Rank Inadequacy: A Partially Ordered Set Approach For Multivariate Data Analysis

Chris Brunsdon*[1]

[1]National Centre for Geocomputation, Maynooth University

Summary

A major issue with 'league tables' of deprivation, university performance and other indicators is that they attempt to impose a rank ordering on multivariate sets of indicators. In this talk, an approach to working with such multivariate descriptors is proposed that attempts encapsulate some degree of comparison between individual entities without presuming that a full ranking is well-defined, through the use of partially ordered sets. The approach is demonstrated on a set of well being indicators in the US.

KEYWORDS: Posets, Deprivation Indicators, Visualisation, Spatial Dependency

1 Introduction

It is quite common practice to assess some social aspect of places in terms of compound indices of well-being, deprivation or some other portmanteau term, whose aspects are measured in terms of several officially compiled statistics – see for example Noble et al. (2006). Typically, if these statistics are denoted as $s_{1i}, s_{2i}, \cdots, s_{mi}$ for locations $i = 1 \cdots m$, then the indicator is a weighted combination of these quantities, say

$$I_i = w_1 s_{1i} + w_2 s_{2i} + \cdots + w_m s_{mi}$$

Often the I_i values over the locations i and are ranked and prsented as a 'league table'. This is problematic since although it is acknowledged that a portmanteau concept is being measured, and that it requires several variables to attempt to encapsulate all aspects of the concept, ranking is only meaningful for scalar (one dimensional) quantities. The weighted sum approach attempts to overcome this by projecting the several statistics onto the real line therefore creating a scalar, but ranking then depends on choice of weights. If the weights change sufficiently, different rankings will be reported. However, choice of weights is often subjective, reflecting an individual's choice of the importance of each variable.

In this paper, an alternative approach via *partially ordered sets* or *posets* (Dushnik and Miller, 1941) is considered - this aims to identify structure in this kind of data allowing for the fact that in some cases, the multi-dimensionality of the indicators results in some pairs of places not being comparable

*christopher.brunsdon@nuim.ie

in a convincing way. These have been applied in a GI science context previously (Kainz et al., 1993) to model relationships between spatial entities, but here they will be used to analyse *attributes* of spatial objects. The final outcome is a visualisation approach that highlights overall trends, but avoids imposing a complete ordering on the attributes where no consensual ordering is achievable.

2 A Practical Example

To illustrate the issues introduced above, consider a data set taken from the U.S. Dept. of Commerce Bureau of the Census in 1977[1] having variables for each of the 48 co-terminous States of the US as listed in Table 1 below

Indicator	Name	Description
s_1	Income	Per capita income (1974)
s_2	Illiteracy	Illiteracy (1970 percept of popn.)
s_3	LifeExp	Life expectancy in years (1969-71)
s_4	Murder	Murder and non-negligent manslaughter rate per 100,000 popn. (1976)
s_5	HSGrad	Percent high-school graduates (1970)

Table 1: US Well-being variables

Suppose these were used to create a well-being index encompassing measures of income, educational attainment, mortality and crime, via the following procedure: firstly standardising $s_1 \cdots s_5$ to z-scores (denoted $z_1 \cdots z_5$ respectively) and changing signs for z_2 and z_5 so that higher scores for Illiteracy and Murder correspond to 'better' outcomes. Secondly combining $z_1 \cdots z_5$ to create the weighted index $w_1z_1 + w_2z_2 + w_3z_3 + w_4z_4 + w_5z_5$.

Now suppose one analyst wishes to weight all of the z-scores equally, so that $w_1 = w_2 = \cdots = w_5 = 1$, giving an index I_1, whilst another analyst places a double weighting on Murder, so that their index, I_2 is defined by with all $w_i = 1$ except $w_4 = 2$. Each ranking gives a 'league table' of US states, as shown in Figure 1 below. On the left the states are listed on the basis of ranking of I_1, and on the right they are listed on the basis of I_2. In the centre of the tables lines join each state in the two 'leagues', to allow comparison of their rankings. Clearly due to lack of concensus in weight choice there is some variation between the tables, with most inconsistencies towards the centre of the tables.

Although in both 'league tables' there is some consensus - most states are near the top, or near the bottom or somewhere central in *both* tables, strict ranking is ambiguous.

