Topics in Data Visualization

A Survey Of Some Techniques For Higher Dimensional Data

Jul 30 2014

Charlotte Wickham

vis.cwick.co.nz

Multi/High Dimensional Data

Two continuous variables easily examined in a scatterplot,

A **third** added using size, or colour.

Maybe facet on a **fourth**.

What about more than that?

What are we looking for?

- simple relationships between variables
- low dimensional structure
- clusters

We'll think about a having a collection of observations, where each observation consists of measurements on multiple variables.

We are interested in clusters of observations, and relationships between variables.

```
> atmos
Source: local data frame [41,472 x 11]

         lat   long year month surftemp  temp pressure ozone cloudlow cloudmid cloudhigh
1   36.20000 -113.8 1995     1    272.7 272.1      835   304      7.5     34.5      26.0
2   33.70435 -113.8 1995     1    279.5 282.2      940   304     11.5     32.5      20.0
3   31.20870 -113.8 1995     1    284.7 285.2      960   298     16.5     26.0      16.0
4   28.71304 -113.8 1995     1    289.3 290.7      990   276     20.5     14.5      13.0
5   26.21739 -113.8 1995     1    292.2 292.7     1000   274     26.0     10.5       7.5
6   23.72174 -113.8 1995     1    294.1 293.6     1000   264     30.0      9.5       8.0
7   21.22609 -113.8 1995     1    295.0 294.6     1000   258     29.5     11.0      14.5
8   18.73043 -113.8 1995     1    298.3 296.9     1000   252     26.5     17.5      19.5
9   16.23478 -113.8 1995     1    300.1 297.8     1000   250     27.5     18.5      22.5
10 13.73913 -113.8 1995     1    300.1 298.7     1000   250     26.0     16.5      21.0
```

observation is a spatial grid cell in a month of a year,
variables are `surftemp, temp, pressure, ozone, cloudlow, cloud mid and cloudhigh`

Scatterplot matrices

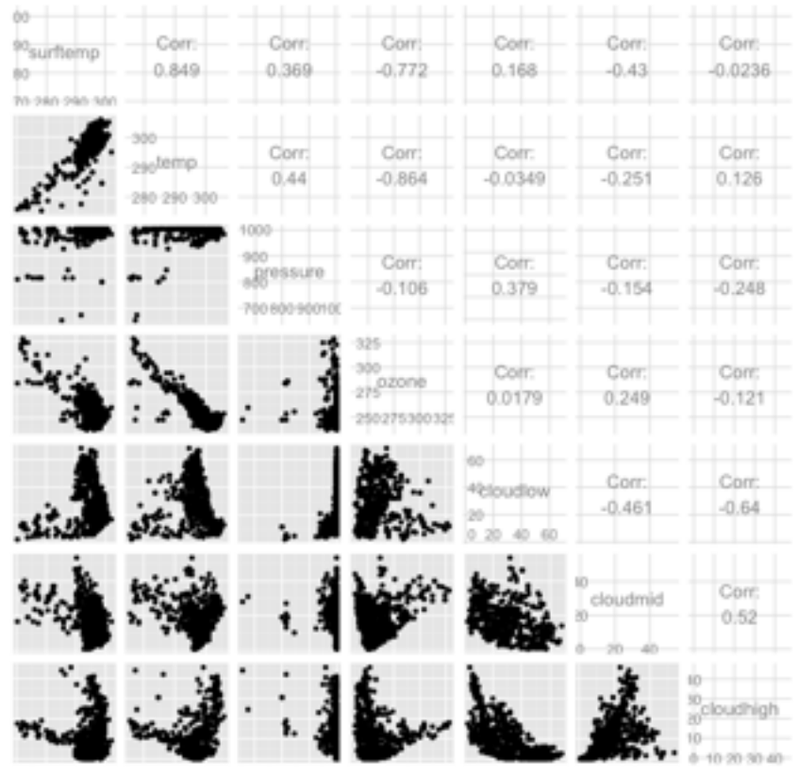
Parallel coordinates

Tours

Glyphs

Heatmaps + Seriation

Scatterplot matrix



Good at:

finding relationships between pairs of variables

clusters, if clusters are well defined with only one or two variables

Bad at:

