Graphs and Charts

Session II

Vik Gopal

Jan 13, 2017

Introduction

Making Graphics

- Before we proceed, it is important to remember that the process of making a graphic is highly iterative.
- The graphics that we finally choose to present to our reader were not created at the first attempt.
- As we proceed through an analyses, we go back and forth between asking questions and making plots.
- Even to answer a single question, we may end up making several plots, adjusting point size, colour, font, etc.

The Purpose of a Graphic

The final one(s) that we choose to display, however, should *reveal* something about the data.

From Exploratory Data Analysis, by John W. Tukey

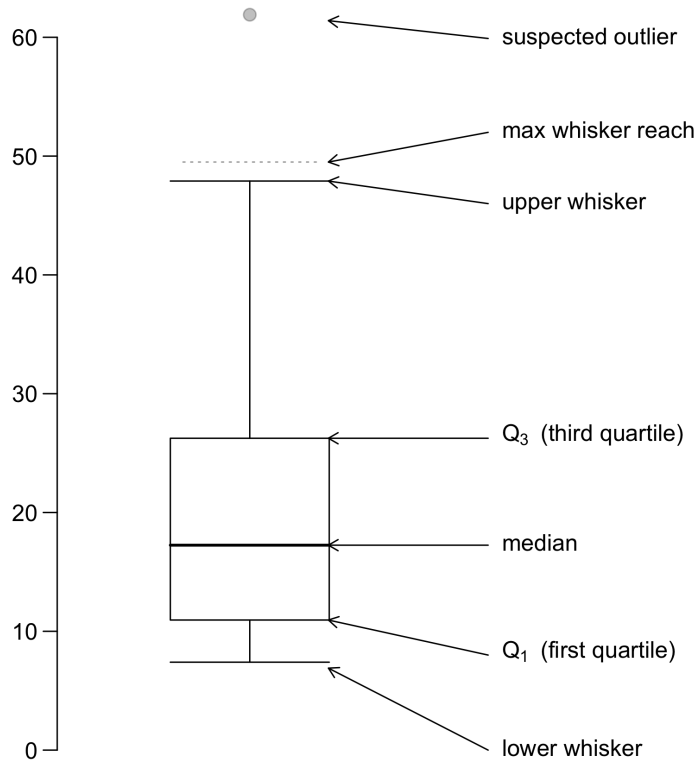
> **The greatest value of a picture** *is when it forces us to notice* **what we never expected to see**.

We may not always achieve this, but we should strive to do so, if only to save our reader the time of looking through plots that depict what we already know.

Boxplots

- Boxplots are a visual representation of 3 of the five numbers in the five-number summary.
- They identify the median, lower and upper quartiles, and suggest which points could be **outliers**.
- An **outlier** is an observation that is very different from the majority of the data. An observation is defined to be an outlier if it falls more than $1.5 \times IQR$ below the lower quartile or more than $1.5 \times IQR$ above the upper quartile.

Ingredients of a Boxplot



- On the left are the elements of a boxplot.
- The distance from the "max whisker reach" to the third quartile is exactly $1.5 \times$ IQR.

Boxplots Versus Histograms

- A boxplot does not portray certain features of a distribution, such as distinct mounds and possible gaps in the data.
- If a distribution is indeed unimodal, then a boxplot does give an indication about the skew of a distribution.
- Boxplots are useful for identifying potential outliers, and for comparing groups with respect to their "center" and "spread".
- I find it meaningful to label the outliers (if there aren't too many). This provides context to the plot, and puts a name to the extreme observations.

Describing a Boxplot

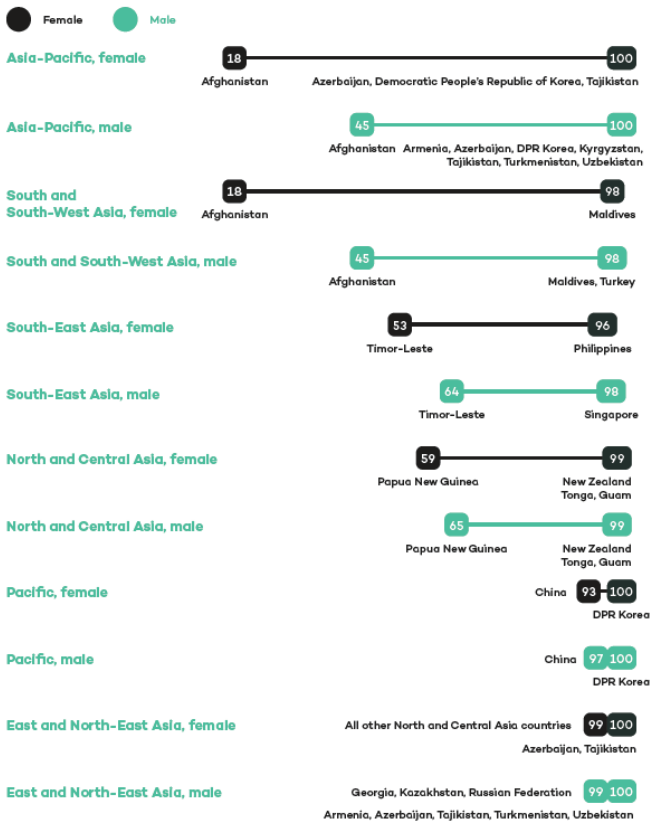
When presented with a boxplot, a good guide is to mention the following:

- Report the median as accurately as you can.
- If there are outliers, mention how many there are, and on which side of the median they are.
- If you are presented with more than one boxplot, try to compare their medians and inter-quartile ranges.

In addition, of course, please note anything that stands out in your data.

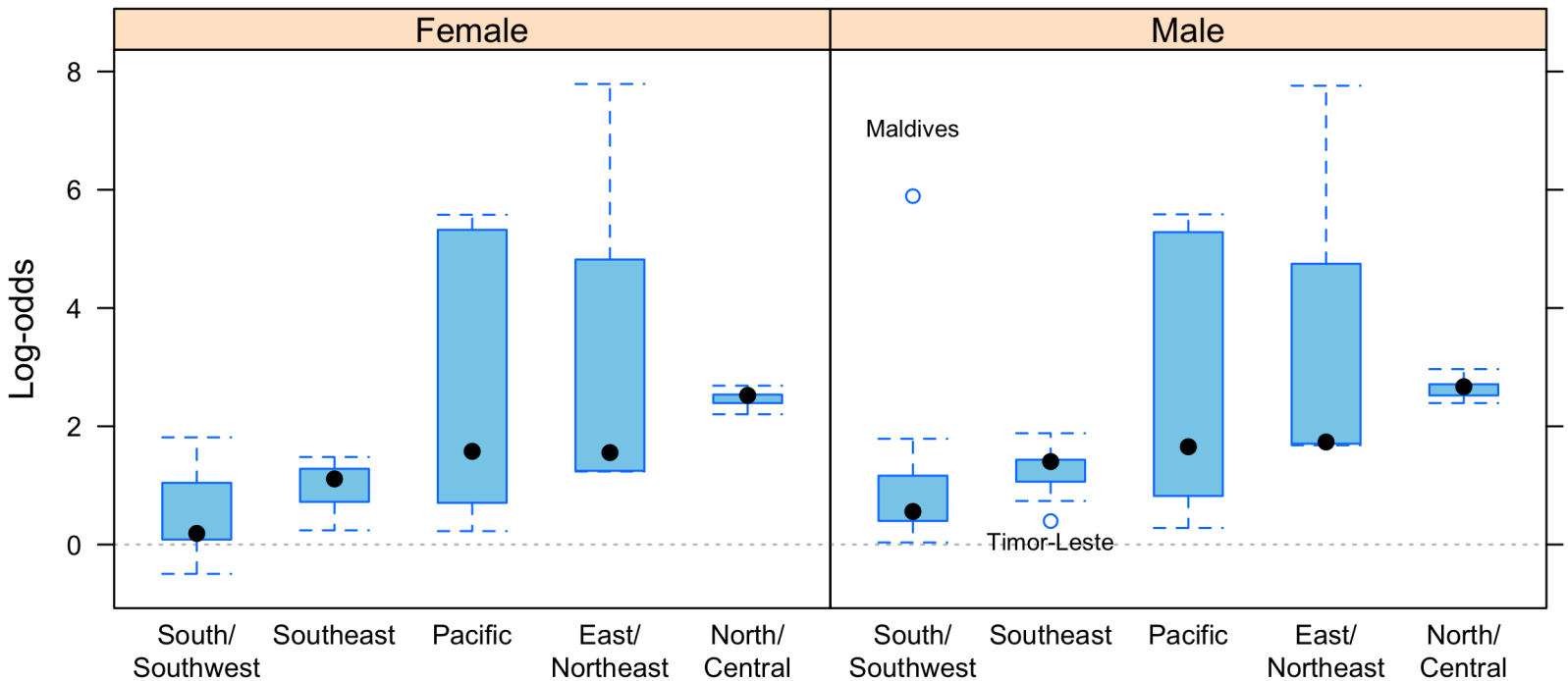
Literacy Rates

Example 1 (Literacy Rates)



- The graphic on the left was taken from *Gender Equality and Women's Empowerment in Asia and the Pacific*, UNESCAP, Figure 10.
- It displays the range of literacy rates (proportions) for the different regions and genders.
- It shows that different regions have different spreads.
- Let us use a boxplot to show what is happening in the middle of the distribution for each region as well.

Literacy Boxplots



Literacy Boxplots Construction Details

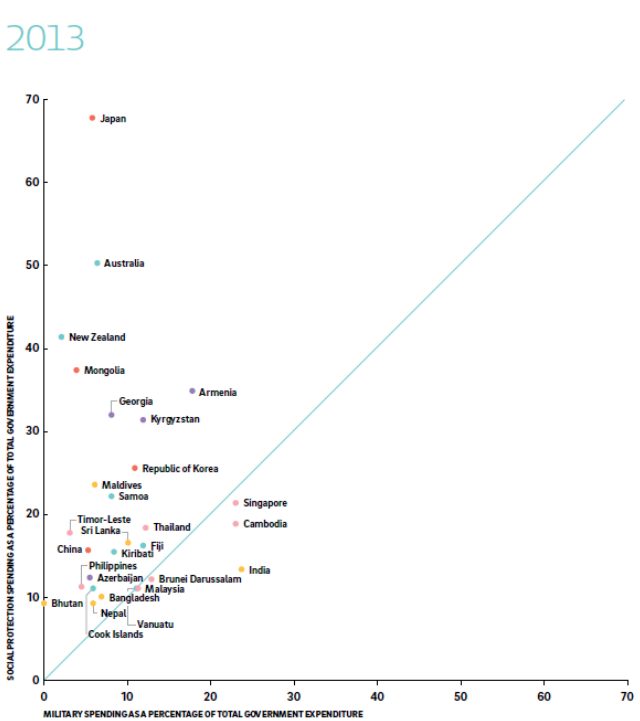
- The following variables were retrieved/derived from the UNESCAP database:
  - Literacy rates (by gender) for 2015.
  - Population for age group 15 and above (by gender) for 2015.
- Most countries have a high literacy rate, so we use the log odds to separate them.
- We have "started" all the counts.
- In the plot, the regions are sorted by increasing median value.
- As a reference, note that
  - Log odds of 0 corresponds to 50% literacy (dashed line).
  - Log odds of 2 corresponds to 99.0% literacy.
  - Log odds of 4 corresponds to 99.9% literacy.

Literacy Boxplots Insights

- The literacy rates for North/Central are exceptionally high. The IQR is the narrowest.
- The rates in the East/Northeast region are also very high; all countries in that region have a literacy rate of at least 99%.
- Although the median for Pacific is similar to the East/Northeast, there is a large IQR for those islands, with some countries close to 50% literacy.
- It is clear to note that about half of the females in the South/Southwest regions have a literacy rate of 50% or lower. South/Southwest regions have the lowest literacy, except for Maldives males.
- Apart from the Pacific and North/Central regions, the distribution of literacy rates for females is skewed more to the left than that for the males.
    - We have used what is known as a trellis plot to present the Males and Females side-by-side. It facilitated the comparison; more on this later!

Social Protection and Military Spending

Example 2 (Social Protection and Military Spending)

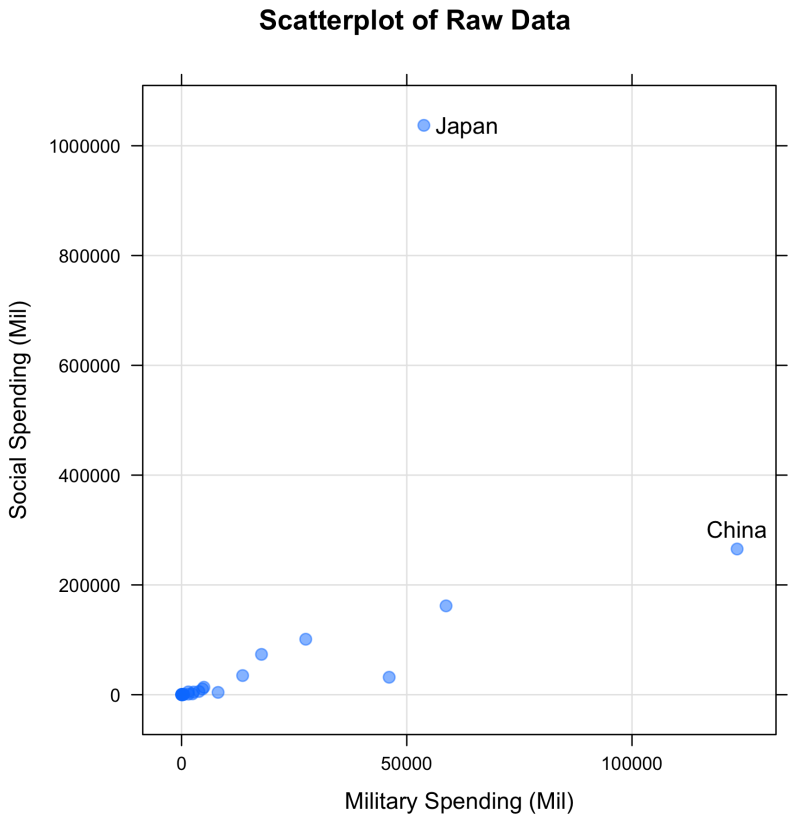


- The plot on the left was taken from *Time for Equality* report, SDD, Figure 2.6.
- It shows social protection on the y-axis, and military spending on the x-axis, both as a percentage of GDP for the year 2013.
- One of the insights from the graphic was that more than half of the countries spent more on social protection than on military in 2013.