As an alternative, suppose principle components were used to inform the weighting. The loadings for all principle components based on the z-scores (with signs for Illiteracy and Murder reversed) are given in table 2 :

The loadings of the first component suggest that this is a weighted overall well-being measure, placing slightly less importance on Income than the other scores. The second component contrasts Income

[1] U.S. Department of Commerce, Bureau of the Census Statistical Abstract of the United States *and* County and City Data Book (1977).

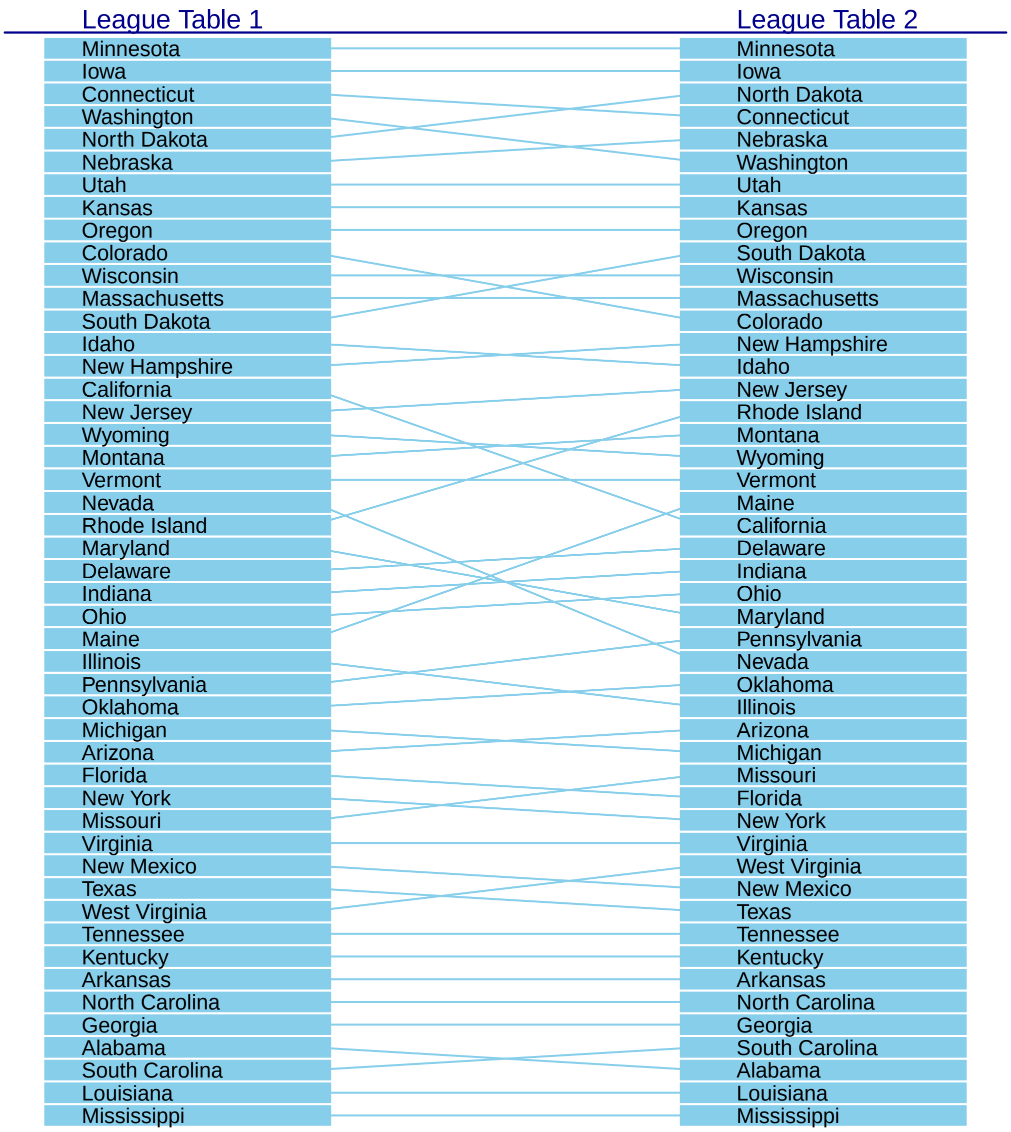

Figure 1: Comparison of ranks

	PC1	PC2	PC3	PC4	PC5
Income	0.37	0.75	0.54	-0.05	-0.14
Illiteracy	0.49	0.02	-0.30	-0.65	0.50
LifeExp	0.47	-0.33	0.32	0.57	0.49
Murder	0.45	-0.53	0.24	-0.27	-0.62
HSGrad	0.46	0.24	-0.68	0.42	-0.31

Table 2: US Principal Component Analysis Weights

and High School Graduate rates against the other scores. Were the first principle component to be used as an alternative index, this accounts for 69.0% of the variance suggesting that a reasonable amount of the structure in this data requires more than one dimension to be represented - this can also be seen in Figure 2 - in which the first and second principal components of the data are plotted.