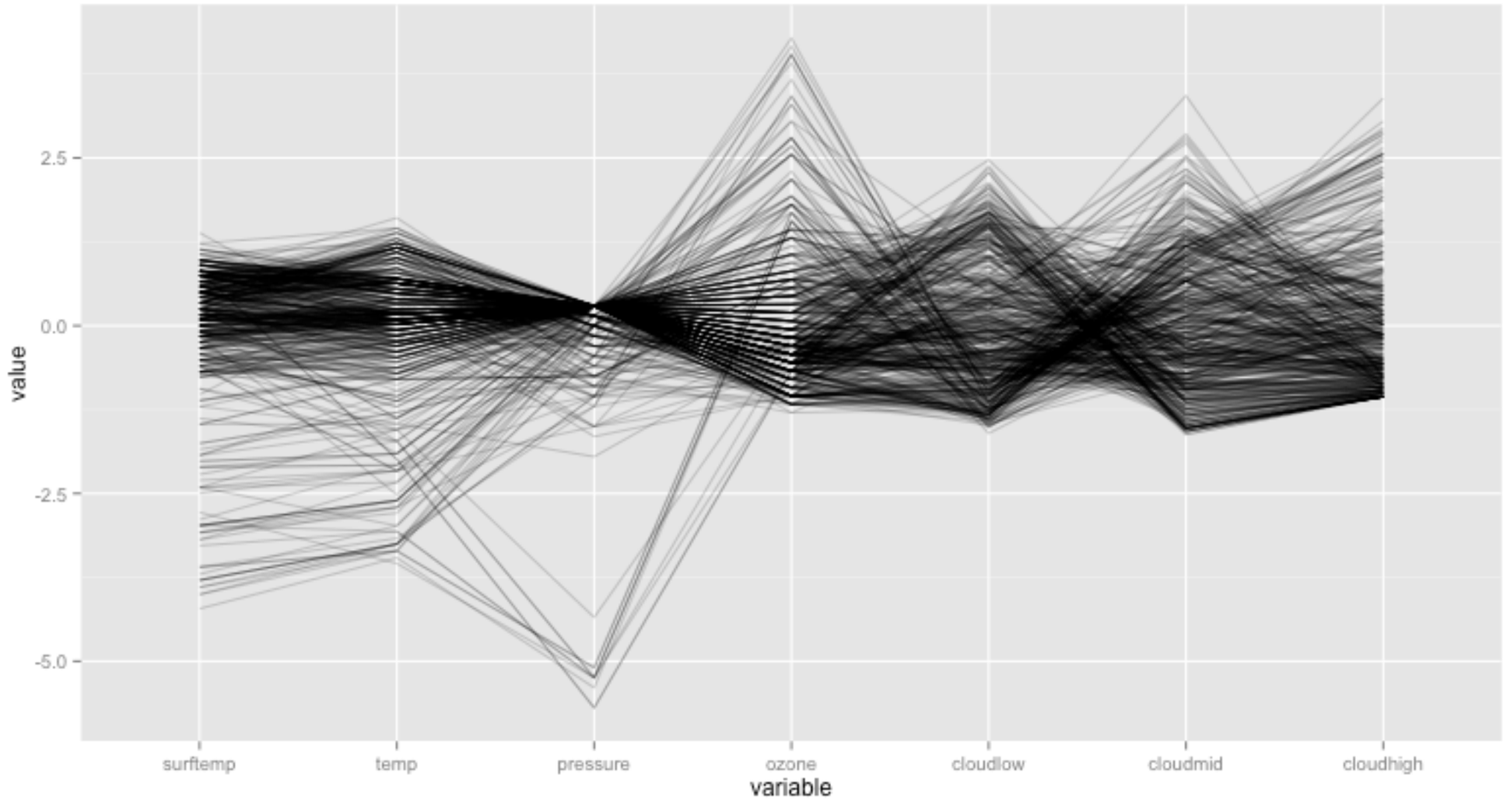
finding low dimensional structure

finding clusters that aren’t easily defined with only one or two variables

Parallel coordinates

Explain on board:

basic idea, put each variable on it's own vertical axis, draw observations in as lines.