Using data from the UNESCAP database for 2010, let us see if there is an association between

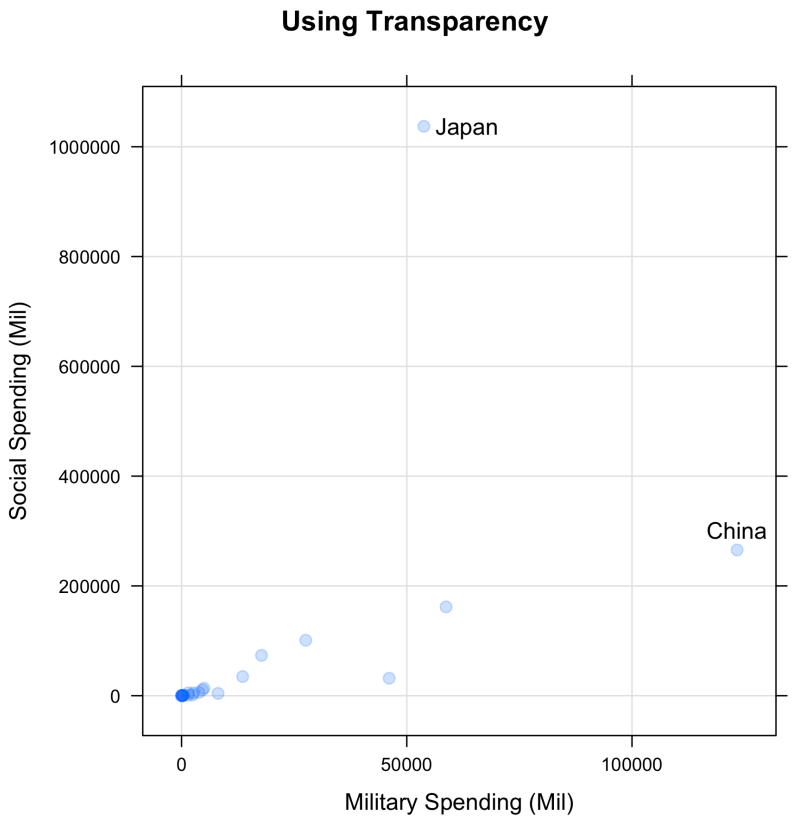
- Public social protection expenditure (including healthcare)
- Military spending (million USD)

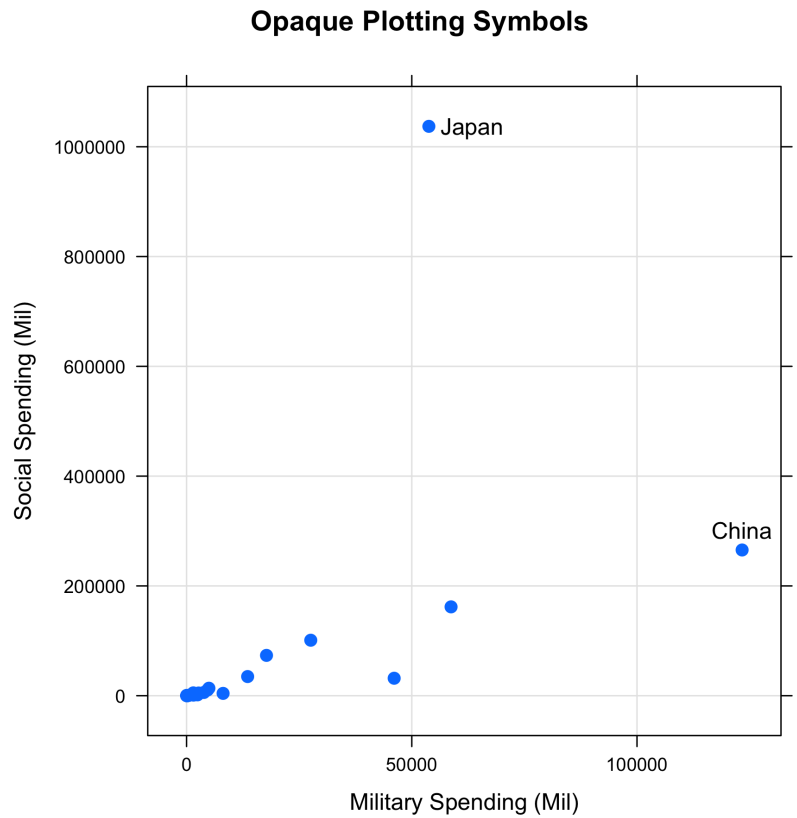
A Scatterplot of the Raw Data



- Japan spends a great deal on public social protection.
- China spends a lot on military.
- What can we say about the rest?
- As an aside, notice how there are several points bunched together at the bottom left corner? How do we deal with those visually?

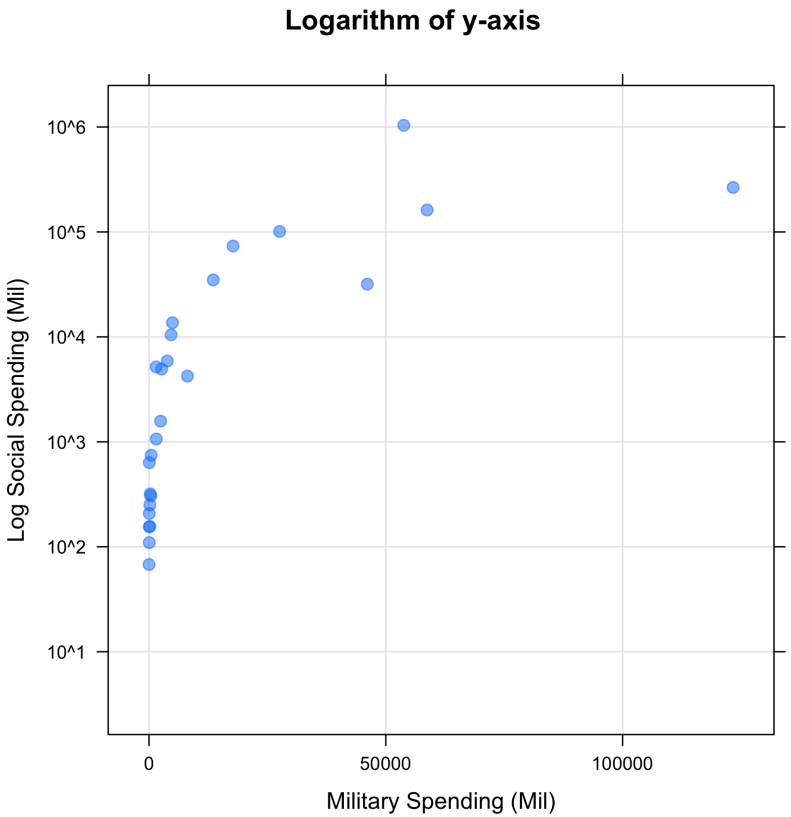
Using Transparency to Convey High Data Density





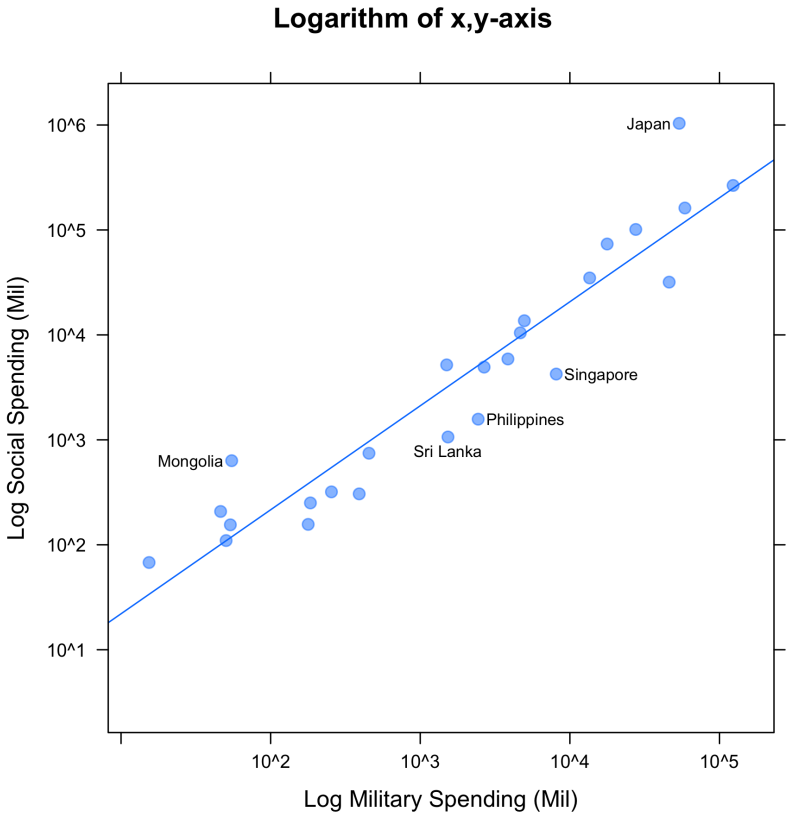
In the plot on the left, the level of transparency was such that if 5 points overlapped, it would appear as an opaque point.

Taking Logs of the *y*-variable



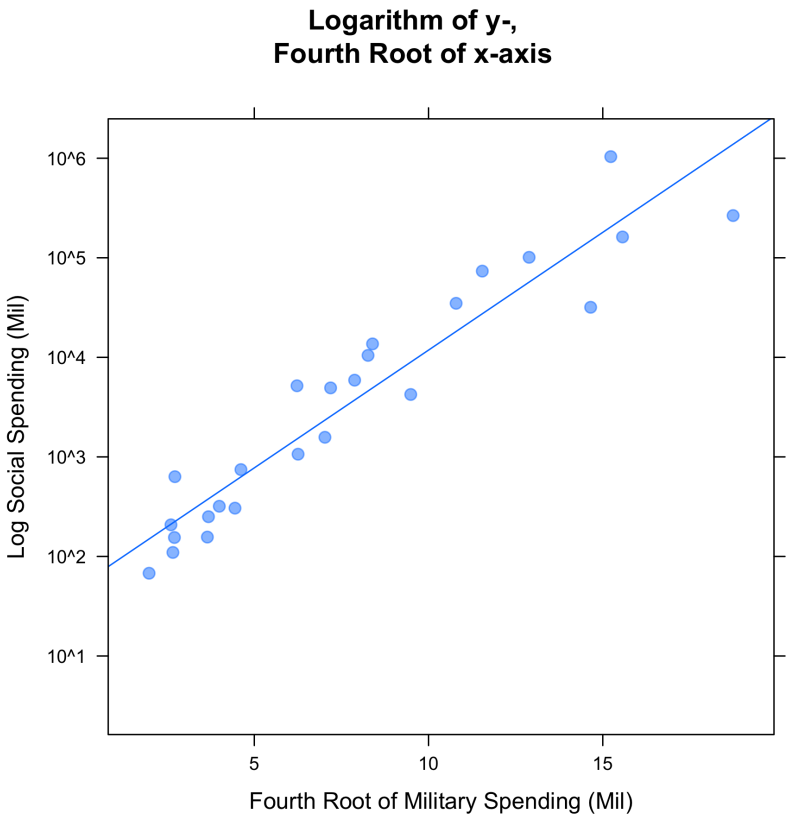
- This looks better in that we can differentiate the social spending somewhat.
  - A good guide to whether (any) transformation will be beneficial is to take the ratio of max/min. If it is greater than 3, start thinking about transforming your data.
- However, the countries with low military spending are still bunched together.
- Can we "straighten" this plot further?
- We need to amplify the differences at the lower military spending values a little more, de-emphasize the differences at higher military spending.

Taking Logs of the x- and y-variables



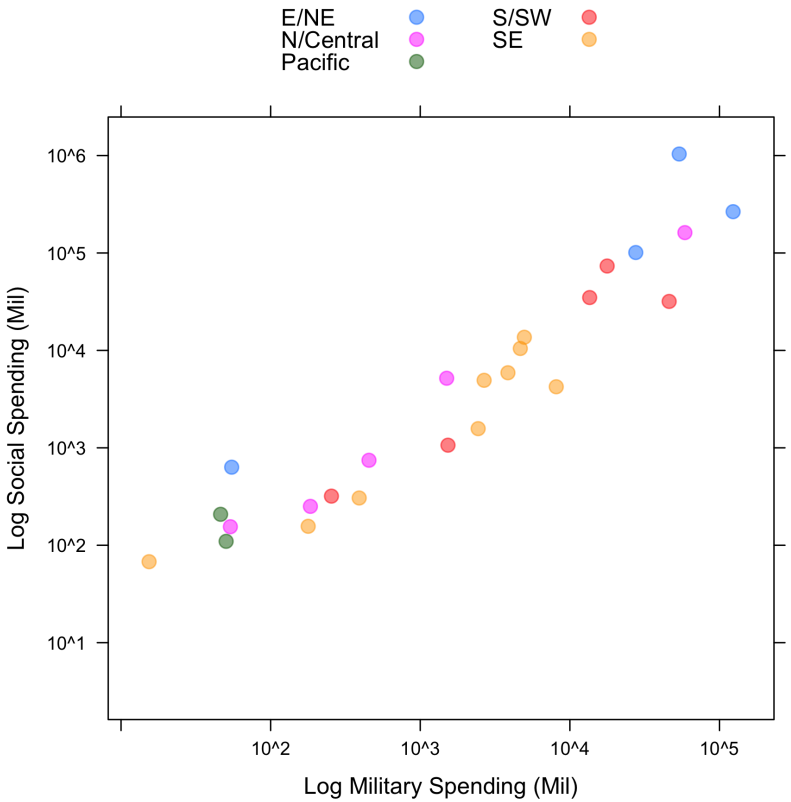
- The gradient is approximately 1, which means that every 10% increase in military spending is *associated* with a 10% increase in social spending.
- The countries that deviate the most from this guideline are
  - Japan and Mongolia (much more social protection spending)
  - Singapore, Philippines and Sri Lanka (much less social protection spending)
- Hmmm.. have we found a law?

An Optimal Line



- For those of you who noticed, it might bother you that there is still some curvature in the previous plot.
- We would do better to take $\sqrt[4]{x}$, but it is harder to interpret the relationship.
- Let's stick with the log-log plot for now, and incorporate the regional information.

A Final Plot?



- There does not appear to be any clear groupings or clear patterns that arise by colouring the points according to their regions.
- Hence, the best plot to summarise this data is probably the earlier one, on slide 23.

Why Smooth?

- Sometimes, it is not appropriate to search for a straight line relationship.
- Sometimes, we need to filter out the noise in the data, so that we can see the "general" pattern in the data; this pattern may not always be a straight line.
- When we look at a smooth, we are looking to show off the general behaviour, not the details. We do not want the details to overshadow the general behaviour.
- Of course, once we have a smoothed version of the data, we can peek under the hood at the residuals.

Moving Averages

- There are numerous techniques to smooth a sequence of data.
- We shall describe and demostrate one of the most computationally simple techniques.
- The main paradigm to smoothing is that information from nearby points can be combined together. We shall see this once again in section on spatial smoothing.
- The technique we shall discuss here is moving averages.

Computational Details

Example 3 (Moving Average, Order 3)

Suppose we have a sequence of observations at 8 time points. The MA(3) filter replaces every observation with the mean of the three nearest observations.