Considering issues of variability due to choice of weights (due to subjective choice of weighting scheme) and variability in the data beyond that acheivable through a one-dimensional weighted index (as in the principal components example) it seems that that in general the data is not meaningfully and unambiguously rankable - and that comparison between ranks will have results that are influenced by the (subjective) choice of weights.

Thus, in this paper an alternative approach is proposed. In the above example, although *some* US states clearly have a better degree of well-being than others a strict ranking may not be meaningful, and a consensus as to a *definitive* ranking may not be reachable. Thus, an approach to comparison using *partially ordered sets* or *posets* is considered here. A poset is a set, together with a comparison operator which may be applied to some pairs of elements in the set, but not neccessarily all. In this paper, it will be outlined how geographical data sets such as that in the above example may be represented as a poset, and how this may be used as a tool to identify structure (including some aspect of stratification) in the data, without going to the extreme of providing a strict ranking where this may not be appropriate. In addition, the degree to ranking *is* a valid option may be assessed.

3 Posets: an Overview

In this section, the basic ideas underlying posets will be introduced. Suppose there is a set $\mathcal{P}$ and a binary relation $\preceq$, defined between elements of P. Then $\{\mathcal{P}, \prec\}$ is a poset if $\prec$ satisfies, for $a, b, c \in \mathcal{P}$

1. $a \preceq a$ (Reflexivity)
2. If $a \preceq b$ and $b \preceq a$ then $a = b$ (Antisymmetry)
3. If $a \preceq b$ and $b \preceq c$ then $a \preceq c$ (Transitivity)

If either $a \preceq b$ or $b \preceq a$ then a and b are *comparable.* However, it is not required *all* a, b pairs are comparable: there could be a, b pairs for which neither $a \preceq b$ nor $b \preceq a$.

If in a particular poset *all* a and b are comparable, it is known as a *totally ordered set* or simply an *ordered set*, and in this case $\preceq$ is simply a comparison operator. For example if $\mathcal{P} = \{1 \cdots 10\}$ and