Parallel coordinates

Lot’s of choices:

- how to scale each variable
- how to order the axes
- color?

Tours

A **projection** takes a high dimensional vector and maps it to a lower dimension.

A linear projection is like a weighted average.

Principal components is one way to find weights, but there are infinitely many others!

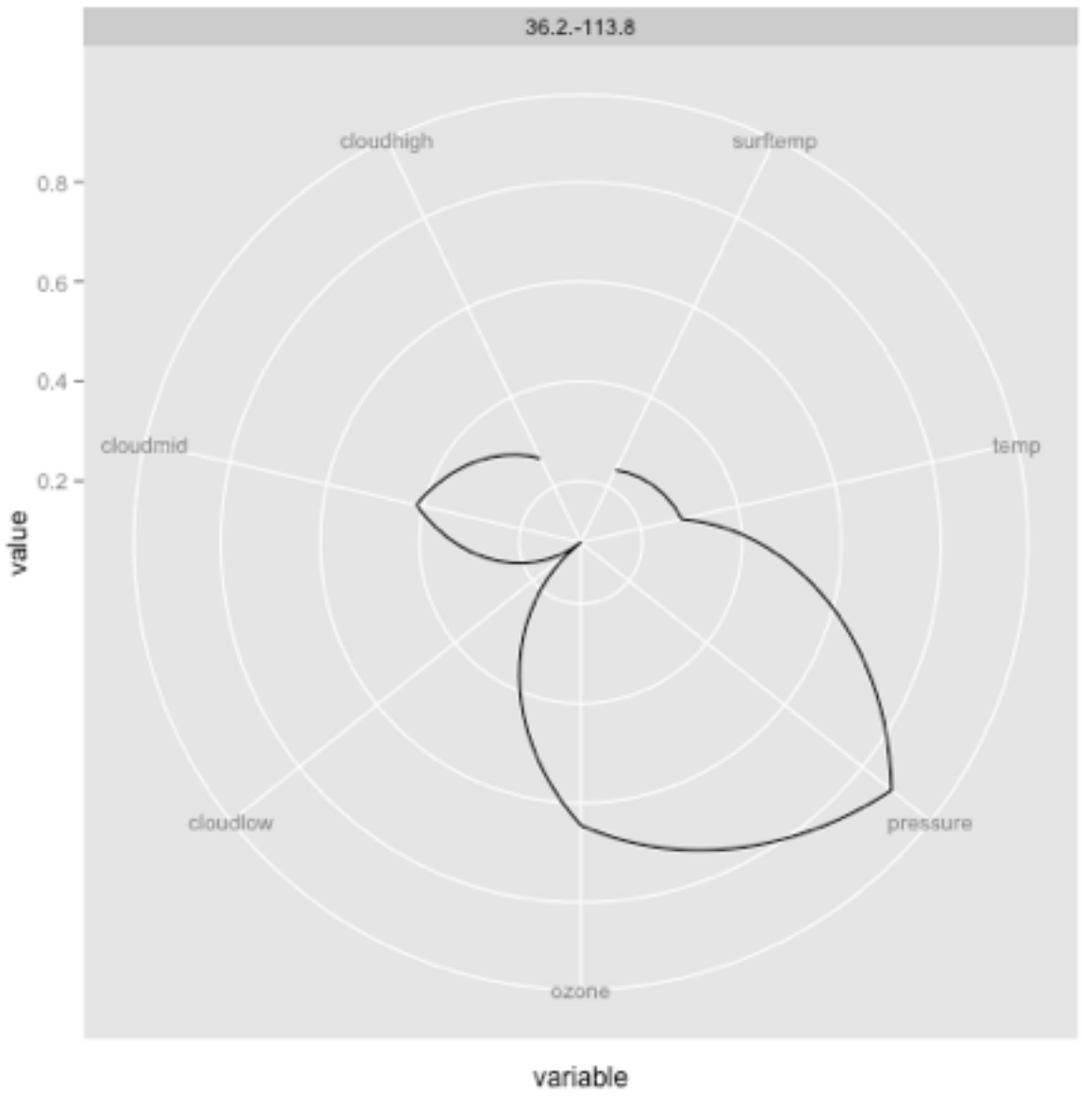
Tours take you on a tour of this infinite space of projections.

