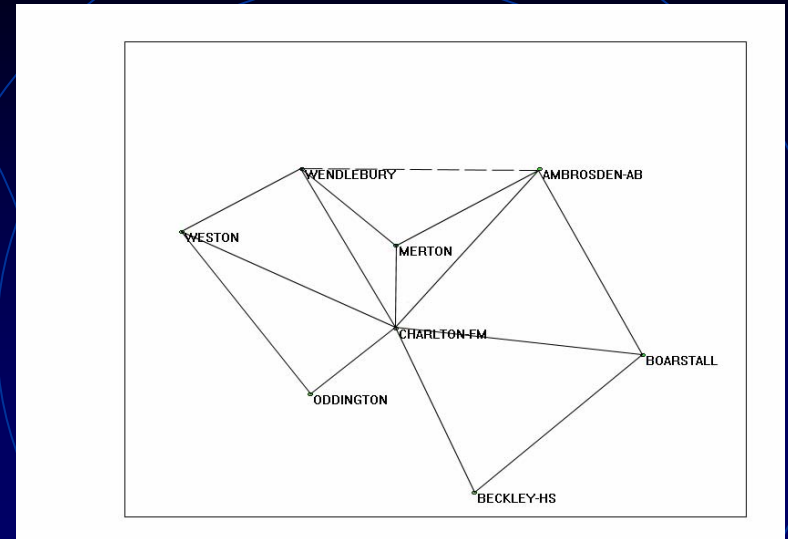


OTMOOR AT THE END OF THE MIDDLE AGES

-  Medieval villages
-  Moated site
-  Approximate extent of late medieval open-field cultivation
-  Areas of settlement desertion or contraction
-  Monastic house
-  Woodland
-  Lodge
-  Park pale



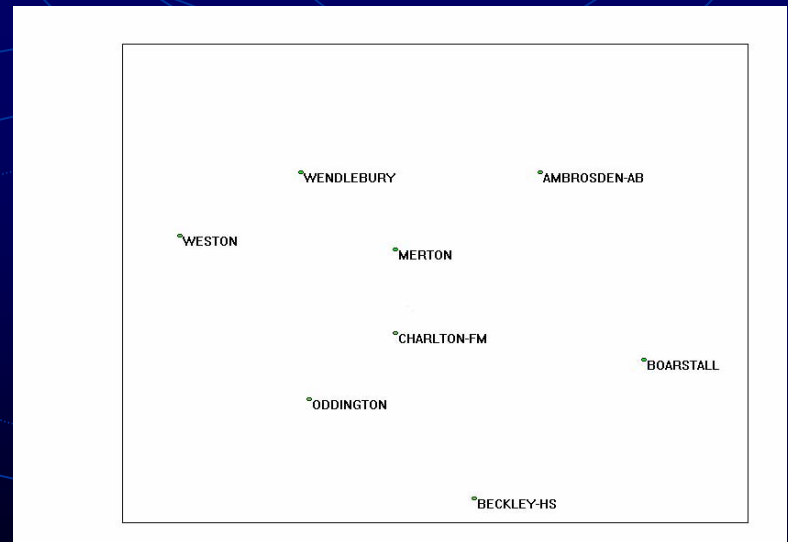
MAPS:



Top left: Map of the 8 parishes and their boundaries and villages

Top right: "Contiguity map" of the parishes

Opposite: Best MDS map from inter-marriage data



“If we could construct a standardized intermarriage rate between parishes for some period of time prior to the large-scale mobility characterizing the current world , we might hope to use it as a measure of similarity from which the map of the parishes could be constructed. “ (Kendall)

Basic Data:

a count made from Parish Records between 1600-1800 of

- *how often a spouse from one parish married a spouse from that or another parish.*
- This forms inter-marriage matrix **M**.

Basic Data Matrix:

	BH S	CF M	O	M	WS	WN	AAB	B	OW	Total
BH S	303	10	0	0	0	0	4	6	104	427
CF M	6	297	8	8	1	3	10	3	109	445
O	3	20	132	1	3	2	7	0	32	139
M	0	12	1	82	3	2	7	0	32	139
WS	0	2	1	2	260	2	4	0	78	349
WN	0	6	2	4	3	98	5	0	89	207
AAB	3	5	4	5	0	4	432	3	168	624
B	3	0	0	1	2	0	4	115	65	190

TABLE 1: Inter-parish Marriage Frequencies, Otmoor data

BUT ... Asymmetry ...

- This is an asymmetric matrix and cannot be scaled as it stands by the distance model, which is symmetric.
- Various possibilities exist to deal with asymmetry:
 - **scale each** triangular (LT, UT) matrix separately (as in inflow-outflow in Social mobility)
 - **sum or average** the corresponding entries
 - **fit an asymmetric** model.

HOWEVER ...The characteristics of the data are deficient and fail several conventional pragmatic tests of suitability for scaling

- **Sufficient data are needed to constrain the solution. [DCR > 2]**
 - But Data compression ratio here < 2
- **The best-constrained : all values should be distinct , but**
 - In both UT & LT, this is not so:
 - only 9 or 10 distinct values vs 28 entries
 - Quarter of data have the value 0 (LT 25%, UT 28%)
 - Many of the data are tied – which raises starkly the question of how to treat ties in obtaining the solution.