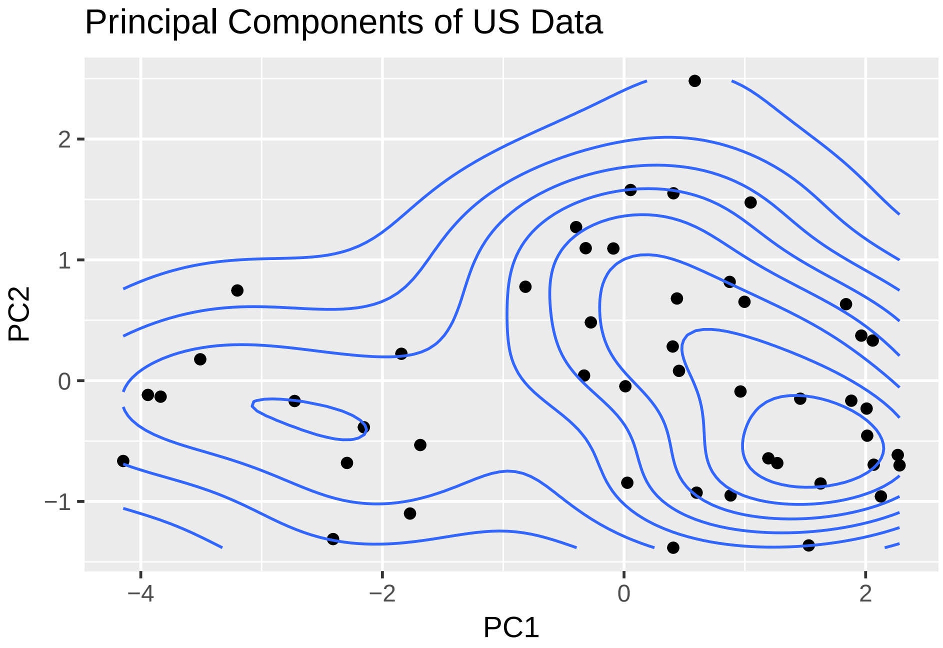

Figure 2: Principal Components 1 and 2

$\preceq$ is simply the numerical comparison operator $\leq$ then $\{\mathcal{P}, \preceq\}$ is both a poset and an ordered set. With an ordered set, ranking is always possible - any sorting algorithm may be applied using $\preceq$ as the comparison operator used in choosing whether to swap a and b in an ordered list. However, here attention will be focussed more on situations where $\{\mathcal{P}, \preceq\}$ is a poset but not an ordered set.

3.1 Posets: Further notation and vocabulary

Although the definitions above are sufficient to specify a poset, some further terms are useful to simplify further discussion and outline further ideas. A number of further relational operators may be defined in terms of $\preceq$.

1. $b \succeq a$ if and only if $a \preceq b$
2. $a \prec b$ if $a \preceq b$ and $a \neq b$
3. $a \succ b$ if $a \succeq b$ and $a \neq b$

Also some further terms are defined:

- A *chain* $\mathcal{C} \subseteq \mathcal{P}$ is a set such that all a, b in $\mathcal{C}$ are comparable. Note that a chain is therefore an ordered set. A chain is *maximal* if no other chain $\mathcal{C}'$ exists such that $\mathcal{C} \subset \mathcal{C}'$.
- The *depth* of a poset $\{\mathcal{P}, \preceq\}$ is the length of its longest chain.
- An *antichain* $\mathcal{A} \subseteq \mathcal{P}$ is a set such that no distinct a, b in $\mathcal{A}$ are comparable. An antichain is *maximal* if no other antichain $\mathcal{A}'$ exists such that $\mathcal{A} \subset \mathcal{A}'$.
- An element $a \in \mathcal{P}$ is a *maximal element* if there is no element $b \in \mathcal{P}$ such that $a \preceq b$. The *maximal element set* is the set of all such elements.
- An element $a \in \mathcal{P}$ is a *minimal element* if there is no element $b \in \mathcal{P}$ such that $b \preceq a$. The *minimal element set* is the set of all such elements.
- If a, b are distinct elements in $\mathcal{P}$ and $a \preceq b$ and there is no c in $\mathcal{P}$ such that $a \preceq c$ and $c \preceq b$ then a is said to *cover* b - denoted by $a \lhd b$ or $b \rhd a$.

3.2 Relating Posets to US Well-being Data

Above a number of formal definitions are given - in this section they will be related to multivariate data sets and applications in indexing and ranking. The key idea is to attribute each US state with a vector $(z_1, z_2, \cdots, z_5)$ of the 5 z-scores as defined in section 2, and designate these as elements in $\mathcal{P}$, together with a partial ordering relation $\preceq$ intended to compare states, when comparison has meaning. The definition proposed here is that if a and b are states represented by the vectors $(z_{1a}, z_{2a}, \cdots, z_{5a})$ and $(z_{1b}, z_{2b}, \cdots, z_{5b})$ then

$$a \preceq b \text{ if and only if } z_{1a} \leq z_{1b} \text{ and } z_{2a} \leq z_{2b} \text{ and } \cdots z_{5a} \leq z_{5b}$$