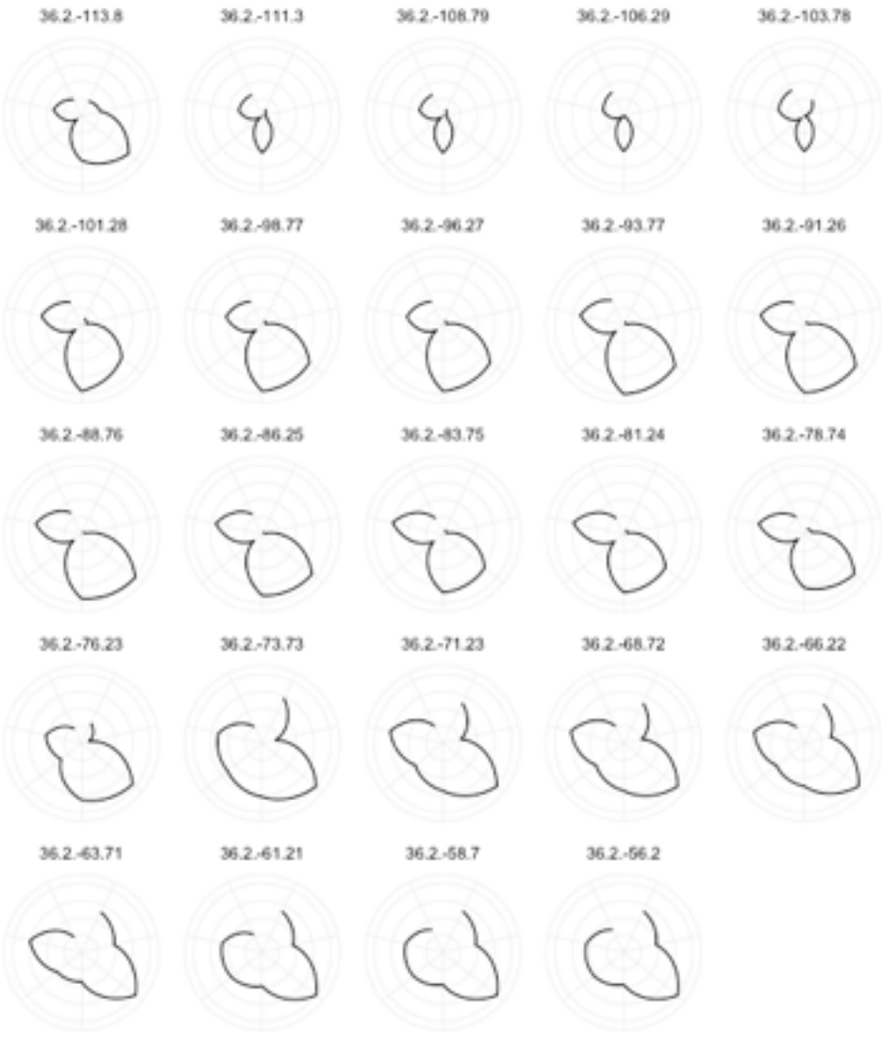
guided - try to move to a more “interesting” projection