Is non-metric MDS going to be robust enough to yield a stable and interpretable result in face of these shortcomings?

Well, yes actually ...!

- TRY

A: Treat off-diagonal entries as they stand as data dis/similarities $\delta_{j,k}$ (i.e. separate LT, and UT)

B: Define standardized intermarriage index [SMI] as $L(j,k)$.

C: Standardize entries to take account of parish population totals (diagonal intra-marriage) and within-parish rates.

$$L_{jk} = \frac{M_{jk}M_k}{M_jM_{jj}}$$

C: To make them symmetric, he defines 2 forms:

Additive (sum):

$$S^1_{jk} = L_{jk} + L_{kj}$$

Multiplicative (GM)

$$S^2_{jk} = \sqrt{L_{jk} L_{kj}}^{10}$$

To produce:

Data Table: S1 (LTM) S2 (UTM)

	BHS	CFM	O	M	WS	WN	AAB	B
BHS		0.026	0.000	0.000	0.000	0.000	0.010	0.023
CFM	0.054		0.064	0.063	0.005	0.025	0.020	0.000
O	0.020	0.168		0.010	0.009	0.018	0.008	0.000
M	0.000	0.160	0.020		0.017	0.032	0.031	0.000
WS	0.000	0.012	0.023	0.039		0.015	0.000	0.000
WN	0.000	0.058	0.035	0.063	0.032		0.022	0.000
AAB	0.020	0.044	0.017	0.086	0.017	0.048		0.016
B	0.045	0.011	0.000	0.009	0.014	0.000	0.038	

S1 & S2 Measures:

- distributions of the S1 and S2 measure values yield much better behaved data than the original frequency data:
 - There are 21 or 18 (S1 & S2) distinct values vs the possible 28 entries – more than a doubling .
 - Considerably fewer of the S1 values data have the value 0, though S2 values include over a third (36%). So tied data and their treatment is still an issue.

Approaches to Tied Data

- **RESCALING TRANSFORM:** these data are presumed to be an (unknown) function of actual distance, so safer to choose more conservative **ordinal (monotonic) function** as the re-scaling transformation
- but use **two approaches to tied data:**
 - **S: Secondary (rigorist) approach to ties** (keeping equal data with equal fitting values) or penalising in Stress1 for their infraction):
 - **P: Primary (relaxed) approach to ties** (allowing tied data to be untied) without contributing to stress.

Logic of Data analysis ...

1. The data treated as proximity/distance-like quantities
2. The purpose of **non-metric MDS** is to ordinally re-scale these data so they become as close as possible to being distances of a describable map in two dimensions = “the solution”
3. Difference between the re-scaled data and the actual distances of the “solution” map (Stress) shows how far we from achieving a perfect scaling – zero stress means it is perfect.
4. This step is achieved using the **MINI-SSA** program in **NewMDSX**.

Logic of analysis , cont ...

4. In this unusual case, we actually know what the solution should be – the 2D centroid “target” configuration.
5. So we need to compare various solutions to see how close they are to it.
6. Any Euclidean Distance model configuration/ solution can be submitted to a similarity/affine transform:
 - [rigidly] rotating the pattern, flipping or reflecting the axes of the pattern, and uniformly changing the scaling /zoom
7. These are implemented using the **PROCRUSTES** program in NewMDSX.
 - Procrustes was an unpleasant inn-keeper outside Athens who had only one bed, and fitted his guests to them by either stretching them or lopping off their limbs until they were in close conformity to the bed.

Analysis of LT & UT data

UT & LT scaled twice, with both P&S Approaches. Stress1 (badness-of-fit) values:

	PS (relaxed)	SS (keep ties)
LT (sum)	.109	.162
UT(Geo. Mean)	.072	.158

- Note that:
 - LT Stress < UT Stress
 - PS Stress < SS [bound to be]
 - Differences in types of stress (PS vs SS) are more than the differences between measures (Add. vs GM)
- **The best-fitting solution is for UT-PS.**

Which is best ?

- In the normal way, this would be sufficient to decide that (UT-PS) solution is best (in terms of fit to the data), and this solution would count in Kruskal's famous (if dangerous) pragmatic evaluation figures as between "good" and "fair" fit!
- But the more important question for us is: ***Which configuration is closest to the "real" configuration?***

How well does each configuration fit the “Real Map”?

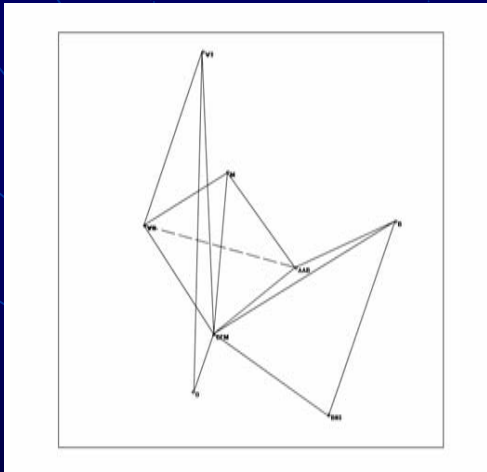
Use "Real Map" as Target Configuration and use Procrustes Rotation to put each solution into max. conformity with it. The Procrustes program (goodness-of-fit) S measure gives degree of fit.