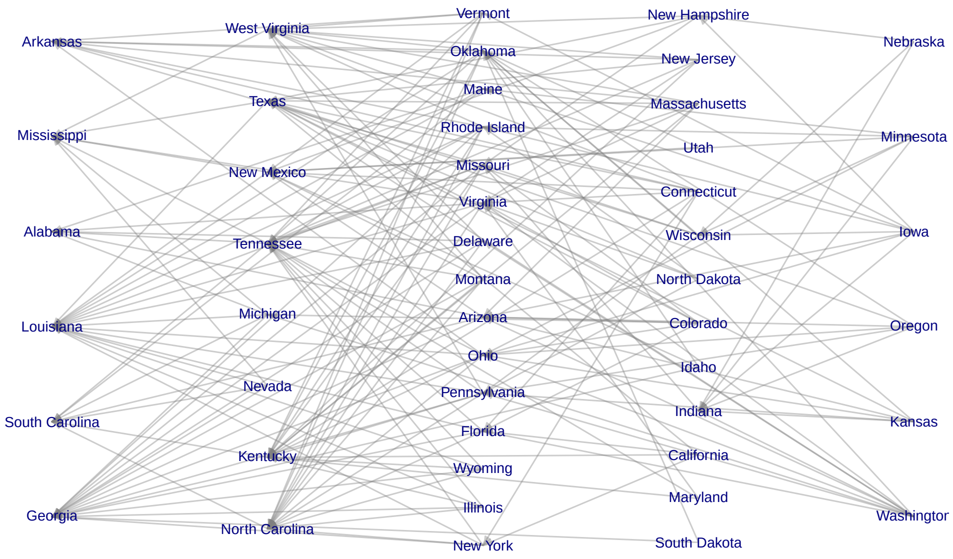

Figure 3: Hasse Diagram (Peeled Minimal Elements)

That is, $a \preceq b$ implies that *every* z-score for state a is less than or equal to that for state b. If neither $a \preceq b$ nor $a \succeq b$ than the two states are not comparable. This will occur when some of the z scores for state a are are greater than or equal to the equivalent scores for state b, but some are not.

3.3 Visualising Posets

Posets are often represented by *Hasse Diagrams*. These are diagrams representing the elements of a poset for which the $\preceq$ relation holds. The diagrams take the form of a network where paths are directed (that is a path from a to b differs from a path from b to a) and where, if nodes a amd b satisfy $a \preceq b$ then there is a path from a to b on the network. This can be achieved if there is a directed edge from a to b if $a \rhd b$. Also, if $a \preceq b$ then on the diagram, a will appear to the aligned with, or to the left of b. Note that in general, there is not a *unique* Hasse diagram for a given poset. One particular Hasse diagram for the US well-being data is shown in Figure 3.

Here, a *peeled minimal element* algorithm is used - so that the rightmost column of states form the minimal element set - and the preciding column consists of the minimal element set after the previous elements are removed, and so on. This implies that Missippi, Arkansas, Louisiana, Georgia, South Carolina and Alabama are such that there are no states that they rank above on *all* of the 'scores'. Note that this does not imply that the leftmost column are the maximal element set - note that Utah and Maryland are elements of this set, but are in the second column from the left.

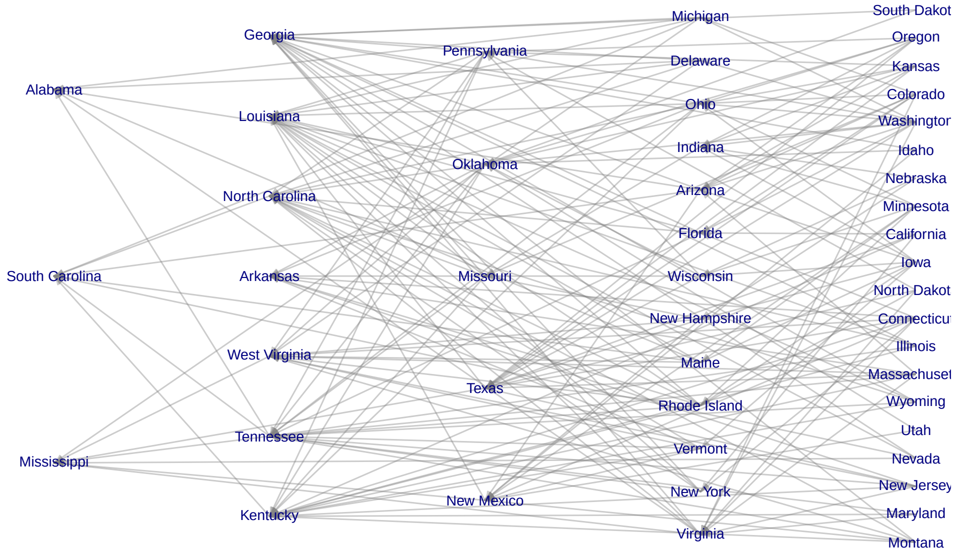

Figure 4: Hasse Diagram (Peeled Maximal Elements)

An alternative Hasse diagram arrangement is a *peeled maximal element* algorithm, which progressively assigns columns from left to right, by recursively identifying the maximal element set and removing it. The resultant diagram for the US states is shown in Figure 4.