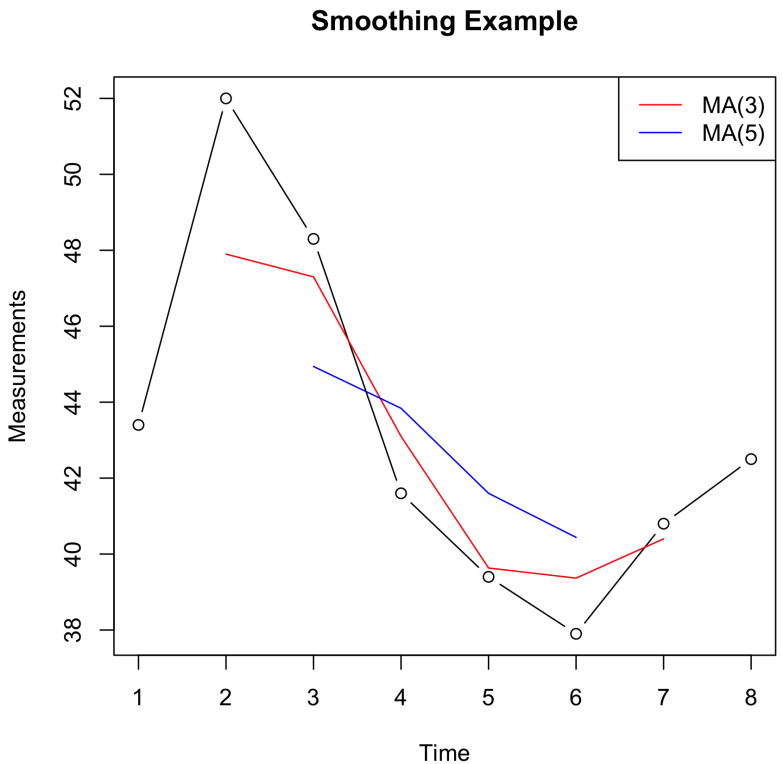
Time point	1	2	3	4	5	6	7	8
Raw data	43.4	52.0	48.3	41.6	39.4	37.9	40.8	42.5
MA(3)	–	47.9	47.3	43.1	39.6	39.4	40.4	–

For instance, at time point 2, the computation is

$$\frac{43.4 + 52.0 + 48.3}{3} = 47.9$$

It is easy to see how to compute MA(5), MA(7), and so on from a given sequence.

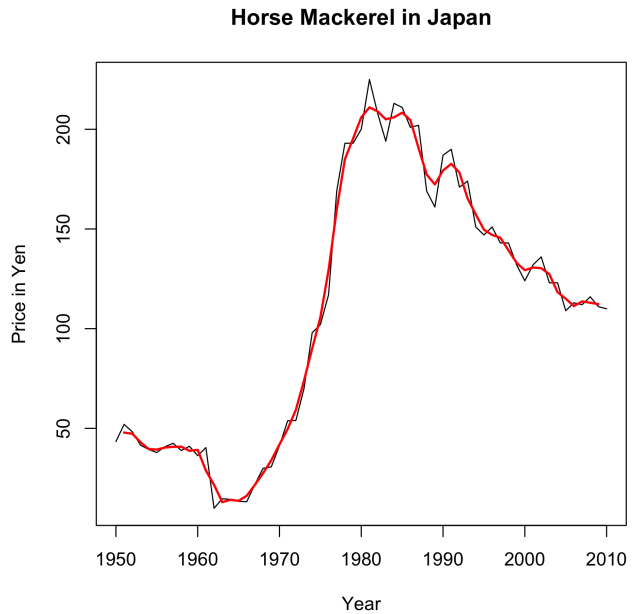
Plot of MA Examples



- Notice how we lose the information at the ends when we use more terms in the moving average.
- In addition, the amount of smoothing changes. When comparing an MA(7) to the MA(3), the former will have much fewer kinks in the plot.

Price of Horse Mackerel in Japan since 1950

Example 4 (Horse Mackerel)



- With the smoothing, it is easier to visualise the pattern of price over time.

Summary

- It is fair to ask several questions at this point:
  - How many terms to use?
  - What to do with the end-points?
  - Can we optimise some criteria to find the best smooth?
- For the purpose of exploring the data, it is best to try a few plots, and then follow what we've been doing so far – inspect those time points that deviate the most.
- Remember, all descriptions of the data are incomplete; but we hope to uncover a litte about the data each time.

Tuberculosis Prevalence, 2013

Example 5 (TB Prevalence)

- The UNESCAP statistial database provides data on the prevalence of TB in countries for the year 2013.
- The data is provided as the number of cases per 100,000 individuals in that country. We can also extract the population for the 58 countries for that year from the same database.
- There are two countries with missing TB counts for that year:
  - Northern Marina Islands
  - Micronesia

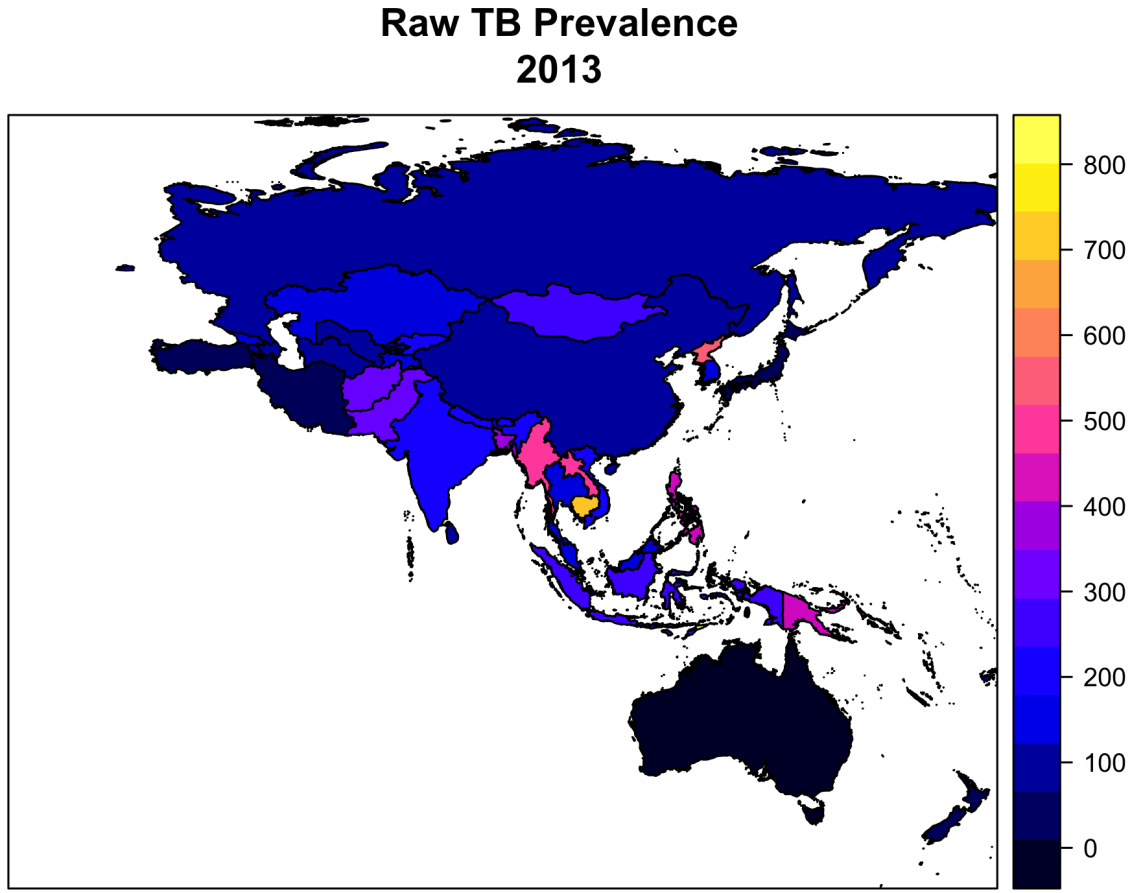
  We shall remove them from further consideration.
- How can we summarise the information in this dataset?

Working with Spatial Data

- Working with this data is unique in a few ways:
  - The data is spatially tagged.
  - It pertains to disease counts.
  - It is collected across countries with different population sizes.
- When dealing with spatial data from countries, it is customary to use a **choropleth** to represent the information.
- A choropleth is a map where areas are shaded according to a key, and each shading or colour represents a range of values.

Raw TB Prevalence

Here is a choropleth for the TB data.



Visual Improvements

We can improve on the previous plot in several ways:

- The number of shades used is a little overwhelming to discern any pattern.
- Without close reference to the legend, the choice of colours is not intuitive or indicative of severity of TB in a country.

Colour Palettes

- A colour palette is a range of colours.
- Typically these can be used in plots to distinguish between groups of data, e.g. males and females, ethnicity, etc.
- In spatial data, they are used for continuous data as well, by binning the range of continuous values and using a different colour for each bin.
- In spatial plots, it is recommended to use no more than 5 or 6 colours.

Which Palette To Use?

- Colour can be thought of as a three-dimensional concept consisting of

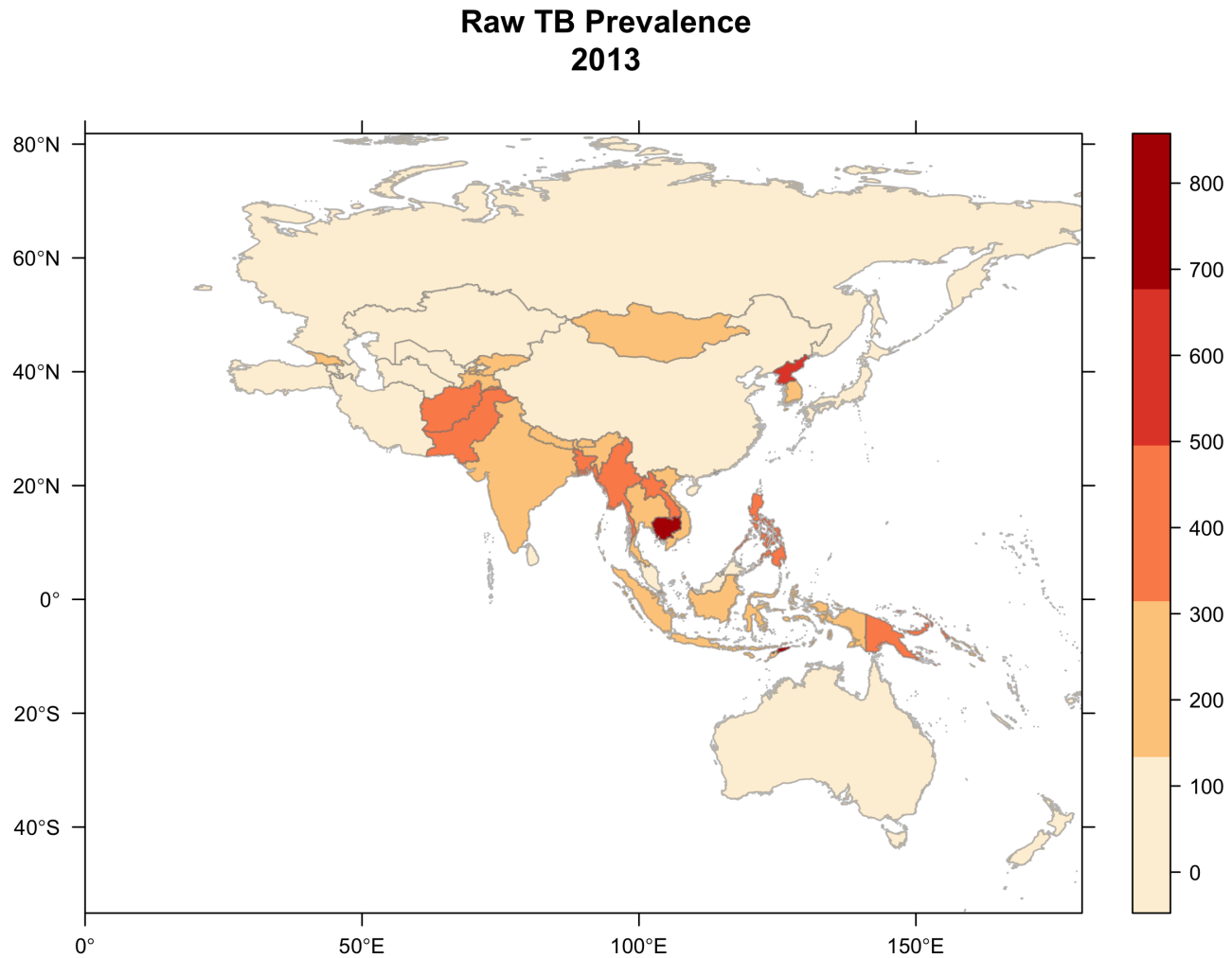
  hue: e.g. red, green or blue.
  value: e.g. light versus dark.
  saturation: e.g. dull versus vivid.

- Viewers tend to perceive *differences* between colours most readily when changing hue, and perceive *ordering* most readily when lightness (value) is changed.
- For instance, a map viewer can quickly tell that a green region is somehow different from a red region, but can more readily report that the light green region is "lower" than the dark green region.
- If we wished to convey increasing severity of TB prevalence, we would use a single hue, with several values (darker indicating more severe, and lighter indicating less severe).

Sequential Versus Diverging Patterns

- We have laid out that when we want our reader to identify which values which are higher or lower than other values, we should use sequential values. This is because lightness (values) is best suited for sequential perception.
- If instead we would like our map reader to identify ordering in two directions, then we should use a diverging pattern, with different hues.
- For instance, if we assign blue to regions having lower than average rates and red to regions having higher than average rates, then the reader can quickly separate the blue from the red regions.
- If we vary the lightness of each hue, then the readers can also order the regions with the two divergent directions.

Using a Sequential Colour Palette



Benefits of the Sequential Colour Palette

- That looks clearer!
- Light colours correspond to lower prevalence of TB.
- Dark red corresponds to higher prevalence of TB.
- We can immediately tell that Cambodia, Timor-Leste and North Korea have very high prevalence.
- SEA and South Asia seem to have higher TB prevalence than the rest of Asia-Pacific.

Choosing the Intervals

- If you look at the intervals (or breakpoints) in the legend of Slide 42, we can see that they have been chosen to be equal in width.
- Can we improve our plot with a judicious choice of intervals?

