Case-Based Reasoning with Population Data & The Lexis Surface

Morning Session: Mortality

Jon Minton

19 June 2018

Welcome!

About your instructor: Jon Minton. My background

- Undergrad: Engineering
- Taught postgrad: Critical Theory/Cultural Studies
- PhD: Quant Sociology & Human Geography (Welfare Reform)
- Postgrad: Housing Policy (Interviewing)
- Then: Systematic Reviewing (Health Sciences: NICE)
- Then: Health Economics (Health Sciences: NICE)
- Then: Urban Studies

My perspective

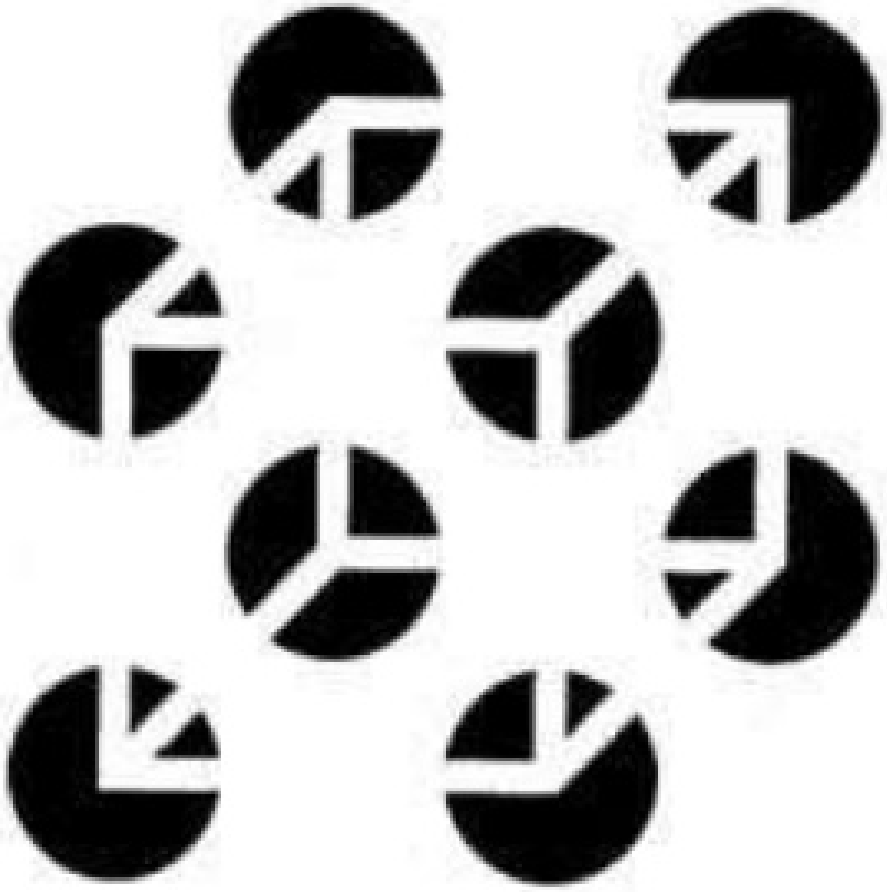
- Accidental Demography
- Quant / (Some) Qual (a bit)
- Physical Sciences / Social Sciences / Humanities
- Health / Sociological

Core aim

To promote the following ideas:

- Much of the quant/qual divide is really case-based vs. variable-based
- Quantitative research isn't about numbers, it's about patterns
- Demography (**demos**: the people; -**graphy**: describing) is the grandmother of the social sciences
- People are good at images but bad at numbers
- People are good at complex **gestalts**; bad at linear sequence
- The Lexis Surface (**maps of age-time**) allows data to be explored much as maps of space can be explored