Although this gives a slightly different picture, both show the complexity of the relationships in a way that a weighted indicator and league table approach systematically masks. The first diagram highlights states with lower well being - in the sense that in the leftmost column states that have at least one indicator that is worst than any other state, whereas the second shows those with a better level of well being - these have no states that having at least one indicator that issecond to none.

4 The Geographical Viewpoint

To gain an insight into geographical aspects of these relationships, a very simple approach is to relax the left-right dependency rule in Hasse diagrams, and use the geographical location of each state to determine the location of the node. This is seen in Figure 5. A backdrop of US co-terminous states is also added to provide context.

A clear trend is visible - perhaps a stronger narrative than the Hasse diagrams earlier - showing that in general states in the north west tend to enjoy a better state of well being (at least on the basis of this index) than those in the south, and towards the eastern seaboard.

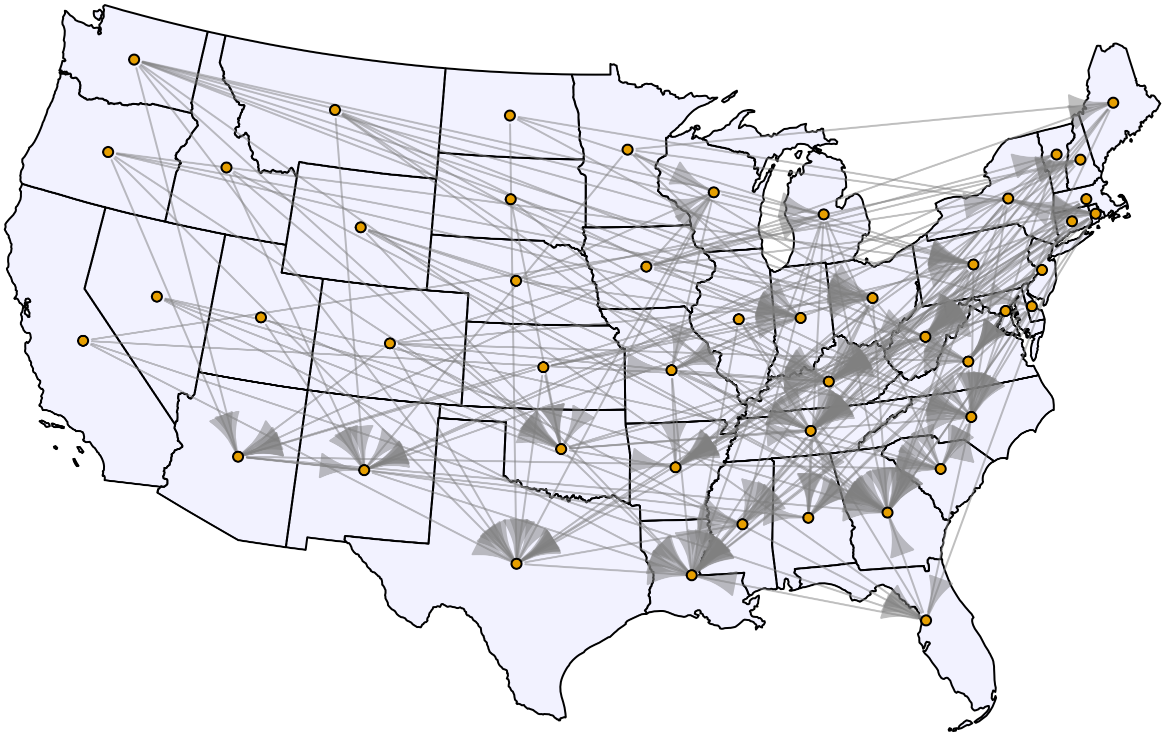

Figure 5: Hasse Diagram (Based on Geographical Location)