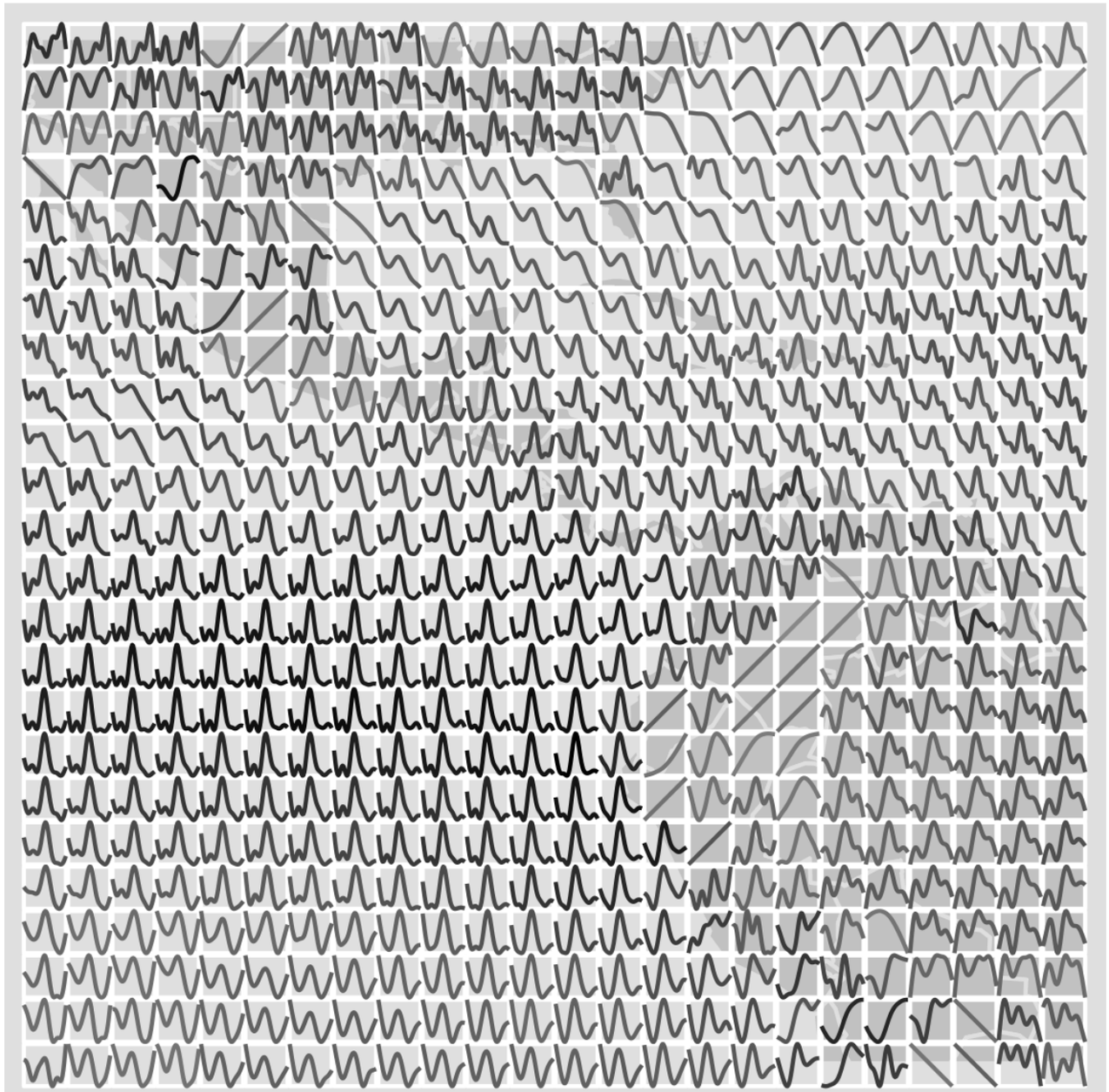
grand - move between random projections

Glyphs

Use a little plotting symbol for each location that captures all the variables.