Example 6 (Choosing Intervals)

Consider a simplistic scenario, where our data had just 5 points and we wished to use **two** colours to distinguish them:

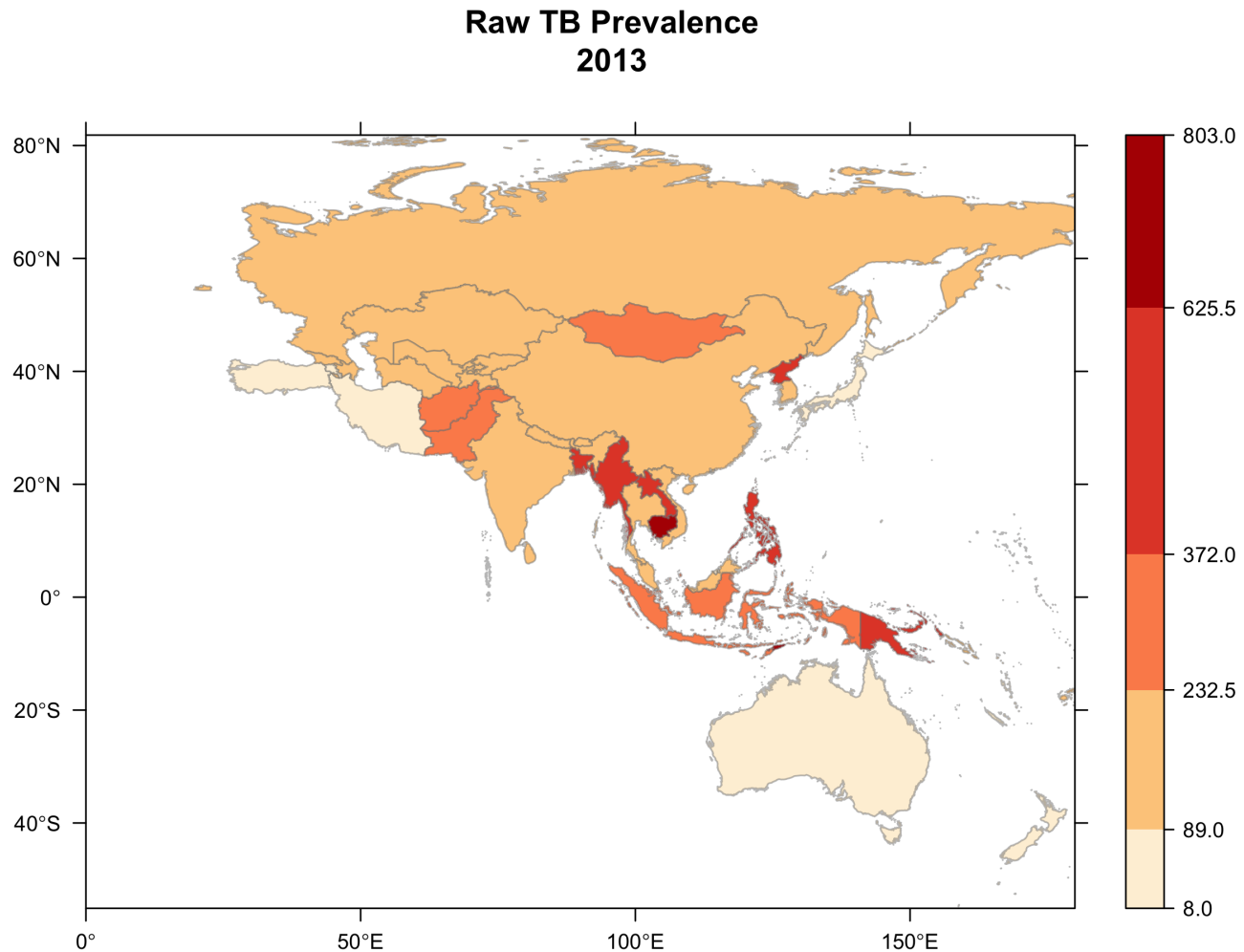
$$1,\ 1.2,\ 0.9,\ 10.2,\ 10.1$$

It would be natural to paint the 1, 1.2 and 0.9 with one colour (or shade), and the 10.1 and 10.2 with another.

We have subconsciously tried to divide them into groups such that within each group, they are *similar*. We could try to do the same with the colours in the map.

Choosing Intervals Smartly

The intervals are also sometimes referred to as breakpoints or cuts.

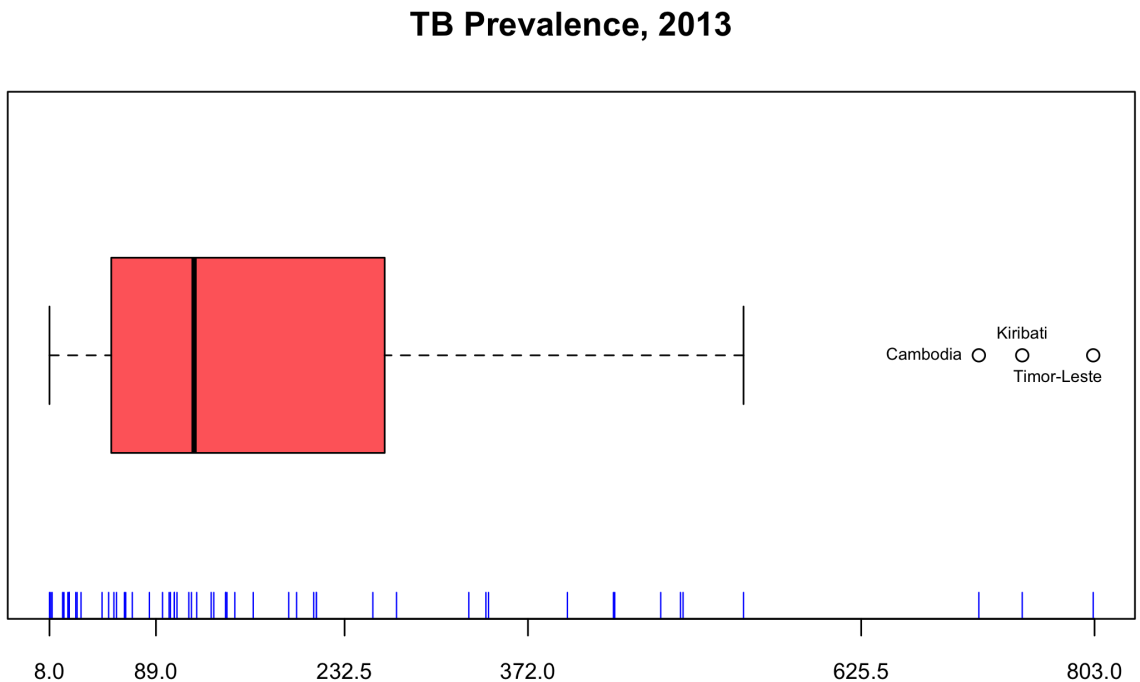


Comparing The Intervals With Before

Non-smart:	[-47.6,133]	(133,314]	(314,496]	(496,677]	(677,858]
No. of countries	31	12	9	1	3
IQR	78.5	63.5	131.0	0.0	43.5
Smart:	[8,89]	(89,232]	(232,372]	(372,626]	(626,803]
No. of countries	20	21	5	7	3
IQR	36.3	58.0	68.0	51.5	43.5

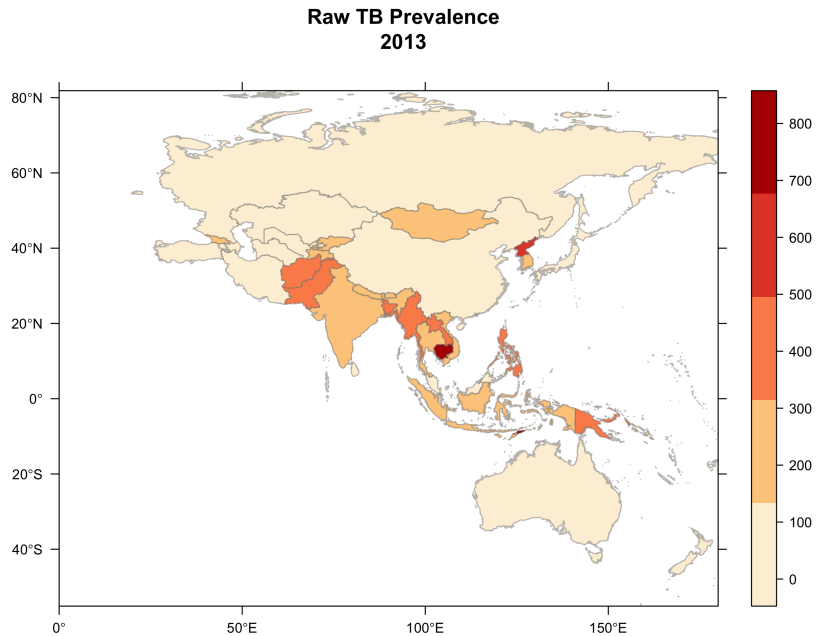
- The largest difference is in the two groups with the lowest TB prevalence.
- The formal criteria was to minimize the sum of the within-group variances, but as we can see, the sum of the IQR's has also been reduced.
- We have managed to create groups where the members are more similar to each other.

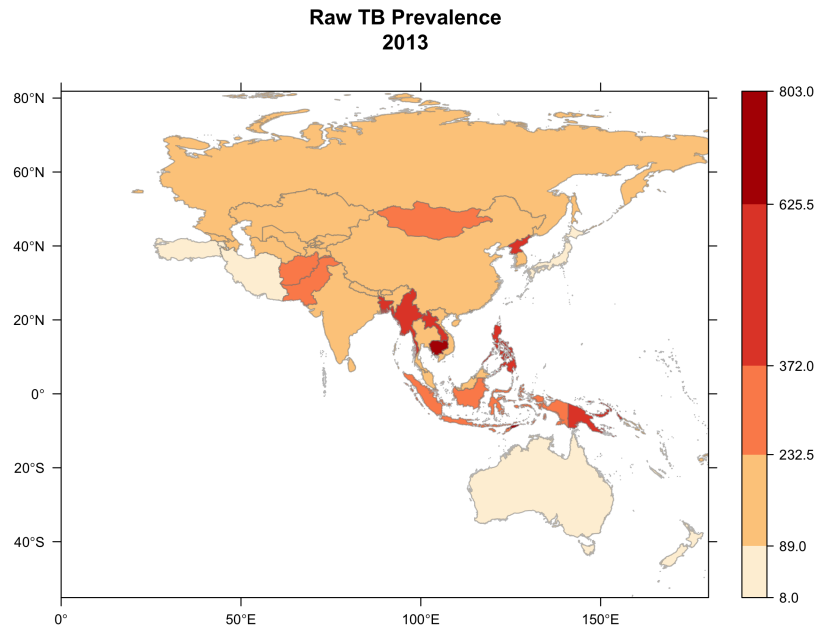
What Have We Achieved?



It is clear that we have painted regions that have similar values with the same colour.

Comparing The Pictures with Before





- Instead of being light orange throughout, the new picture distinguishes the countries with low TB better.
- Cambodia and Timor-Leste remain in the worst-hit category.
- Now that Myanmar, Lao and Phillippines are in the second-worst category, it is more apparent that SEA are a region of interest.

Summary

- Although it is a little technical, we really should stop to consider what intervals to use in our choropleths.
- There are many classification techniques that can be used for this:
  - quantile,
  - kmeans,
  - equal width intervals, and so on.
- There are times when the approaches listed above will not improve the plot much, but that should not stop us from plotting and looking.
- If we were unaware of this aspect of plots, what's there to stop a nefarious analyst from using intervals that throw all countries except one into the "severe" category?!

SIGI Categorisation

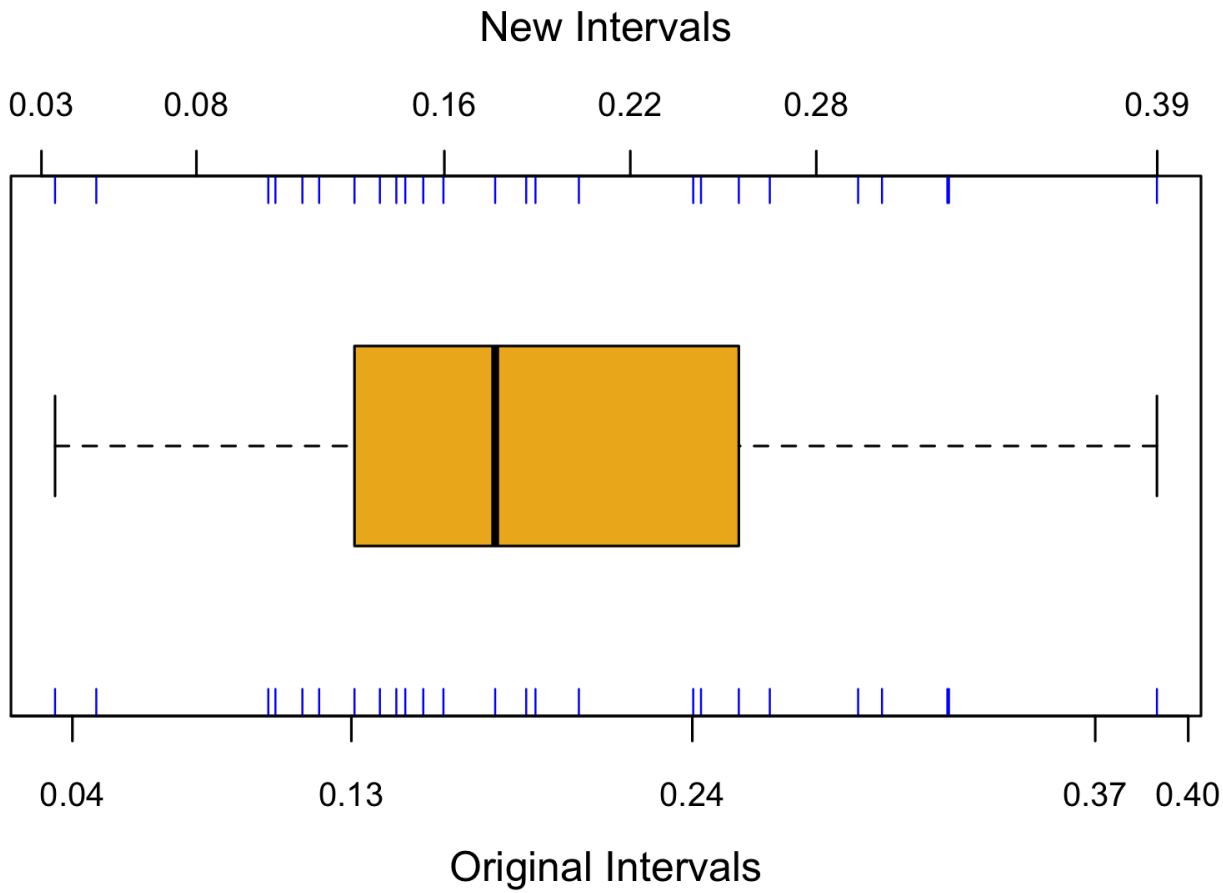
Example 7 (Social Institutions and Gender Index (SIGI) Categorisation)

TABLE 1
2014 SOCIAL INSTITUTIONS AND GENDER INDEX CATEGORIZATION OF COUNTRIES FROM ASIA

Country	SIGI 2014 Category	SIGI 2014 Value
Bangladesh	very high	0.3899
Nepal	high	0.3228
Afghanistan	high	0.3223
Pakistan	high	0.3012
Myanmar	high	0.2935
India	high	0.2650
Timor	high	0.2550
Armenia	high	0.2428
Azerbaijan	high	0.2403
Georgia	medium	0.2034
Sri Lanka	medium	0.1894
Viet Nam	medium	0.1864
Philippines	medium	0.1764
Kyrgyzstan	medium	0.1597
Indonesia	medium	0.1532
Uzbekistan	medium	0.1474
Lao PDR	medium	0.1445
Tajikistan	medium	0.1392
China	medium	0.1310
Kazakhstan	low	0.1196
Bhutan	low	0.1142
Thailand	low	0.1055
Turkey	low	0.1032
Cambodia	low	0.0477
Mongolia	very low	0.0344

- The table on the left was taken from *Gender Equality Report*, Table 1.
- It contains a classification of SIGI values into the following five categories:
  - very low
  - low
  - medium
  - high
  - very high

Re-assigning SIGI Intervals



Which is correct?