	Primary Stress	Secondary Stress
Lower Triangle (Sum)	0.597	0.374
Upper Triangle (GM)	0.451	0.461

- The best-fitting configuration is LT-PS – but is not very good! In some ways UT-PS is better ...
- **BUT ... Primary stress is consistently giving better fitting results than Secondary Stress**

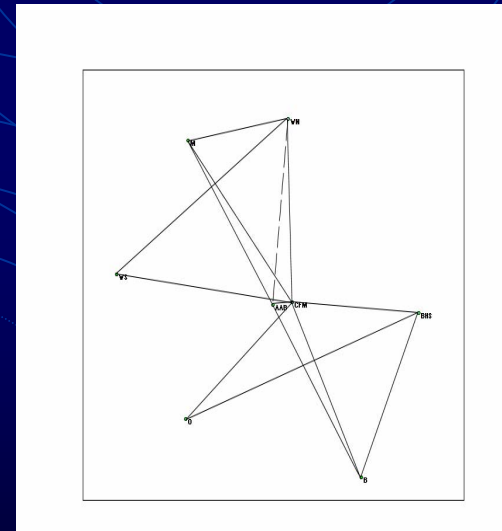
And, making an ellision ...

- At this point , Kendall goes on to consider the effect of standardizing the measures of similarity, and some other refinements, the detail of which we might find a little otiose ... and the discussion gets more complex . We shall abridge it until he reports ... The result?
- Suffice it to say that the *standardized version of the multiplicative measure in combination with Primary Stress* outperforms other alternatives, producing the RH contiguity graph:



SumPR

More planar, though
not best fit



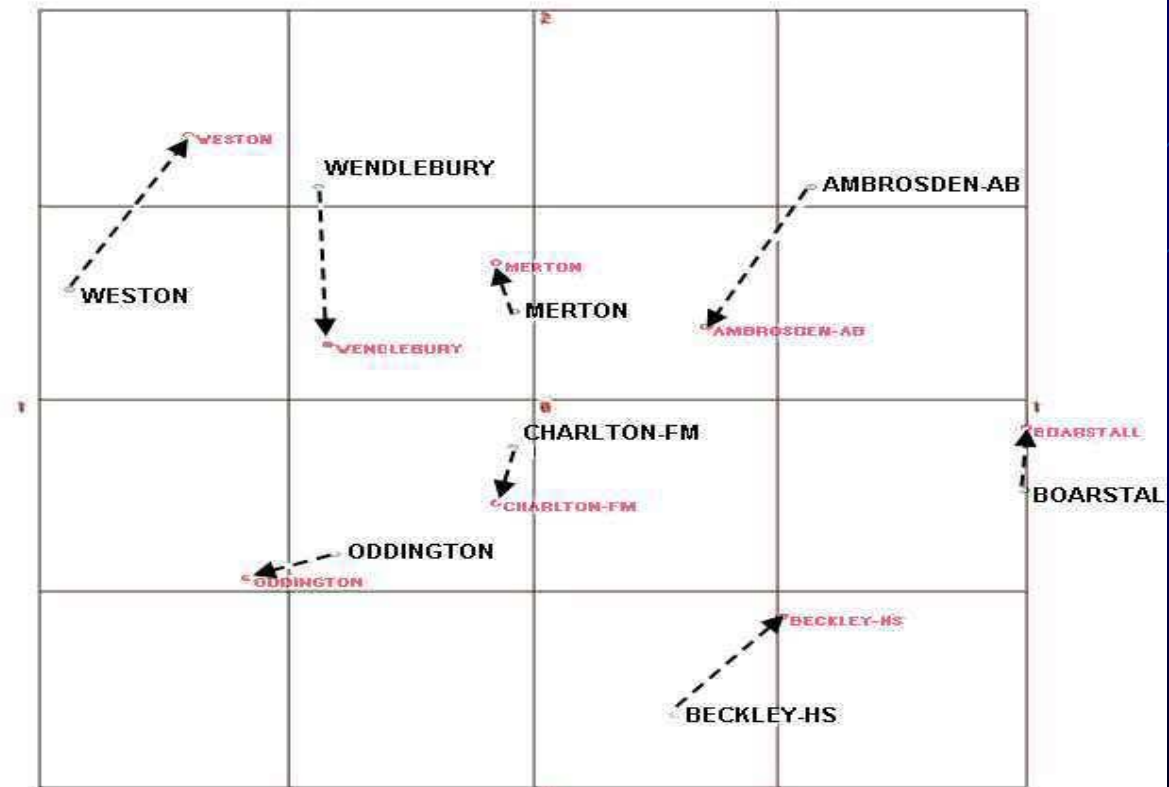
GMPR

Contorted
tho' better fit

Precisely what displacements between GMPR & Real Map?

Locations of all parishes except Weston are within parish boundaries

S2-PRIMARY configuration compared to real configuration



And so ...