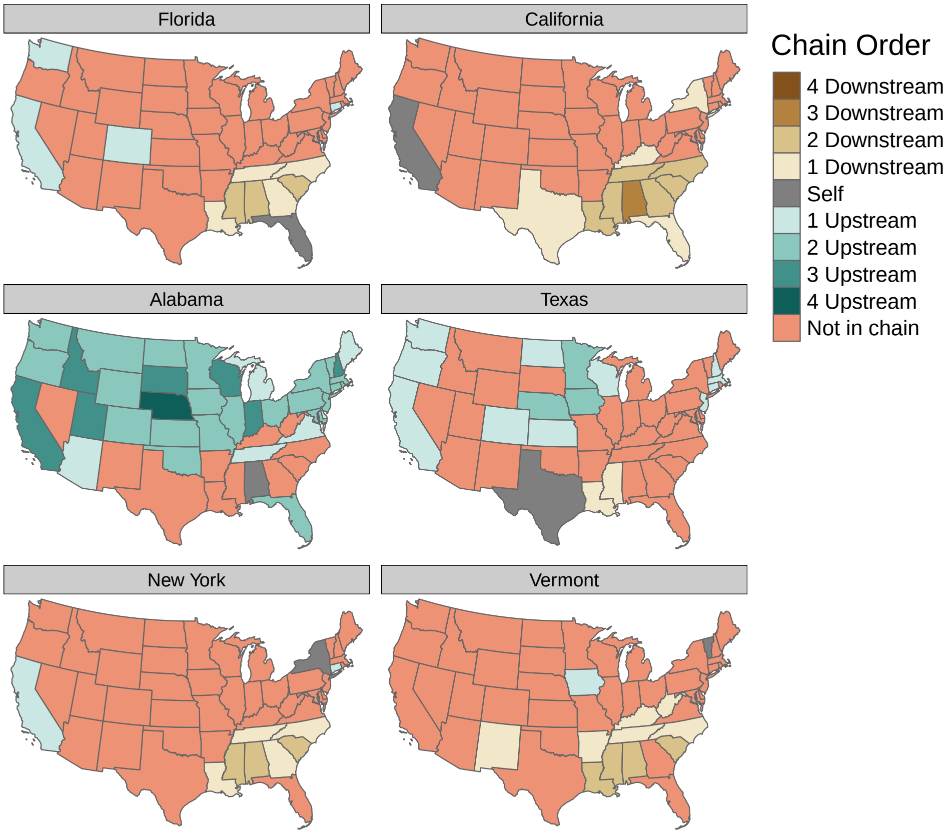

Figure 6: State-focused Relationship Maps

4.1 State-focused Relative Chains

Further insight nmay be gained by considering an individual state in relation to the others in the US. In particular, when considering a the network representing a Hasse diagram, for a given state one can find the length of the path to all of the other states. If we consider state s, there will be a set $\mathcal{C}_s$ of states that are unrelated to s - that is if $x \in \mathcal{C}_s$ then neither of $x \preceq s$ or $s \preceq x$ holds - that is states for which some indices are greater for s and some for which they are greater for x. There will also be a set of states $\mathcal{D}_s$ that are 'downstream' from s - that is there is a path from s to these states on the Hasse diagram. There will also be another set of states $\mathcal{U}_s$ that are 'upstream' from s. There exists a path from these states to s on the Hasse diagram. These distinct sets can be shown on a choropleth map. This is shown in Figure 6 - also using intensity of shading to show the *length* of the path on the Hasse diagram for states in $\mathcal{D}_s$ and $\mathcal{U}_s$ - for a sample of six states - Florida, California, Alabama,Texas, New York and Vermont.

From these maps it is evident that certain spatial trends occur - for example although California has one of the higher well-being score sets the states downstream are predominantly in the south