Assessing SIGI Intervals

- There is a case to be made for the original intervals.
- It lumps all the countries in the middle of the distribution to be medium, then divides the tails into very low/low/high/very high.
- The new intervals, on the other hand, attempt to identify homogeneous groups.

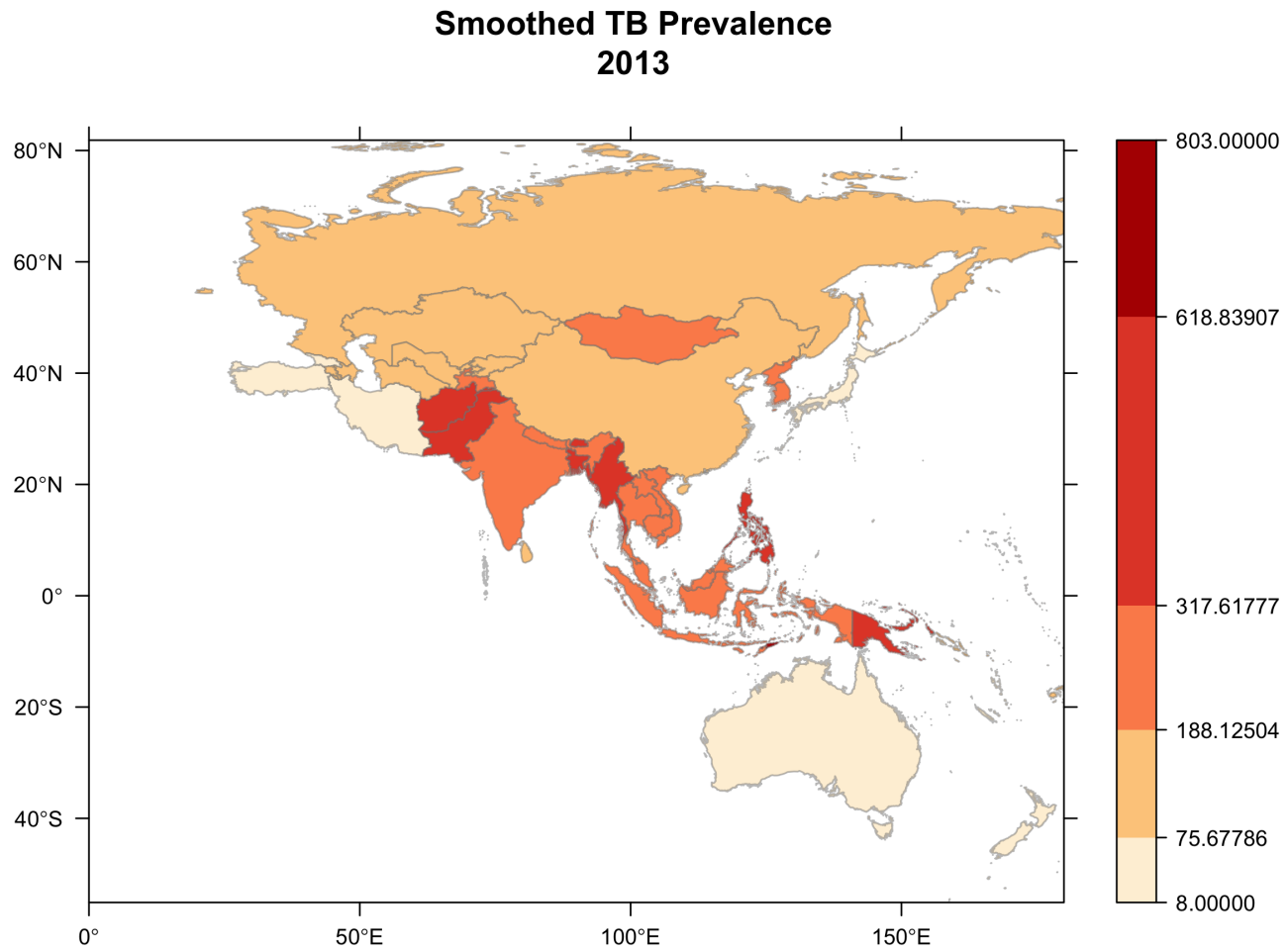
Spatial Smoothing

- We now return to the TB rates example, and discuss how we can stabilize the estimates of prevalence. Why do we need to do so?
- A country that reports a prevalence of 8 per 100,000 is not always going to be that low. Similarly, the rate for Timor-Leste may not be as high as 800 in the next year.
- These values need to be "shrunk" to something more typical. This is what Francis Galton meant when he talked about "regression to the mean".
- When smoothing spatial data, we shall be consistent with how we smoothed time series data – we shall assume that we can use information from nearby countries to improve our estimate of TB rates.
- We need to introduce the notion of the *neighbours* of a country.

What is a Neighbour?

- A neighbour of a country is one that we define to be adjacent to it.
- Adjacency can be defined in several ways:
  - Countries that share a border.
    - For instance, this would make Malaysia and Thailand neighbours.
    - Similarly, Thailand, Lao and Vietnam would be neighbours of Cambodia.
  - Countries whose capital cities are within a specified distance.
  - Countries between whom their volume of trade is above a certain threshold.
- The definition of adjacency could alter your analysis, so be sure that you think carefully about it.

Spatial Smoothing And Discerning Intervals

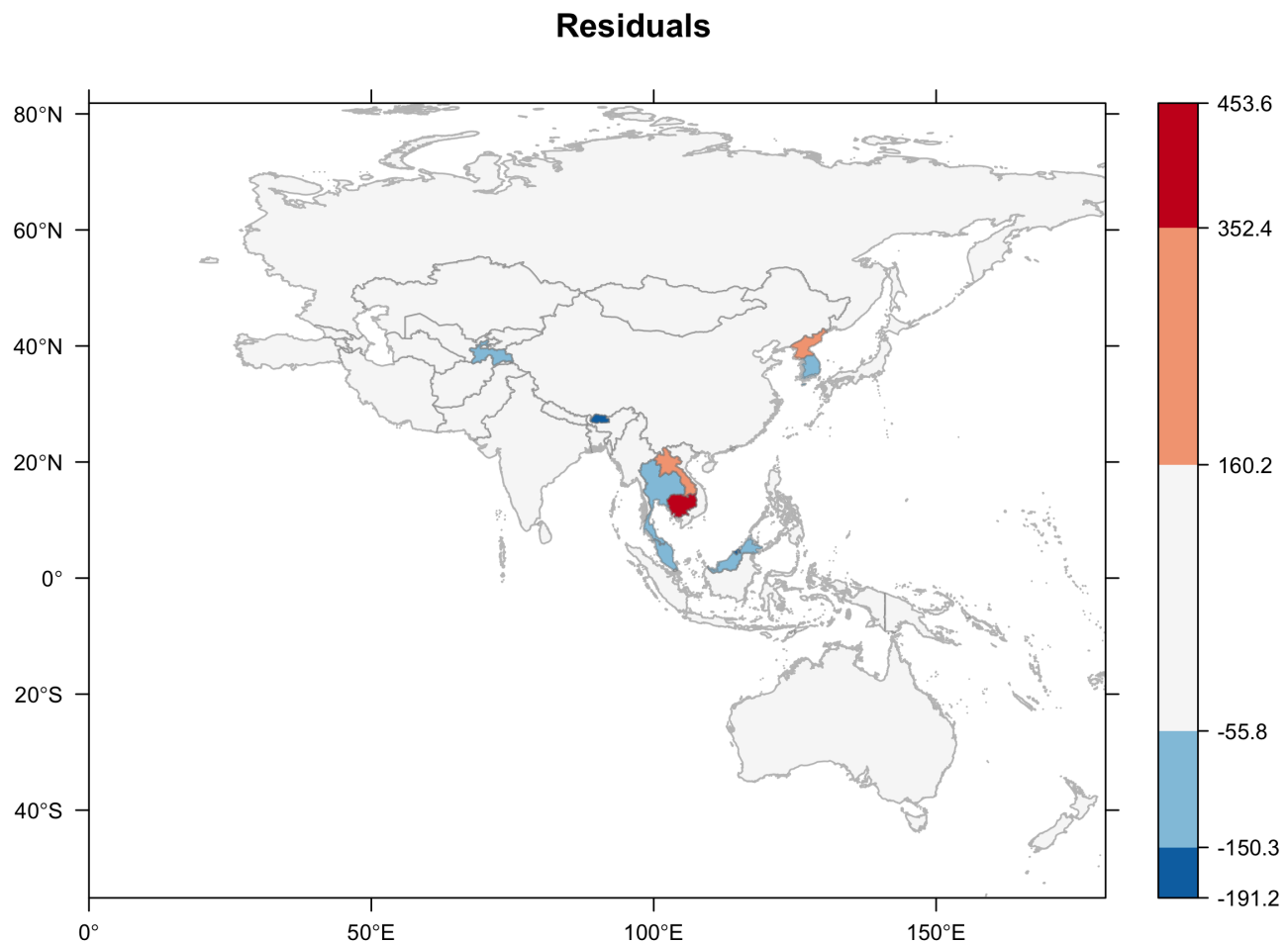


Smoothed TB Prevalence

Compare the image on slide 57, to the one on slide 46:

- The changes are more gradual; the pattern of TB rates increasing towards South, Southeast Asia is much more evident.
- The belt from Afghanistan to Papua New Guinea is clearly the area of concern.
- Now that we have a fit, let's peek under the hood.

Residuals from Smoothed TB Rates



Summary of Residuals

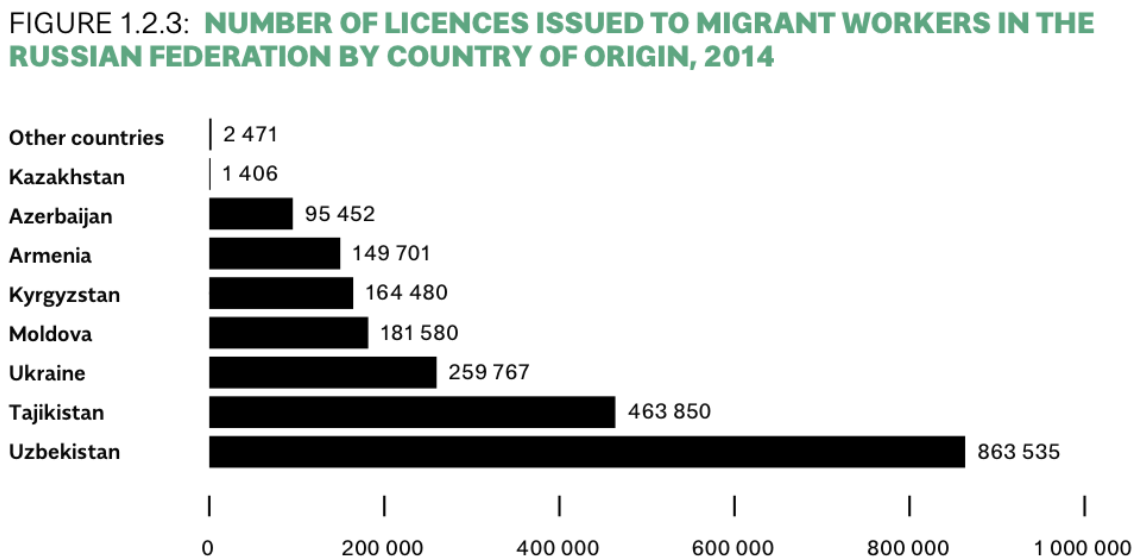
- Contrasting hues (blue and red) are used to depict positive and negative residuals.
- Recall that residuals are computed as

  $$\text{observed} - \text{fit}$$

  Hence red hues correspond to countries that had a higher TB rate than our smooth; blue hues to countries that had a lower TB rate.
- It is clear that overall, the smoothed fit is suitable for most of the countries.
- The greatest deviations are in SEA, where
  - Cambodia and Laos who have extreme TB rates.
  - Thailand and Malaysia have much lower TB rates compared to the smoothed rates for their neighbours.

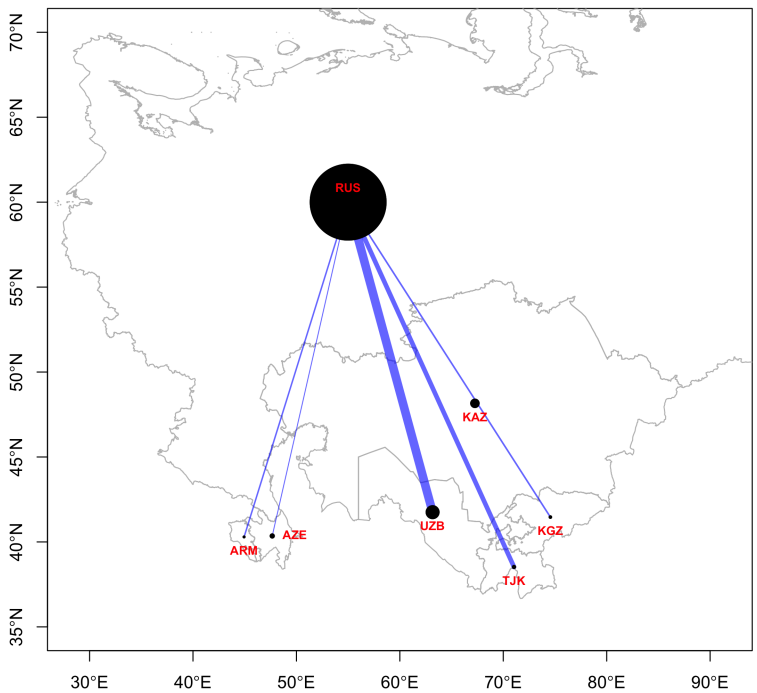
Migrant Licenses Issued by Russian Federation

Example 8 (Migrant Licenses in Russian Federation, 2014)



- In Session I, we interpreted the chart above.
- Can we provide more context to it with a spatial plot?