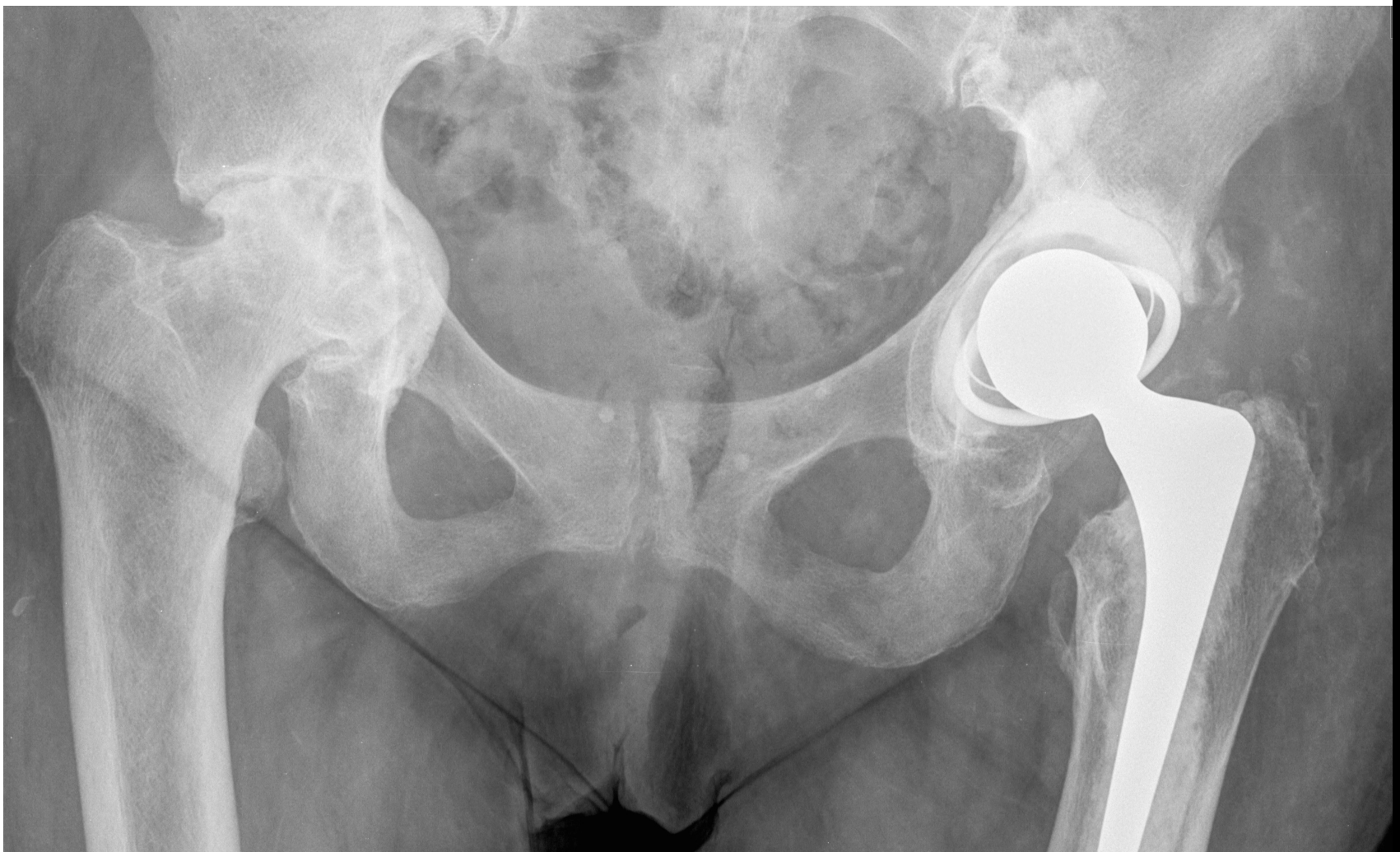
Examples of Gestalts



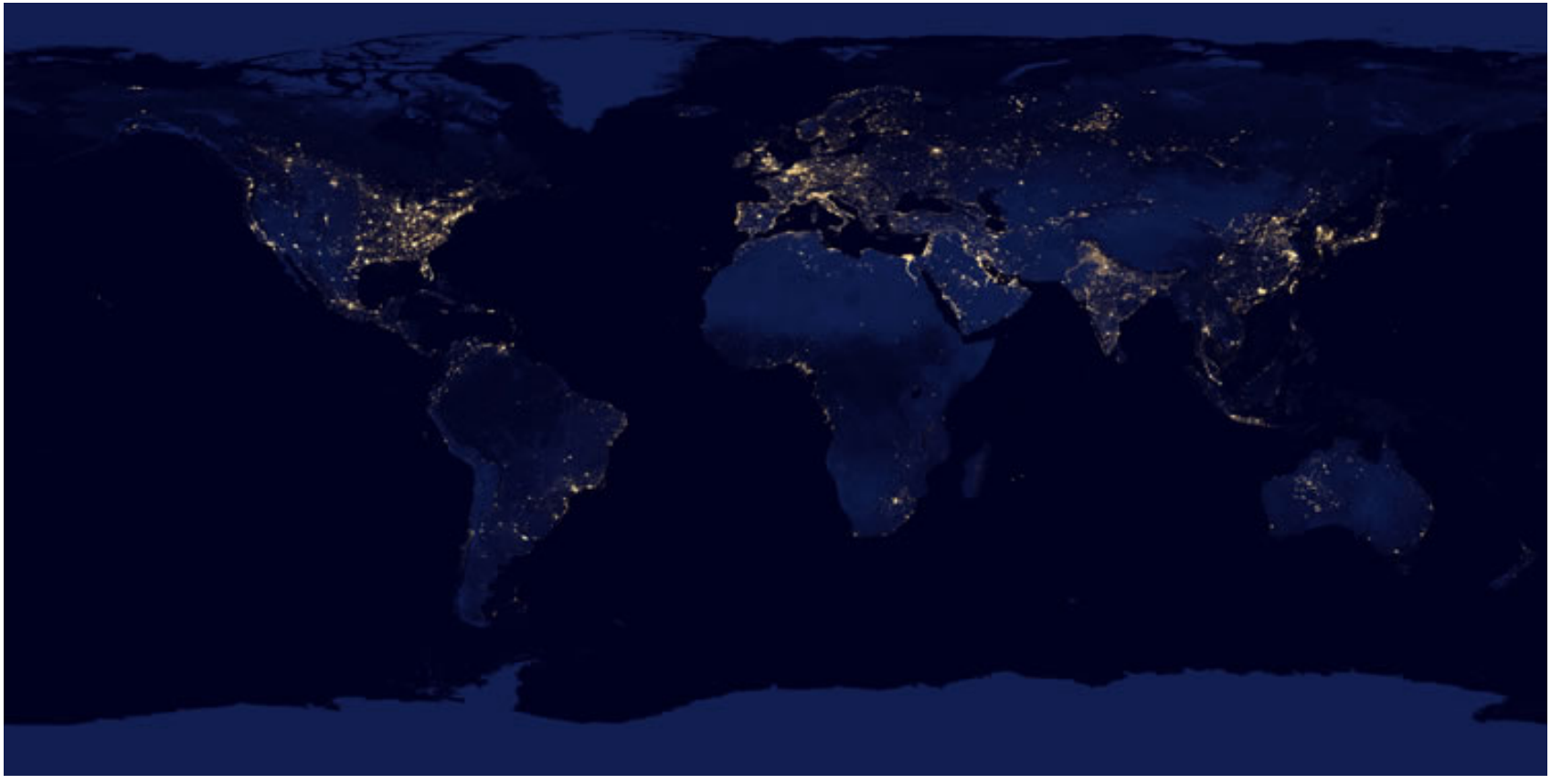