1. Even for such “derivative” data which are small in number and rather unconstrained
2. Basic Non-metric MDS scaling can recover the actual positioning remarkably well, especially under the following conditions:
 - i. Data Rates are **standardized** rather than being raw
 - ii. Measure of similarity is used which **symmetrizes** the data (and is preferably **multiplicative** rather than additive), and, most importantly ...
 - iii. Uses the **primary** rather than secondary approach to ties in the data.
3. These turn out to be important guidelines for other studies where there is no “real map” ...

TOBLER & WINEBERG's "CAPPADOCIAN SPECULATION"

- But what if we did not know the real configuration? Could we then go on to predict or discover real locations?
 - "Philosophically, of course, prediction of this sort is impossible but empirically it is done" *<not least by himself !....>*
Tobler
- "Given the degree of similarity or interaction between places one can predict their location" (Tobler 1974). And he should know since he did it!

– Tobler, Waldo and Susan Wineberg (1971) *A Cappadocian Speculation*.
Nature 231, pp 179-180 , *erm* 20 pages on from Kendall !!

You'll remember your ancient history?

- The study of pre-Hittite Assyrian merchant colonies in Bronze Age Anatolia [1940-1740 BCE] was greatly enhanced by the discovery in 1925 of a hoard of 819 cuneiform tablets detailing trading exchange, whose contents contained reference to 119 towns.



Tobler's data

- **DATA:**

- 754 tablets, (referring to 62 towns)
- Make a count of number of tablets on which each pair of towns occurs
- Yields a 62 by 62 symmetric co-occurrence matrix
 - Sparse -- 187 out of 1891 pairs (~10%)
 - Frequencies vary between one and twelve.

- **Assumptions:**

- places mentioned together frequently are probably geographically closer
- large towns occur more frequently than small towns (frequency of mention is surrogate for population size)

Gravity Model

- Interaction is likely to be a function of both population size and the distance separating them (the gravity model):

$$I_{i,j} = kP_i P_j / d_{ij}^2$$

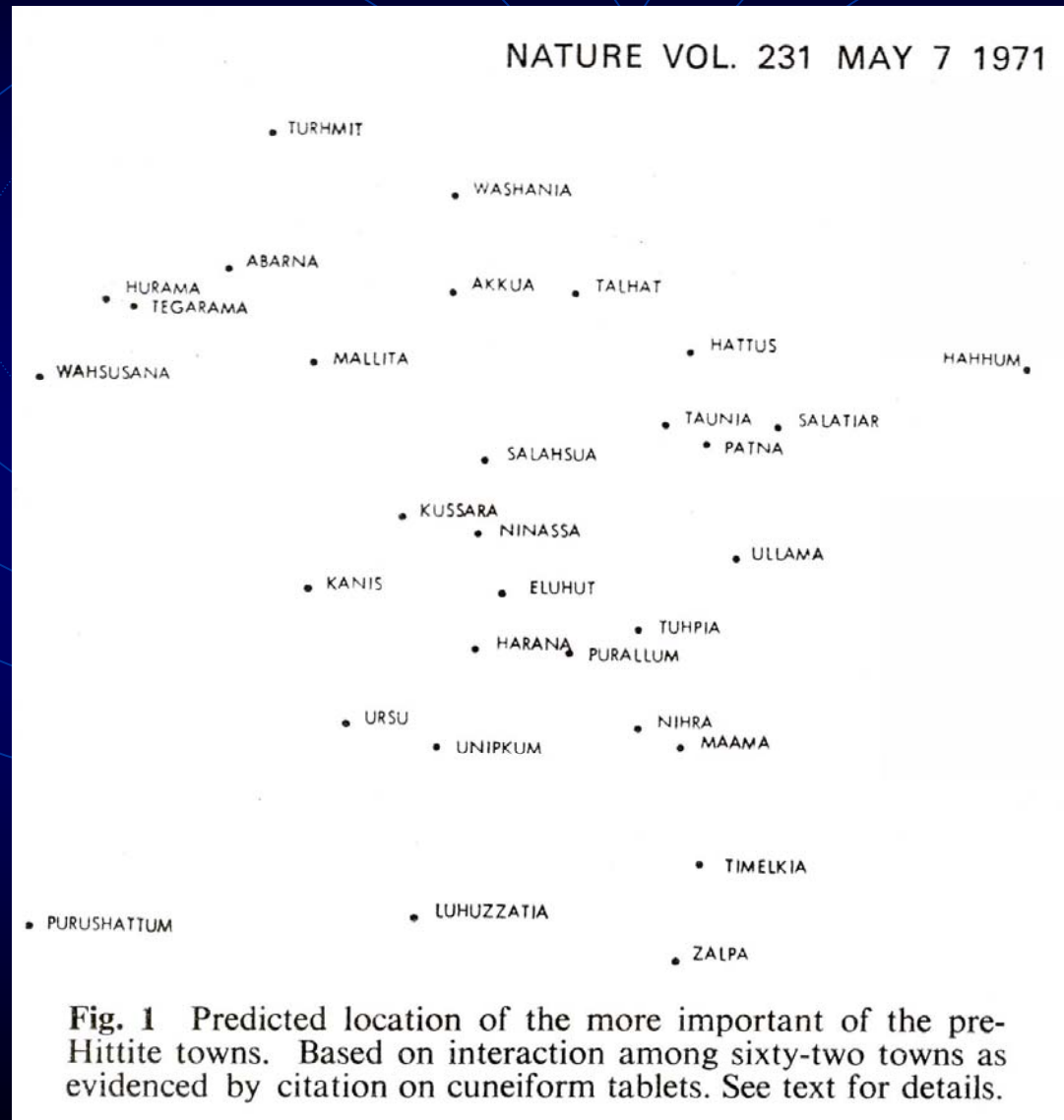
TRANSFORMATION: Guttman's version of SSA) with the non-metric (ordinal) transformation.