Spatial Version of Migrant Licenses Bar Chart



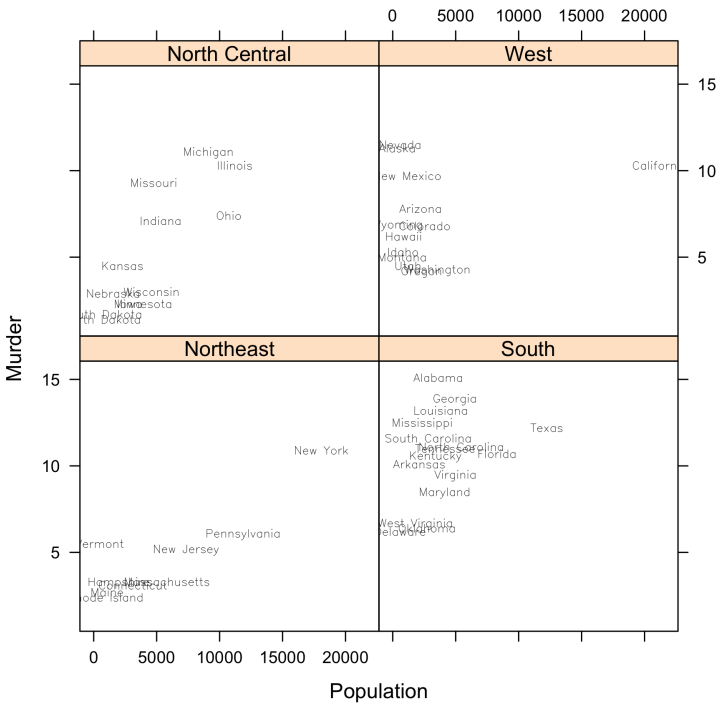
- The size of the circle is proportional to the population of the country.
- The width of the blue line is proportional to the percentage of licenses issued to that country.
- It is clearer that Uzbekistan gets the bulk of licenses, even though they are not the neighbour, and even though they do not have the largest land mass, but probably because they have the next largest population.

Introduction

- The Trellis graphics is a framework for data visualisation.
- It is a powerful and elegant system that was designed with multivariate data in mind.
- The basic idea is that we can reveal (varying) relationships between variables by conditioning on others, and creating a panel (or trellis) of plots.
- For easier comparisons, the panels would usually share the same axes.
- Colours and/or symbols are used to distinguish sub-groups within the data if necessary.

Murder Rates in US States, 1975

Example 9 (Murder Rates in 1975, conditioned on region)



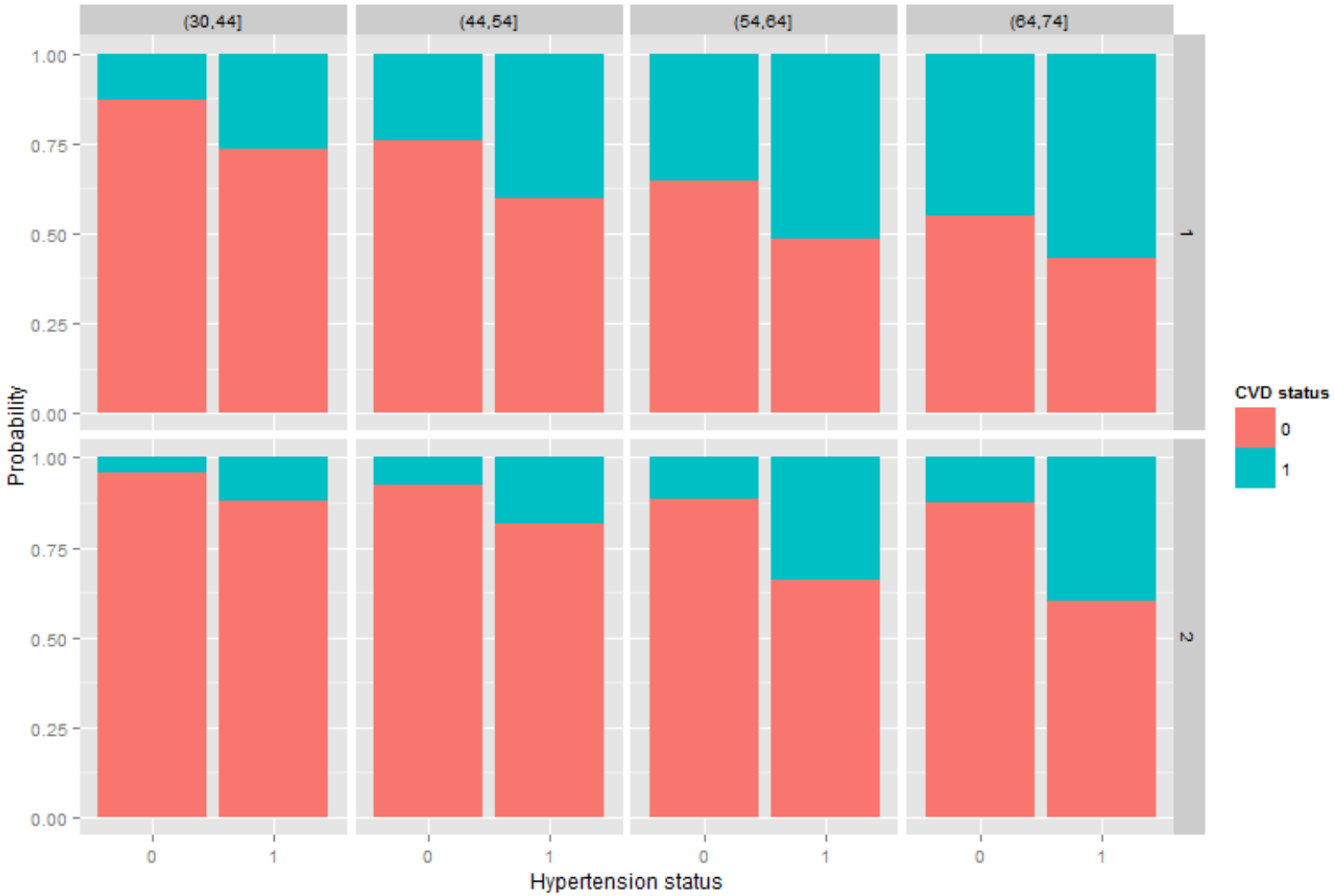
- The data contains information on the murder rates in different states.
- It is clear that the relationship between murder rate and population size is different in the different regions:
  - The relationship looks linear in the North and Northeast, although with different slopes.
  - In the West, the murder rates are above 10 only for 2 – 3 states.
  - In the South, approximately half the states have a murder rate above 10.

The Framingham Heart Study

Example 10 (Framingham Heart Study)

- The Framingham Heart Study was started in 1950 in Boston, USA.
- It involved tracking several patients and monitoring their health status with regard to heart disease.
- Thousands of medical papers have been published using data from this study.
- It is in fact responsible for the term "risk factor".
- How can we summarise the plot on the next slide, which contains information on
  - Age groups
  - Gender (Male = 1, Female = 2)
  - Cardiovascular disease status (Present = 1, Absent = 0)
  - Hypertension status (Present = 1, Absent = 0)

Bar Plots from the FHS



Migrant Stock

Example 11 (Percentage of Migrants in South-East Asia)

The following dataset was provided by Paul Tacon, UNESCAP.

	Country	1990(T)	1995(T)	2000(T)	2005(T)
1	Armenia	18.60	21.50	21.40	15.60
2	Azerbaijan	5.00	4.40	4.00	3.50
3	Georgia	6.20	5.50	4.60	4.50
4	Kazakhstan	21.90	20.40	19.20	20.10
5	Kyrgyzstan	14.20	11.10	7.90	6.10
6	Russian Federation	7.80	8.00	8.10	8.10

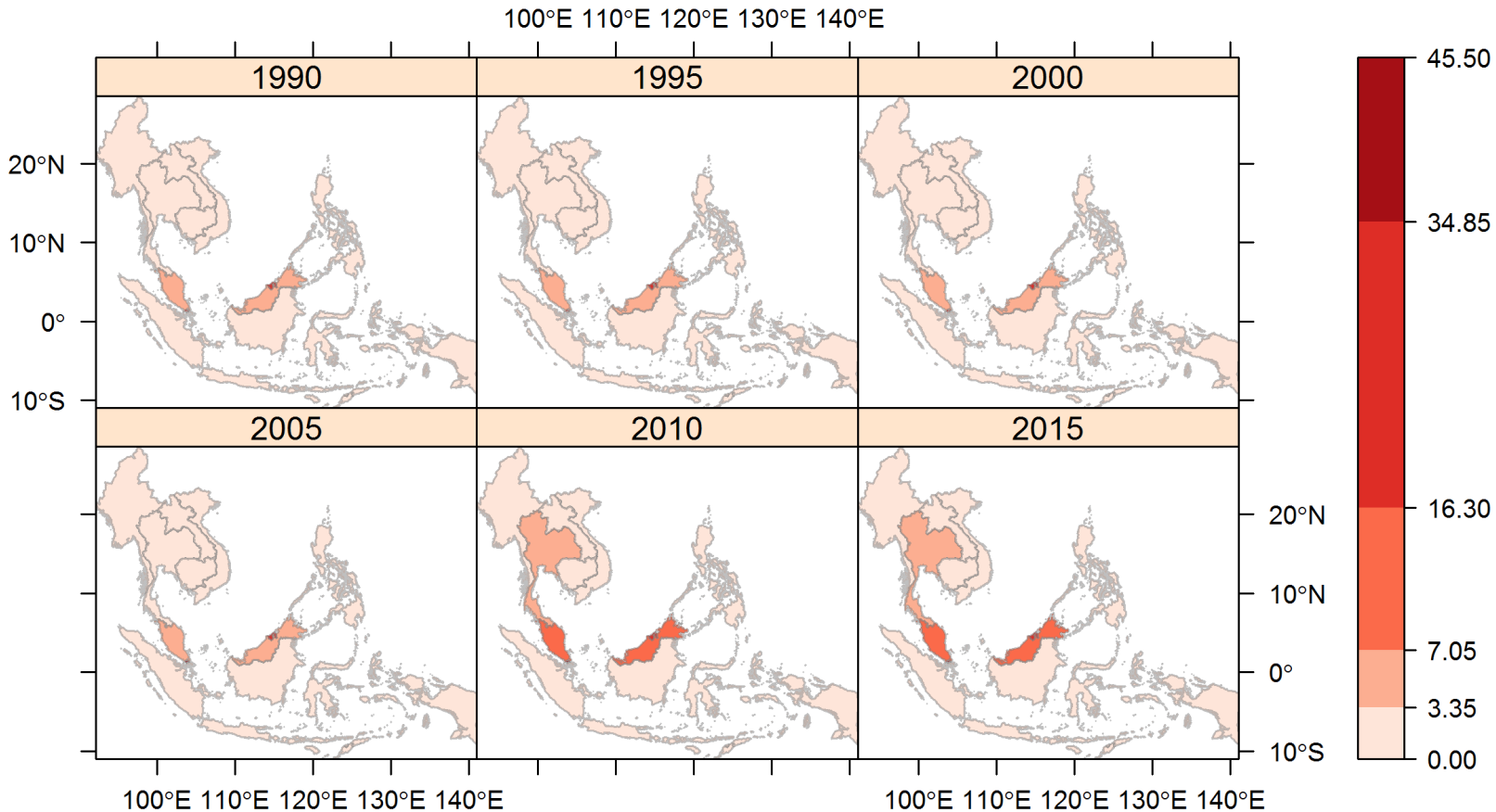
It contains information on the percentage of migrant stock (of the total, male and female populations) for the years 1990 – 2015 (in 5-year intervals).

We shall focus only on the 11 South-East Asian countries in this section.

Spatial and Temporal Aspects

- Our data contains a spatial aspect and a temporal aspect.
- We shall first use trellis plots to depict the change in percentage over time.
- We shall utilise the colour class intervals from the previous section, but we shall drop the smoothing component.
- Later, we shall drop the spatial aspect (partially) to assess if it can reveal the patterns more strikingly.

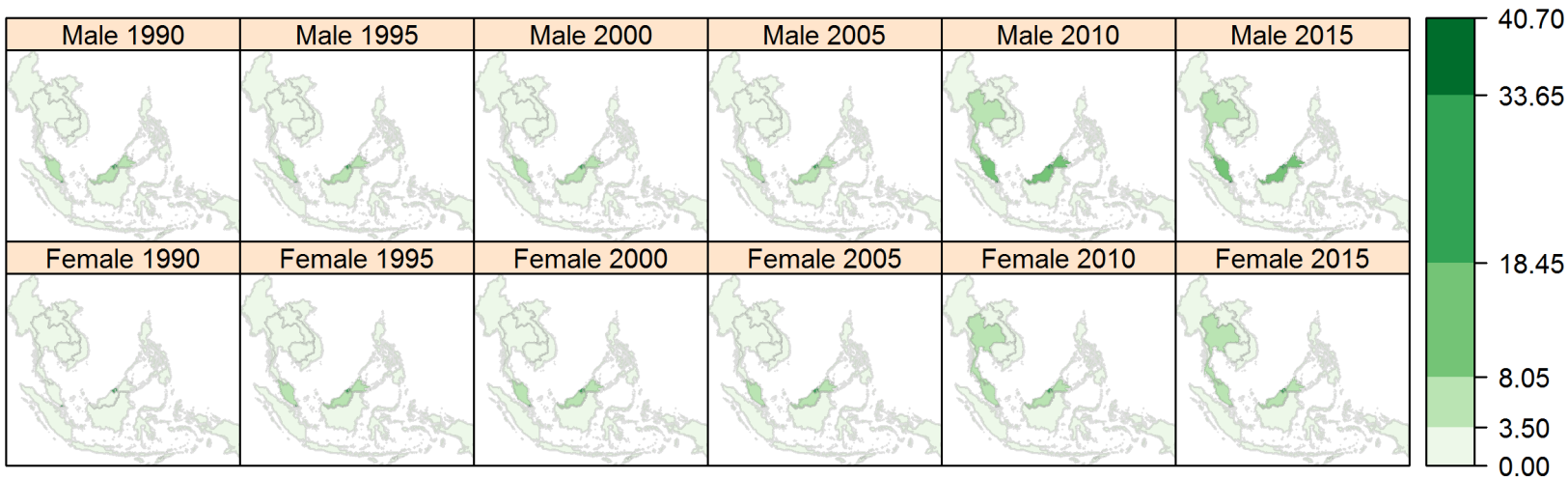
Total Percentage of Migrant Stock



Total Percentage of Migrant Stock – Observations

- For most of the countries in SEA, the percentage of migrants does not cross 3.35% over the 25 years considered.
- Singapore, Brunei and Malaysia begin with the highest percentages, and they remain so in 2015.
- In 2010, Thailand's percentage tips into the next category (7.05 – 16.30%).