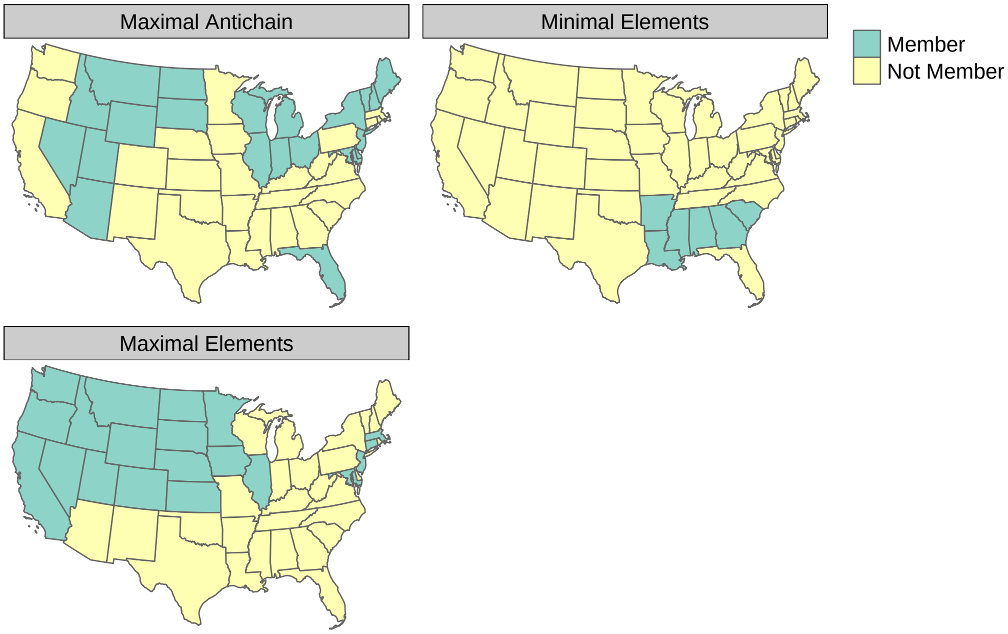

Figure 7: Significant Set Maps

and east of the US. For Florida there are some upstream and some downstream states, but the downstream ones are in the south east and the upstram ones are further west. Alabama has generally low scores, but most of the upstream states are further north.

Another striking point is that the unrelated states in each of the examples appear to be spatially clustered. Finally it can be noted that in these examples, many of the neighbours of the focal state s are members of $\mathcal{C}_s$ - that is, although patterns are observed over wider distances, there are fewer clear relationships between nearby states.

It is also notable that the non-comparable states exhibit spatial clustering. For example, a join-count statistic (Cliff and Ord, 1981) applied to the New York map (testing for clustering of members of $\mathcal{C}_s$ in the New York map) has a value of 3.03 and a p-value of 0.001 suggesting strong evidence of clustering.

4.2 Minimal and Maximal Elements and the Maximal Antichain

In terms of identifying trends, it is also helpful to map the geographical location of the states in the sets of m inimal and maximal elements - and also the those constituting the maximal antichain. There are shown in Figure 7 - again there are clear spatial patterns oin all of these - the minimal element set consisting entirely of coterminous states in the south-eastern US. The maximal elements make two coterminous groups, one in the northwest, spreaading to the midwest, and another in the north east. These tell a similar story to the state-focused maps shown earlier. The maximal antichain again suggests large spatial clusters - this time of incomparable states - where no state betters its neighbours in all of the well-being indicators. Although the visualisation provides strong evidence, those who feel compelled to carry out a formal statistical test may consult the join count statistic tests in table 3.

These tests reinforce the suggestion in an earlier section, that although a full ranking may not be meaningful, and indeed impose spurious detail, there are broader geographical trends in terms of the location of the minimal and maximal element sets. There is also evidence - in the form of spatial autocorrelation of the maximal antichain members - that nearby states are prone to being incomparable. Both of these observations lead to a hypothesised variant on Tobler's First Law (Tobler, 1970) for partially orderable data :

> *"not everything is comparable to everything else, but near things are less likely to be comparable than distant things."*

This has implications in the analysis of weighted indices, particularly when using statistics such as Moran's I - since this is defined in terms of neighbouring values of the variable under investigation. Whereas this may be useful in situations where a scalar quantity is intuitively defined, it may be less interpretable in the analysis of portmanteau indicators, where neighbouring values appear to be less likely to be consensually comparable.

	Join Count Statistic	p-value
Minimal Elements	5.043	0.000
Maximal Elements	4.076	0.000
Maximal Antichain	2.817	0.002

Table 3: Results of Join-Count tests for Spatial Arrangement

References

Cliff, A. D. and Ord, J. K. (1981). Spatial processes: models & applications.

Dushnik, B. and Miller, E. (1941). Partially ordered sets. *American Journal of Mathematics*, 63:600–610.

Kainz, W., Egenhofer, M. J., and Greasley, I. (1993). Modelling spatial relations and operations with partially ordered sets. *International Journal of Geographical Information Systems*, 7(3):215–229.

Noble, M., Wright, G., Smith, G., and Dibben, C. (2006). Measuring multiple deprivation at the small-area level. *Environment and planning A*, 38(1):169–185.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(1):234–240.