Cuneiform Tablet Sites

- virtually none of the locations are known for certain, except Kanis [Kayseri] Hattus (Bogazkoy) and possibly Akkua
 - Other sites have been estimated, notably by Orlin (1970), based on itineraries rather than on interaction
- **MODEL: Euclidean distance**



Tobler Solution

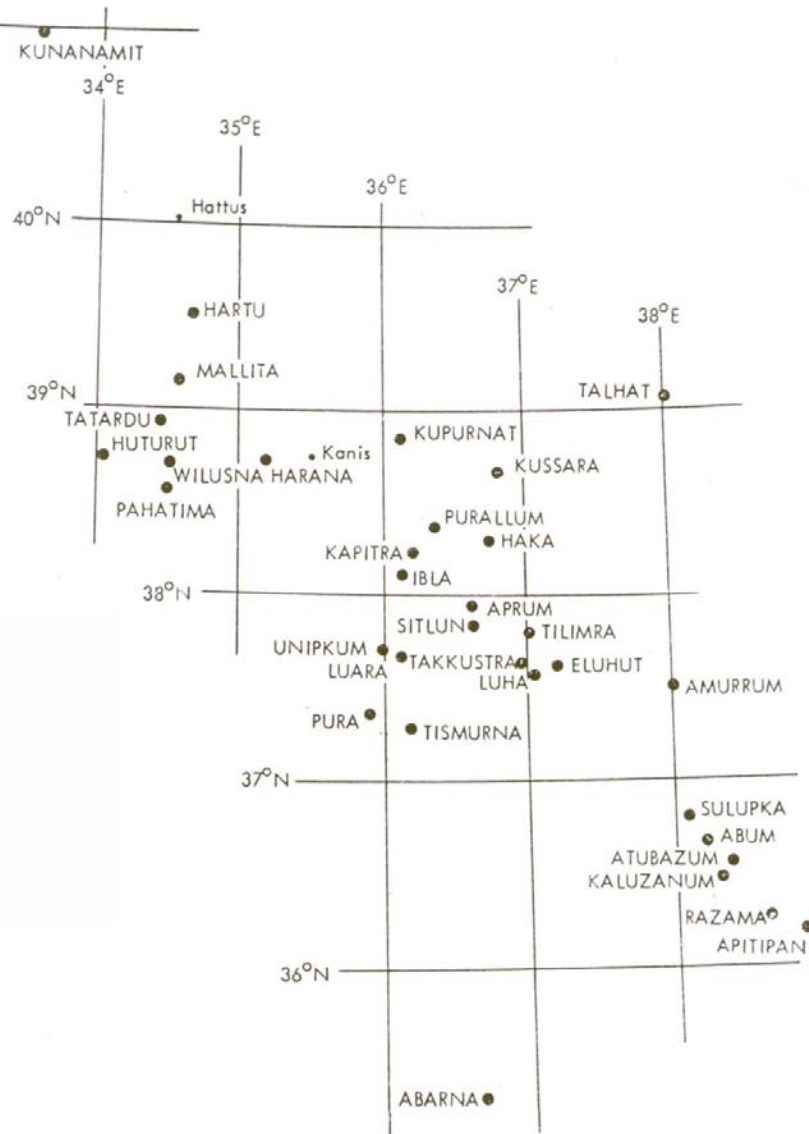


Orlin & Tobler

- Orlin's and Tobler's locations do not agree.
- If there were an agreed framework of known sites, then the information in Tobler's data could be used to estimate the location of unknown sites
- failing this Tobler takes Orlin's estimates of the location of 29 towns which are common to both studies as an external fixed framework (akin to the fixed Real configuration in Kendall's case) and fits the full 33 sites.

Tobler in Orlin

NATURE VOL. 231 MAY 7 1971



To conclude this section ...

- It would be pleasant to report that it would be possible to begin excavations on the locations indicated in the map, but this is not a viable suggestion, as the precision of such locations is estimated by Tobler as averaging 50km, which is rather too wide for a rich archaeological site, especially in an area as rich in mounds as central Turkey” as Tobler comments (p41).
- But the principle is sound, and illustrates well an imaginative and fascinating approach to and application of MDS to such data.

And now to something very different ... failing Banks !