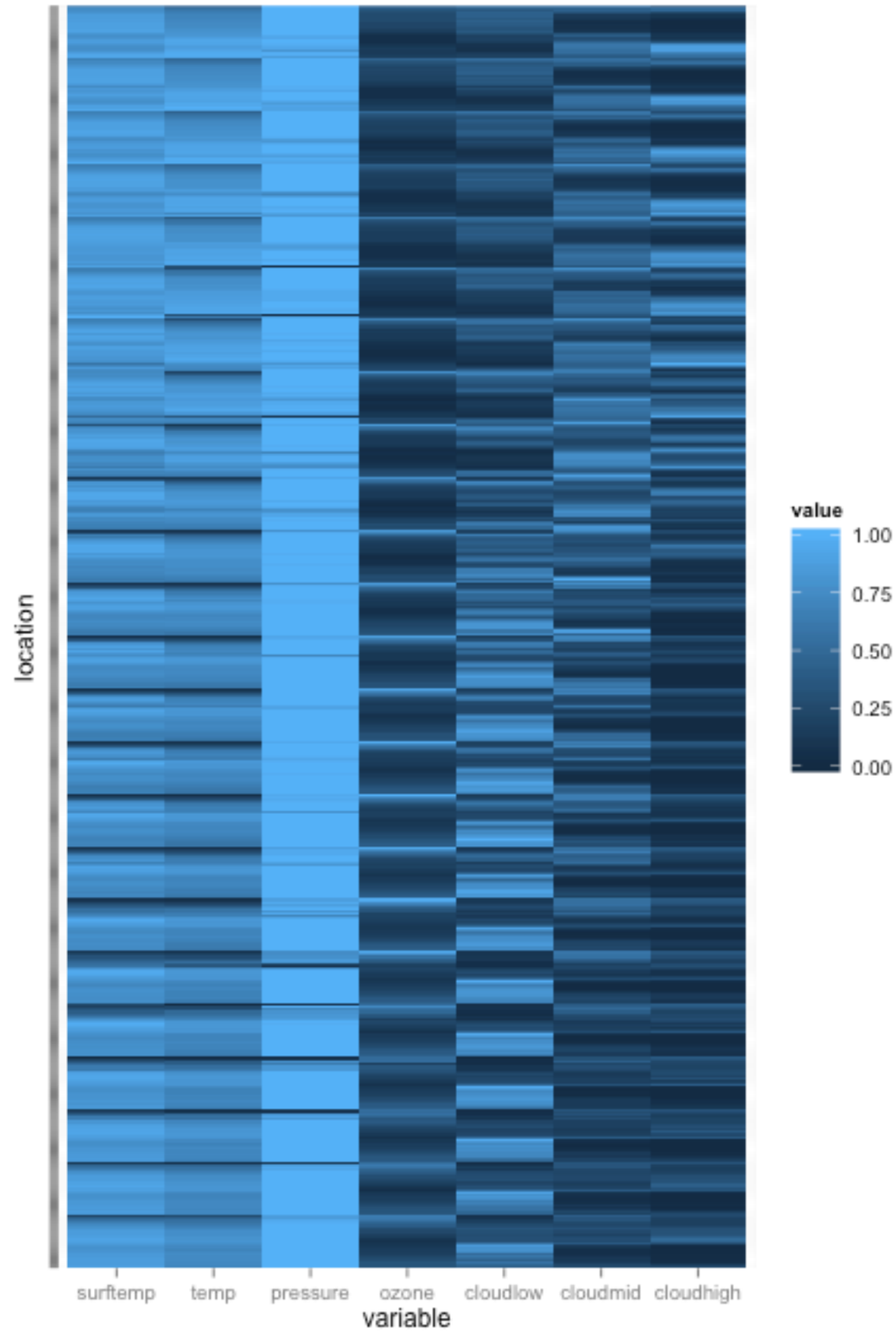




Each glyph here is at the spatial location of a grid point, and is a little time series of temperature.