Male and Female Percentages of Migrant Stock



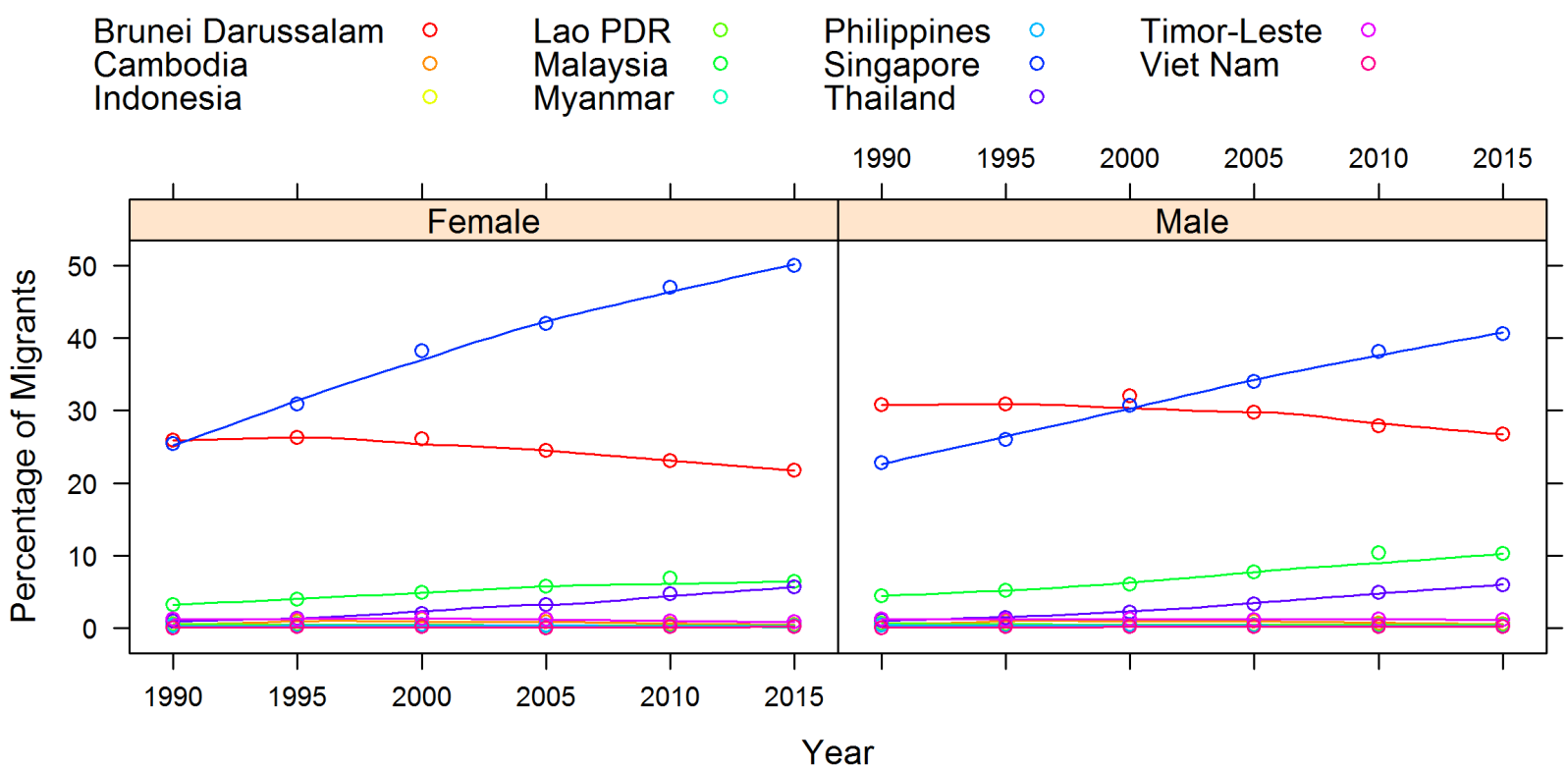
Male and Female Percentages of Migrant Stock – Observations

- The points from the previous graphic remain accurate.
- In addition, there is a point that, for females in Malaysia, the percentage of migrants starts in the lowest category, and then only increases to the next. The males, on the other hand, go on to the next higher category.

Comments

- The trellis plots on the previous slides reveal a little (but very little) about the SE Asian countries.
- However, there are several ways to improve. For instance,
  - Singapore is barely visible.
  - Since the majority of countries' percentage change quite little, perhaps we should study how much the percentages of Singapore, Brunei, Malaysia and Thailand change over time.
- We can do so by focusing on the time series nature of data that we have.

Male and Female Percentages of Migrant Stock



Male and Female Percentages of Migrant Stock – Observations

- The percentage of females in Singapore has doubled. It now stands at 50% of the female population. The corresponding percentage for males has increased by 50%. Both genders started out at approximately the same level.
- For Brunei, the percentage is decreasing very gradually.
- In Malaysia and Thailand, the percentage of males is always greater than that for females.
- In Thailand, the rate of change appears to increase after 2000.

Male and Female Percentages of Migrant Stock – Improvements

How can we improve the previous plot?

- Taking logs to de-emphasis the differences betweeen Singapore/Brunei and the rest.
- Labelling Singapore, Brunei, Malaysia and Thailand on the lines.
- Studying the residuals, i.e. the deviations from the smoothed lines.

Condensed Mosaic Displays

- Mosaic displays represent the counts in a contingency table directly by tiles whose area is proportional to the cell frequency.
- A *condensed mosaic display* is similar to a bar chart.
- The width of each column is proportional to the marginal frequency of the column variable.
- The height of each tile is determined by the conditional probabilities of the row variable in each column.
- Thus the area of each box is proportional to the cell count, and complete independence is shown when the tiles in each row all have the same height.
- A mosaic display can be generalised to more than 3 variables.

Hair and Eye Colour

Example 12 (Hair and Eye Colour)

Consider the following table, which contains counts of the hair and eye colour of students.

Eye \ Hair	Black	Brown	Red	Blond
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

Any pattern to the counts is not clear from the table.

Hair and Eye Colour

Example 12 (Hair and Eye Colour)

- It is now clearer to see that most students had brown hair.
- Of the students who had red hair, their eye colours were more or less equally distributed.
- For students with black and blond hair, their eye colours were dominated by brown and blue respectively.

Enhanced Mosaic Displays

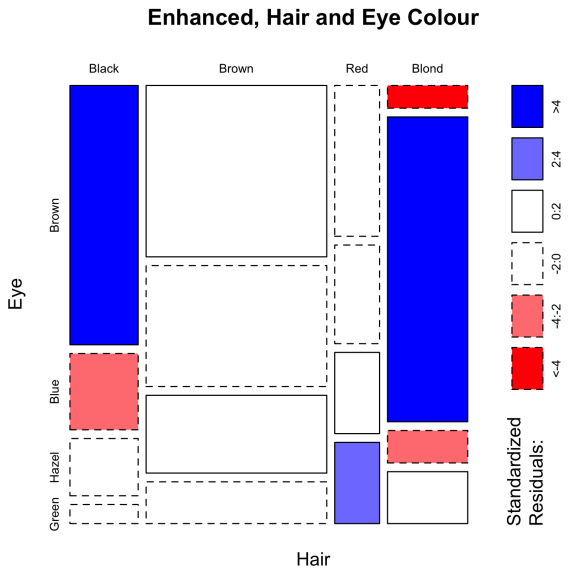
- The enhanced mosaic display uses colour and shading to reflect the sizes of the residuals from independence.
- The residuals for each cell are computed as

$$d_{ij} = \frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}}$$

  where m_{ij} is the expected cell count in the situation that the variables were independent, and we had observed the same marginal totals.
- When making the plot, cells with positive residuals are shaded blue, while those with negative residuals are shaded red.

Hair and Eye Colour

Example 13 (Enhanced Display, Hair and Eye Colour)



It is now clear that, compared to the expected counts under independence,

- there is a surplus of black-haired students with brown eyes,
- there is a surplus of blonde-haired students with blue eyes.

Migrant Workers in Russia

Example 14 (Russian Migrant Workers)

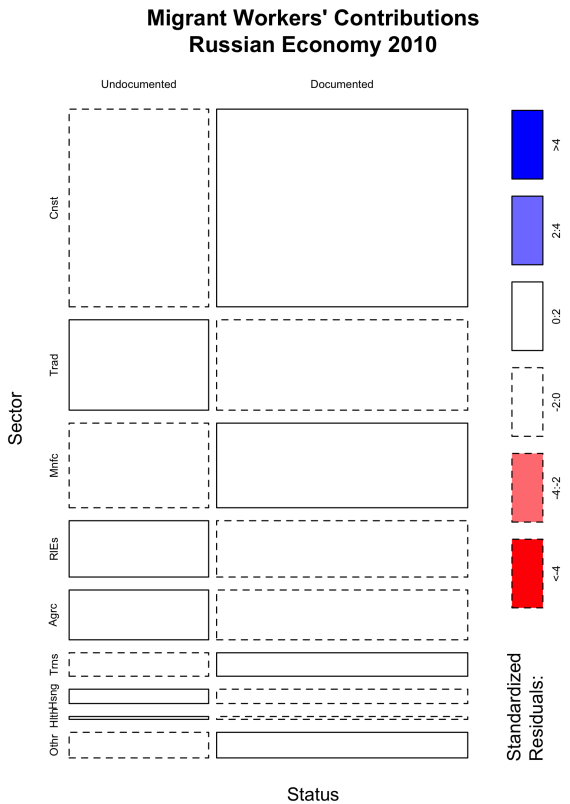
TABLE 3.1.2: **ASSESSMENT OF THE CONTRIBUTION OF MIGRANT WORKERS TO THE SECTORS OF THE RUSSIAN ECONOMY, 2010**

SECTOR	DOCUMENTED MIGRANT WORKERS	IRREGULAR MIGRANT WORKERS	TOTAL MIGRANT WORKERS (DOCUMENTED + IRREGULAR)	TOTAL EMPLOYMENT BY SECTOR	TOTAL MIGRANT WORKERS AS PERCENTAGE OF TOTAL EMPLOYMENT BY SECTOR	GDP ADDED BY MIGRANT WORKERS (RUB BILLIONS)
Construction	595 151	1 071 272	1 666 423	5 642 000	29.5	946.7
Wholesale and retail trade	272 114	489 805	761 919	13 542 000	5.6	432.9
Manufacturing	254 732	458 518	713 250	13 197 000	5.4	405.2
Real estate transactions, lease, services, financial activities	169 538	305 168	474 706	6 932 000	6.8	269.7
Agriculture and forestry, hunting and fishery	149 923	269 861	419 784	6 609 000	6.4	238.5
Transport and communication	70 592	127 066	197 658	5 430 000	3.6	112.3
Housing and municipal services	43 180	77 724	120 904	2 547 000	4.8	68.7
Health care, education and social services	8 519	15 334	23 853	10 270 000	0.2	13.6
Other	77 052	138 694	215 746	3 734 000	5.8	122.6
Total number of work permits at legal entities	1 640 801	2 953 442	4 594 243			2 610.0
Licences	191 200	344 160	535 360			304.1
Total	1 832 001	3 297 602	5 129 603	67 968 000	6.6	2 914.1

- The table contains information on documented and undocumented (illegal?) migrant workers in Russia in 2010.
- It lists the category of work that they fall into.
- The table is obtained from *The Role of Labour Migration in the Development of the Economy of the Russian Federation*, Table 3.1.2.

Migrant Workers in Russia

Example 14 (Russian Migrant Workers)



- There were more documented than undocumented workers.
- For both documented and irregular workers, the modal category was Construction.
- The sector that both contributed to the least was Healthcare.
- The distribution of documented and irregular workers was identical. The sector of contribution was not associated with the status of the workers.

Migrant Workers in Asia

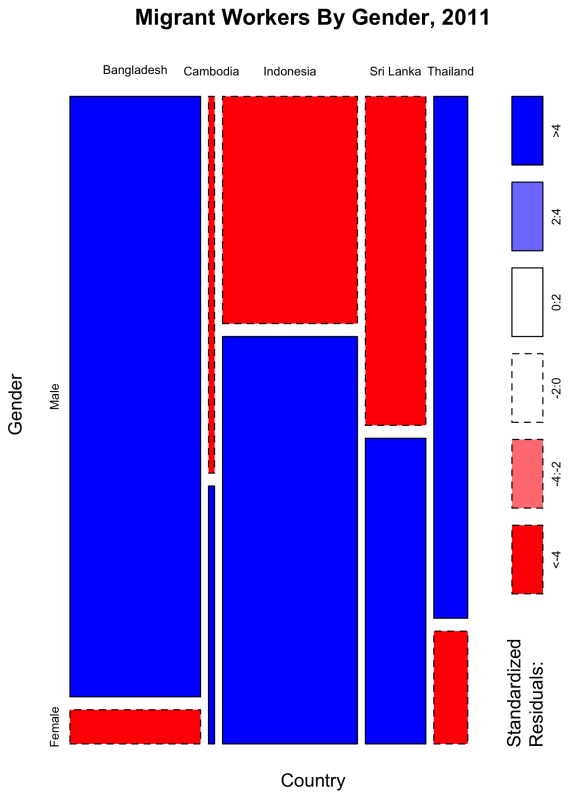
Example 15 (Gender of Migrant Workers)

COUNTRY, TOTAL MIGRANTS AND BY SEX	2007	2008	2009	2010	2011
BANGLADESH					
Total	822 387	875 109	475 278	390 702	568 062
Male	803 293	854 267	453 054	365 864	537 483
Female	19 094	20 842	22 224	24 838	30 579
CAMBODIA					
Total	9 476	7 340	14 928	29 783	26 219
Male	4 611	3 616	4 292	10 501	15 563
Female	4 865	3 724	10 636	19 282	10 656
INDIA					
Total	809 453	848 601	610 272	641 356	626 565
INDONESIA					
Total	696 746	644 731	632 172	575 804	586 802
Male	152 887	148 600	103 188	124 684	210 116
Female	543 859	496 131	528 984	451 120	376 686
NEPAL					
Total	204 533	219 965	294 094	354 716	384 665
PAKISTAN					
Total	287 033	430 314	403 528	362 904	456 893
PHILIPPINES					
Total	1 077 623	1 236 013	1 422 586	1 470 826	1 687 831
SRI LANKA					
Total	218 459	250 499	247 119	267 507	262 961
Male	103 476	128 232	119 276	136 850	136 307
Female	114 983	122 267	127 843	130 657	126 654
THAILAND					
Total	161 917	161 852	147 711	143 795	147 623
Male	137 923	137 325	124 227	121 168	121 391
Female	23 994	24 527	23 484	22 627	26 232

- The table contains a breakdown of migrant workers by their gender.
- It is taken from Table 1.1 of *SDD AP Migration Report*.

Migrant Workers in Asia

Example 15 (Gender of Migrant Workers)



- Bangladesh and Indonesia have the largest amount of migrant workers.
- While Cambodia and Sri Lanka have roughly the same proportion of male and female migrant workers, the division is more skewed for the other three countries:
  - Bangladesh and Thailand have a much higher proportion of males.
  - Indonesia has more females.
- Cambodia has the fewest number of migrant workers.