Molinerero: Spanish Bank Failure

Background & Problem:

Between 1978 and 1983: Molinerero reports:

- the Spanish private banking system went through a deep crisis, but received little publicity, probably because it did not result in redundancies or bank failures , although one bank lost its licence.
- The financial crisis that swept the world in the '70s was , at first, unnoticed in Spain due to its relative isolation from international financial markets , but when it arrived it was particularly deep.
 - In 1977 the Bank of Spain, the regulatory body, created the Deposit Guarantee Fund (Fondo de Garantõa de Depositos , FGD) to act as an insurance system.
- **Out of a total of 108 banks , 51 private banks needed the support of the FGD in this period.**

Several statistical techniques

have been deployed to answer the question of estimating failure-probability of bank-company

- univariate statistical analysis in combination with linear discriminant analysis (LDA)
- logit analysis
- self-organizing neural networks
- data envelopment analysis (DEA),
 - (a non-parametric OR programming approach to predicting productivity) .

... But without any great success.

Molinero proposes MDS models are preferable:

- MDS can be used as [less demanding] alternative to DA & Logit to classify & predict banks as failing or continuing
- MDS intuitively more accessible & comprehensible
- MDS is robust to presence of outliers
 - unlike other models used, and especially Data Envelopment Analysis.

Prediction of Bank failure: DATA

Basic information: 9 financial ratios [FR] by 66 Banks (29 of which failed)

- “Variables” : [FR] standardized ratios: indicators of financial health:
 1. Current assets/Total assets
 2. Current assets – cash / Total assets
 3. Current assets/loans [1-3liquidity]
 4. Reserves / Loans [Ability to self-finance]
 5. Net Income / Total assets
 6. Net Income / Total Equity Capital
 7. Net income / loans [Profitability]
 8. Cost of sales / sales [Cost of sales]
 9. Cash flow / loans [Cash flow]
 - *n.b. their high intercorrelation not a problem to MDS*
- ◆ (“Subjects”): the 66 Spanish banks

Prediction of Bank failure: DATA

Molinerero's previous study (Molinerero & Ezzamel (1991, 1996))

- was "R" analysis of inter-relations of FR variables
- This is a "Q" analysis of inter-relations of subjects (Banks).
- (Input) DATA:
 - 2W1M square symmetric matrix
 - (MDS is indifferent to choice of dis/similarity measure)
 - In fact, used PM correlation measure, r
 - PM r is monotonic with Euclidean distance for n-m
 - PM r can be converted to Euclidean Distance D for metric.
 - keeps comparability with other studies
- TRANSFORMATION:
 - Ordinal / non-metric Monotonic
- MODEL:
 - Euclidean Distance

Analysis of Molinero Data

- Preliminary PCA analysis indicated that 3 dimensions suffice
 - Previous metric studies: between 3 & 7D
 - Non-metric more robust and allows lower dimensionality
 - Badness-of-fit [Young's] Raw Stress₁:
 - 1D 0.211
 - **2D 0.050**
 - 3D 0.027
 - Most analysis done in 2D subspace.
 - Analysis plots the Banks as numbered points
 - “External information” on failure (black vs white!)
 - “External” mapping (using PRO-FIT) of internal information
 - Variables/FRs as Vectors in Bank-space