Heatmaps

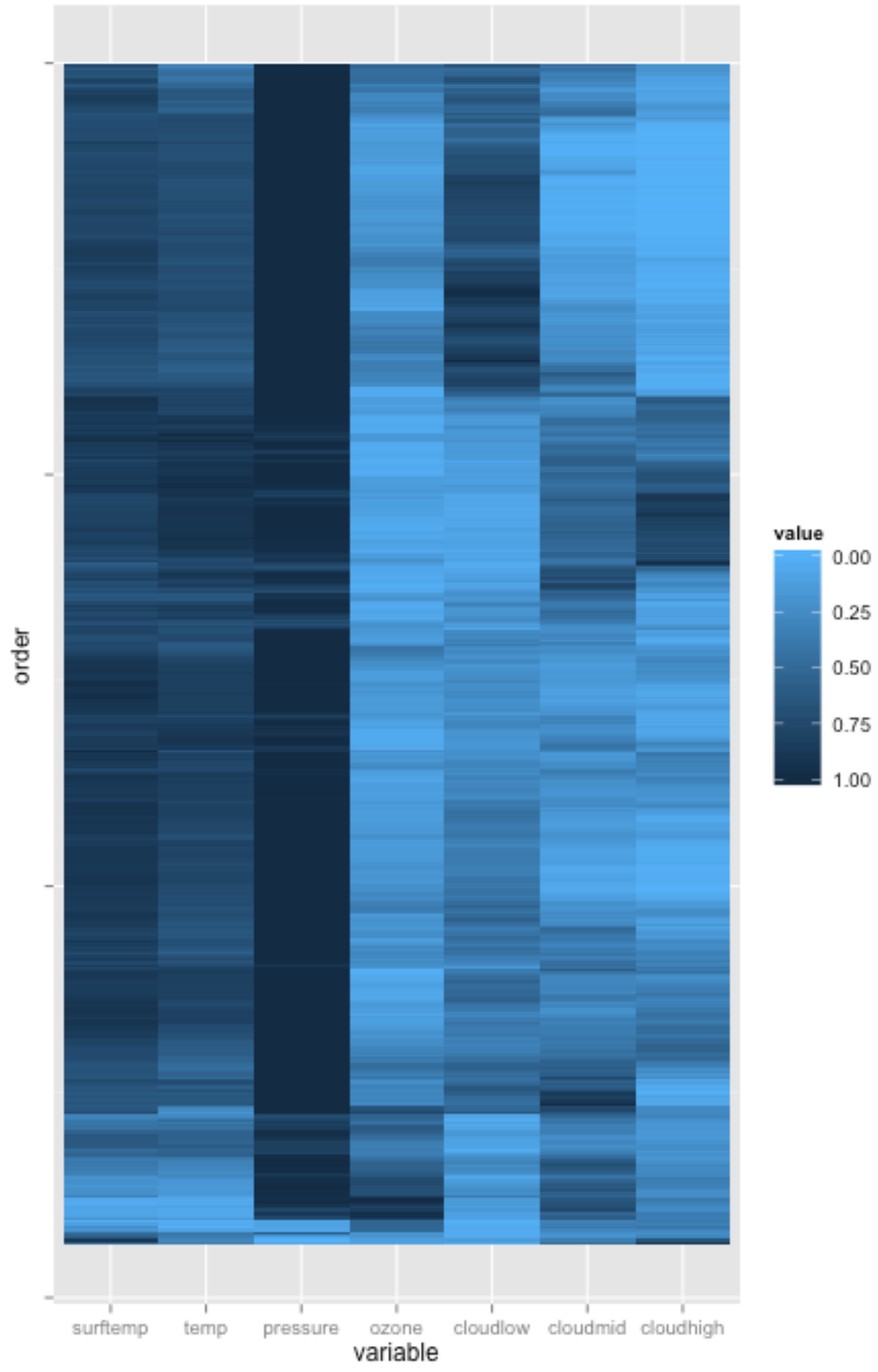
Observations in rows, variables in columns, value (scaled) mapped to hue.



Seriation

Order of observations matters for finding patterns.

Use ordering (and/or clustering) method to find an order.



A little note on maps

Most shapes will need to be drawn with `geom_path, or geom_polygon`

each observation (row in your data frame) is a coordinate,

- group aesthetic identifies single objects
- order aesthetic identifies order of drawing

use coord_map

`ggmap` add layers from google maps, stamen maps, etc.