Migrants in Thai School System

Example 16 (Migrants in Thai Schools)

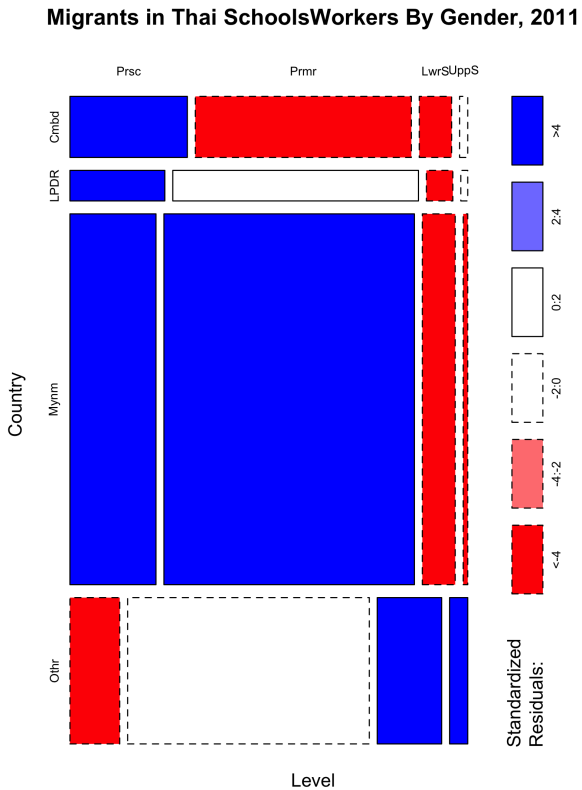
Number of international migrant children in Thai schools by level and country of origin, 2012

LEVEL	CAMBODIA	LAO PEOPLE'S DEMOCRATIC REPUBLIC	MYANMAR	ALL OTHER COUNTRIES	TOTAL
Preschool	2 568	1 038	11 428	2 608	17 642
Primary school	4 726	2 688	33 275	12 666	53 355
Lower secondary	708	289	4 369	3 376	8 742
Upper secondary	178	76	605	950	1 809
Total	8 180	4 091	49 677	19 600	81 548

- The table contains a breakdown of migrants in Thai schools in 2010.
- It is taken from Box 3.2 of *SDD AP Migration Report*.
- The school levels considered are preschool, primary school, lower secondary and upper secondary schools.

Migrants in Thai School System

Example 16 (Migrants in Thai Schools)



The following points are clear to see:

- Most migrants in Thai schools are from Myanmar.
- Most of the immigrants are in primary school.
- There is a surplus of immigrants from Myanmar, Laos and Cambodia in preschools.

Palma Ratio

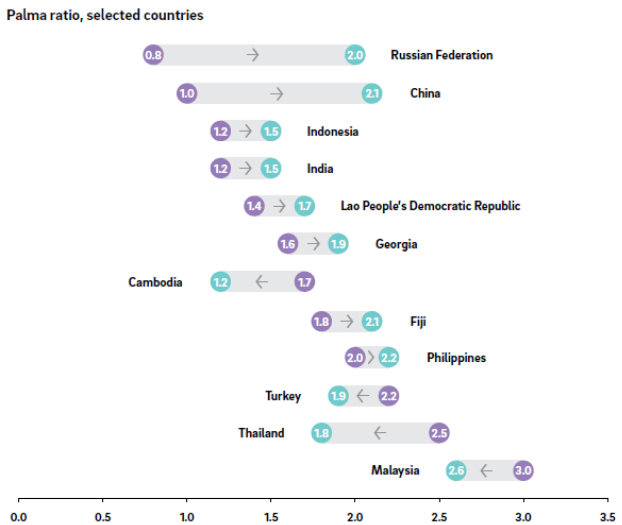
- The Palma ratio is an alternative measure of income inequality.
- It is computed as the following ratio:

$$\frac{\text{Income share of richest 10\%}}{\text{Income share of poorest 40\%}}$$

- It has some advantages to the Gini Index, among which are that it is intuitive, and takes into account the middle class explicitly.
- The following URL contains a deeper discussion of the Palma Ratio: `http://www.cgdev.org/blog/palma-vs-gini-measuring-post-2015-inequality`

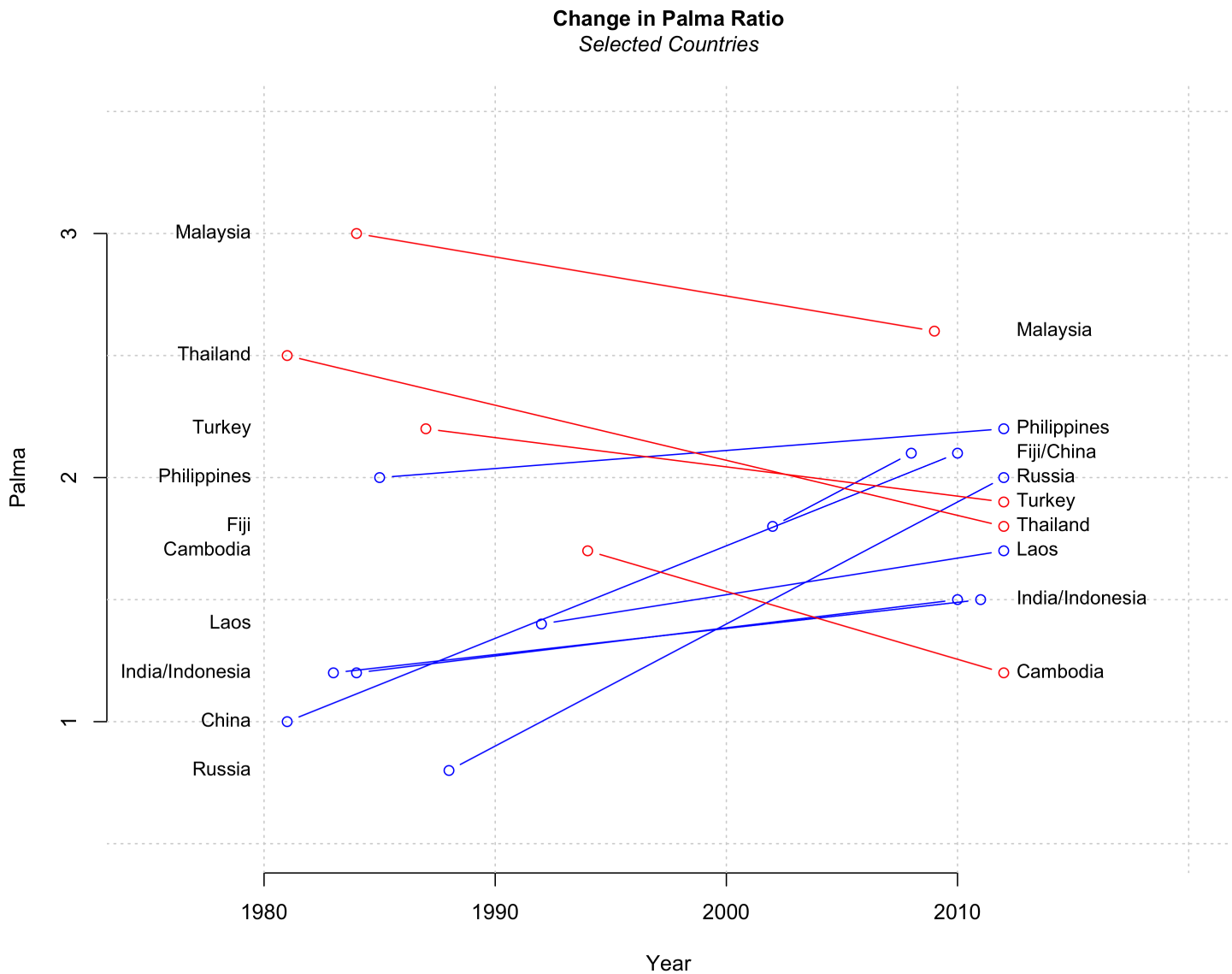
Change in Palma Ratio

Example 17 (Change in Palma Ratio)



- The plot on the left depicts changes in Palma ratio of several countries.
- It is taken from Figure 1.4 of *Time for Equality* report.
- The change takes place over time periods of different lengths.
- What can we say about the countries and the rate of change?

Change in Palma Ratio



Summarising The Changes in Palma Ratio

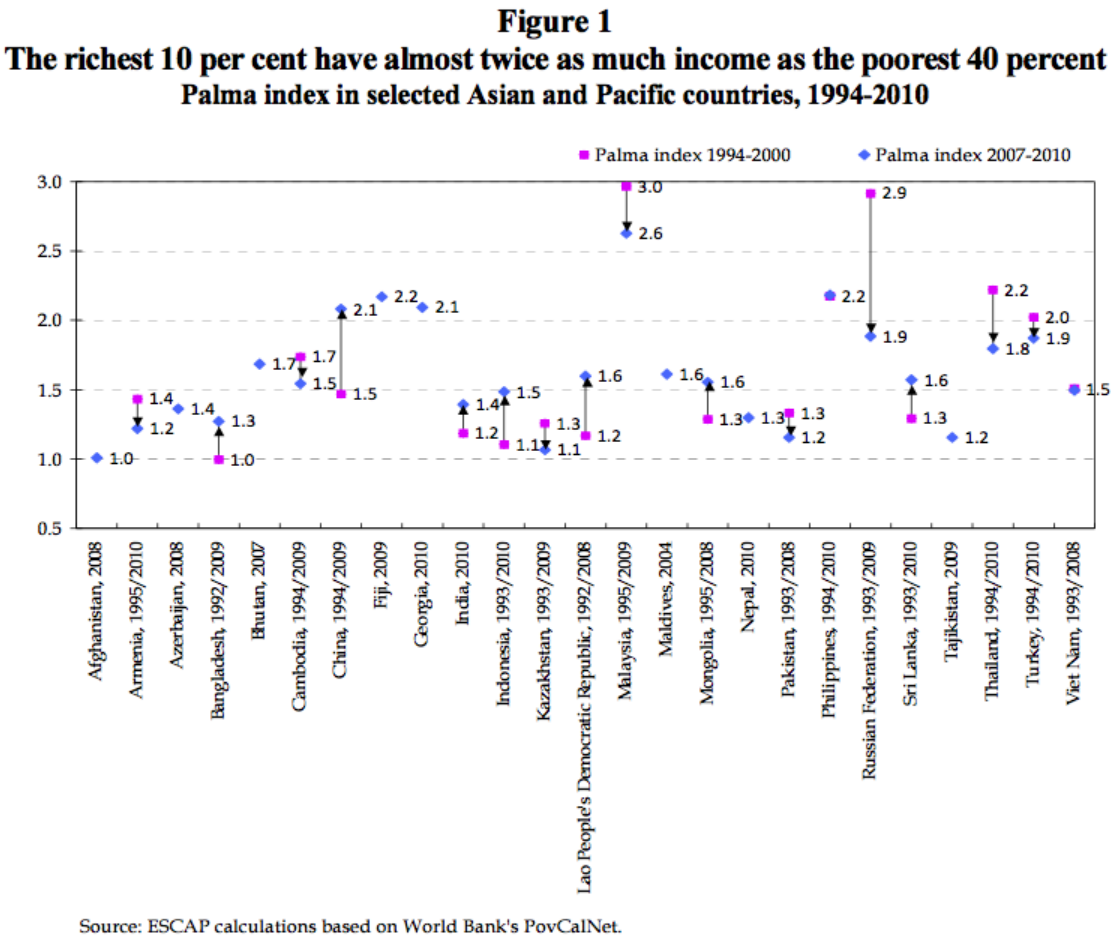
With the second plot, it is easier to note that

- Malaysia has a consistently high ratio, quite different from the rest.
- Cambodia, Thailand, Turkey and Malaysia have reduced their income inequality.
- China and Russia have increased at similar rates; their inequality has approximately doubled.
- India and Indonesia are very similar in inequality and rate of growth.
- Fiji has had a striking increase in a short time.

We are making an implicit assumption that the rate of change of Palma ratio is linear. However, we have little else to go on, and we need to reflect the time factor in the change.

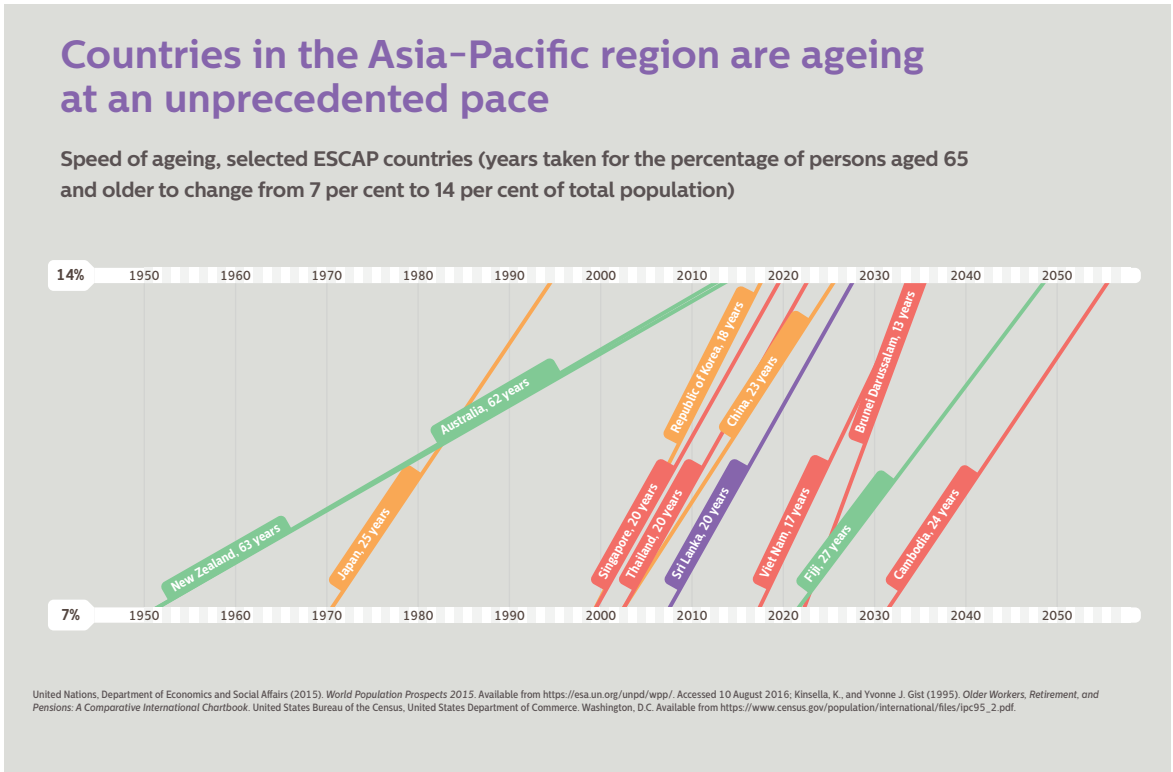
Applicability of This Graphic

The same approach can be applied in this graphic.



Applicability of This Graphic

As you can see, you are in fact already using this approach:



Index