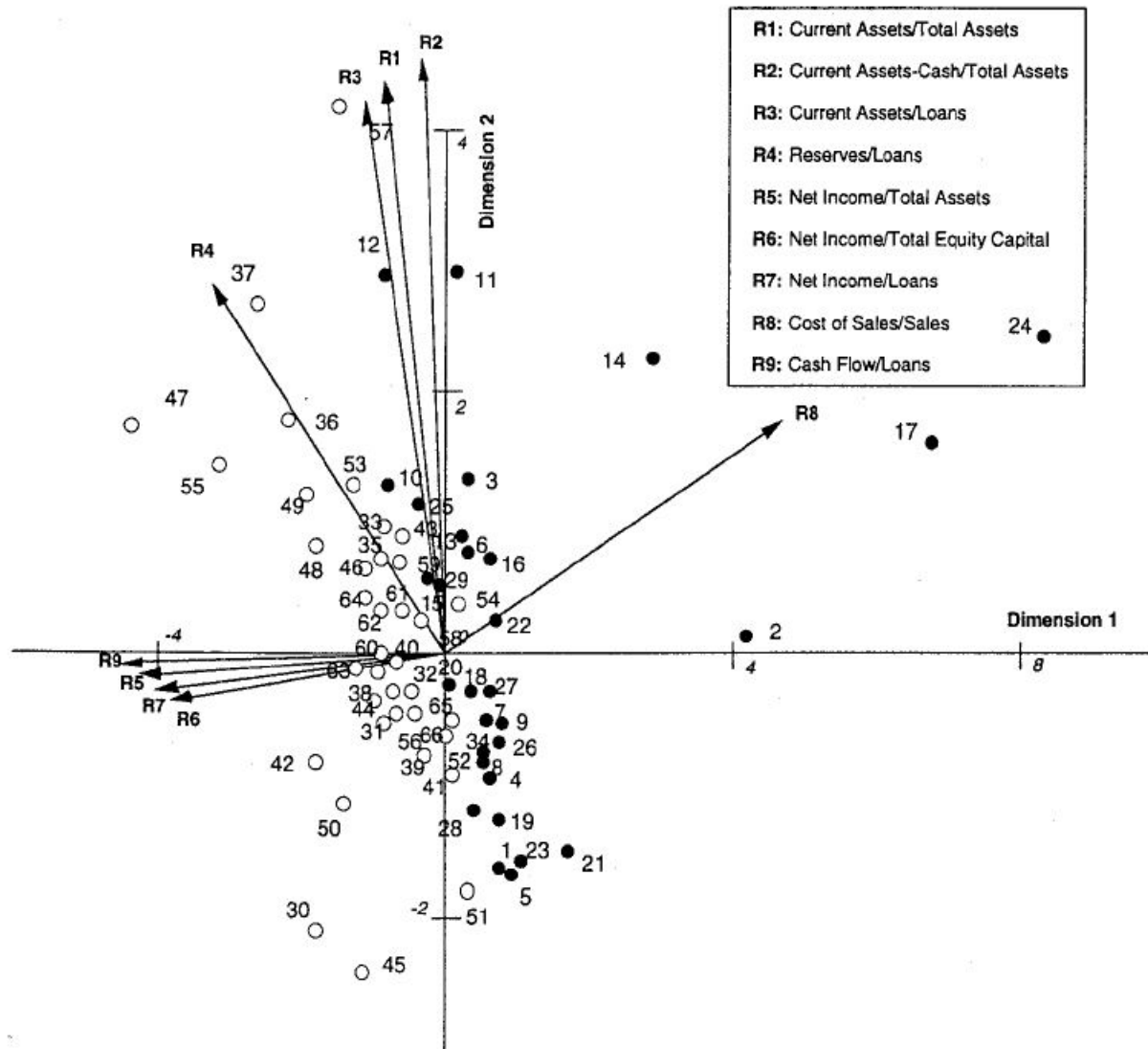


Fig. 1. Multidimensional scaling representation of banks in Spain. Projection on the first two dimensions with profit analysis results. Empty circles correspond to continuing banks, and full circles to failed banks.

Interpretation of MDS Map (1)

(No need to do dimensional interpretation, though in this case ...)

- D1 collinear with FR5,6,7,9 : “PROFITABILITY & Cash-flow”
- D2 collinear with FR1,2,3: “LIQUIDITY”
- Failed banks lie to RHS, Continuing banks toward LHS
 - Suggests D1 is powerful discriminator
- But a *curved* line would better separate Failed from Cont.
 - Suggests *non-linear* discriminator and
 - relationship between D1 & D2 co-ordinates is curvilinear
- So: two Logit analyses performed ... & mapped into the Banks-space

Interpretation of MDS Map using Logit analysis

- Logit using non-linear terms (quadratic & interaction)
 - only non-linear (quadratic) terms needed
 - Logit analysis returns a probability for each point/Bank – “probability of failure/non-failure”; and points in space having same probability of failure can be joined to produce a curved “iso-solvency surface”
 - Curves for 1.0 – 0.75 – 0.5 and 0.25 probability of **non-failure** were chosen
- Representation of Logit analyses in 2D configuration:
 - Configuration simplified by compressing D2
 - Failed banks are black, Continuing banks are white
 - Iso-solvency non-failure curves drawn into configuration ...

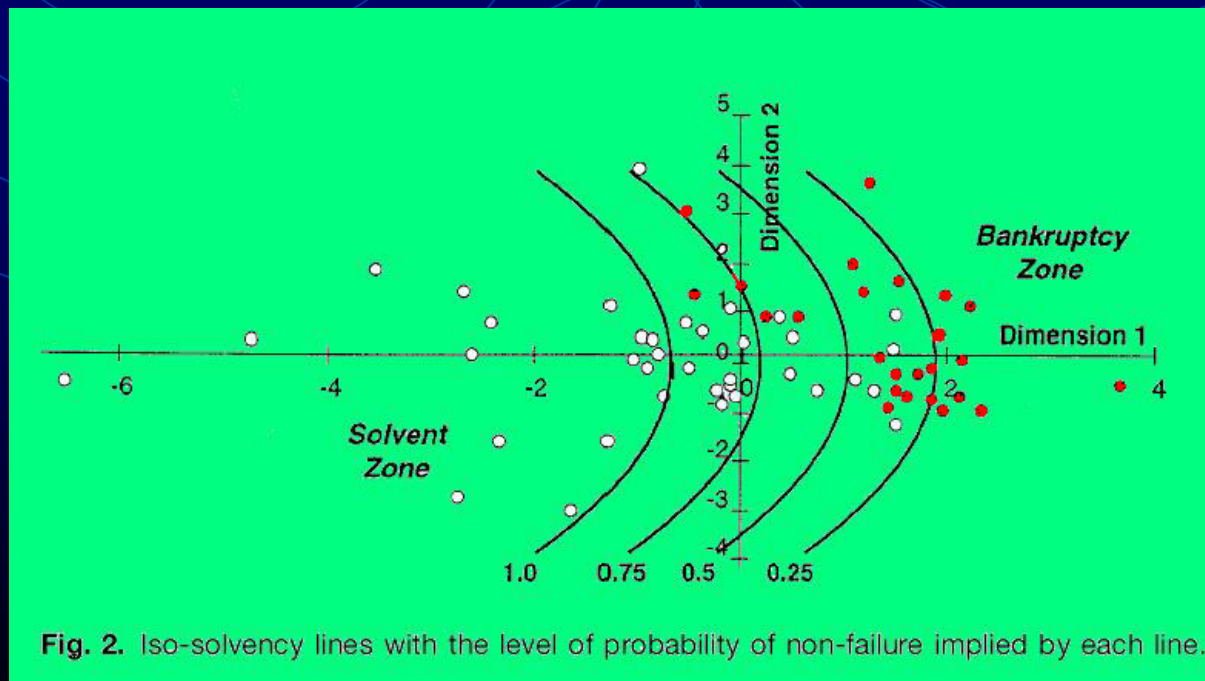


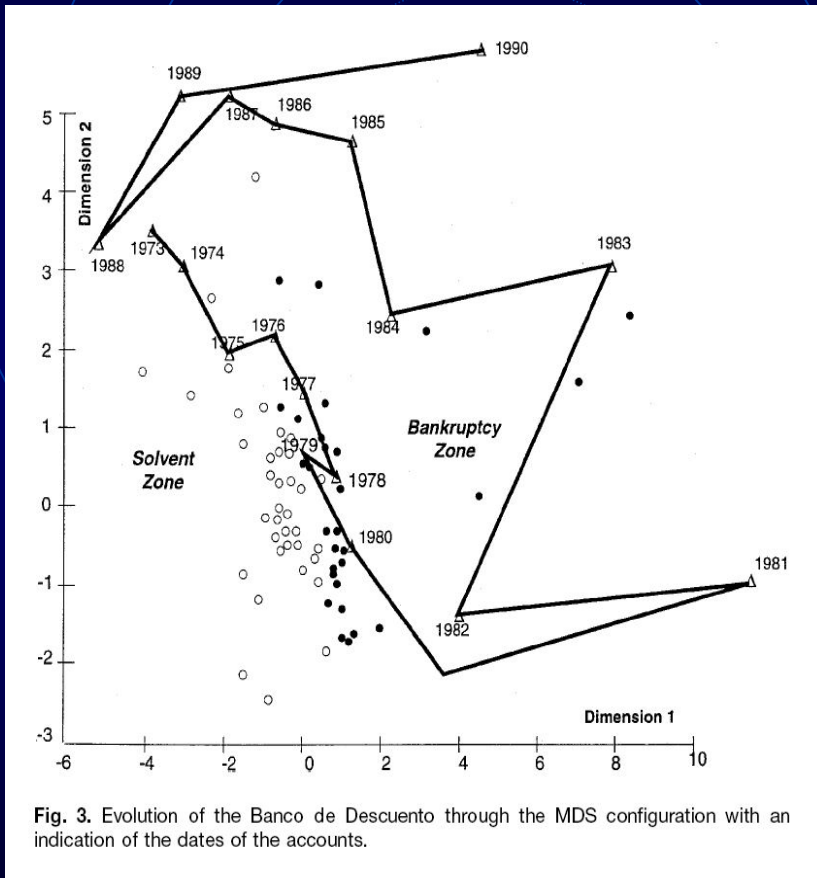
Fig. 2. Iso-solvency lines with the level of probability of non-failure implied by each line.

Impressive, or what? 😊

- no simple frontier (or middle ground) between failed and non-failed banks
- linear discriminant on D1 would not do well in middle ranges (nor in LDA studies), but curvilinear in 2D does excellently
- If take 0.5 as criterion of failure, mis-classification of only 4 banks!

Tracing the “pathway to failure”

What about including other banks (in terms of their 9 FR properties? It can be done, using external MDS (PREFMAP-3): Banco de Desuento' yearly data were entered ...



1973-76: hi in Solvency zone & in D2 (Liquidity)
'77 moves into initial failed bank area, worsens in '78, **fails in '80**; refloats but moves further into failure zone
Liquidity improves
'88 moves back into safe zone
'89 and '90 moves back into failure zone and **bankrupts in 1990**
BdeD was bank with high liquidity ratios

- cause of failure was low profitability
- It could have been identified as “at risk” 4 yrs before failure...

MDS representation helps explain causes of failure & plot the process

MDS has been used:

- Easily assimilable visual representation of MV data
- Two dimensions are often enough (Shepard's "Law")
- Prediction: Assessment of health of new point/bank by external MDS mapping

- PROGRAMS:
 - **SPSS**
 - Proxscal & Categories (some of above are possible)
 - **NewMDSX:**
 - MINISSA/MRSCAL (Bank-space)
 - PROFIT (map FRs as vectors into Bank space)
 - PREFMAP-3 (locate new point/bank into Bank Space)

And now we turn to the Third section of the Talk, **Drugs ...** promising more excitements!

- But some of you are looking anxiously at your watches
- Wondering whether we can perhaps trade the third section for a little liquid refreshment?
- **Of course we can!**
 - and that was actually the intention, because the topic will be among the first applications in the Course, so we may leave it until tomorrow!
 - And those not attending the Course , who are nonetheless interested, can (after tomorrow) see the slides on:
 - <http://apmc.newmdsx.com>

Thank you