Case-based reasoning as Gestalts

Figure 1

Case-based reasoning as Gestalts

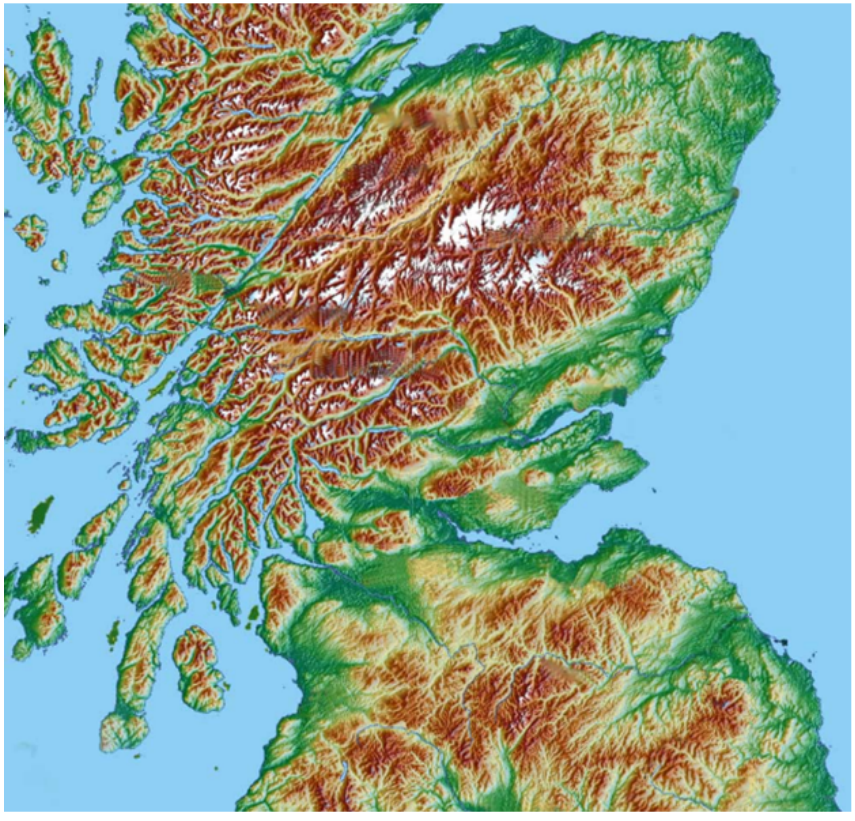


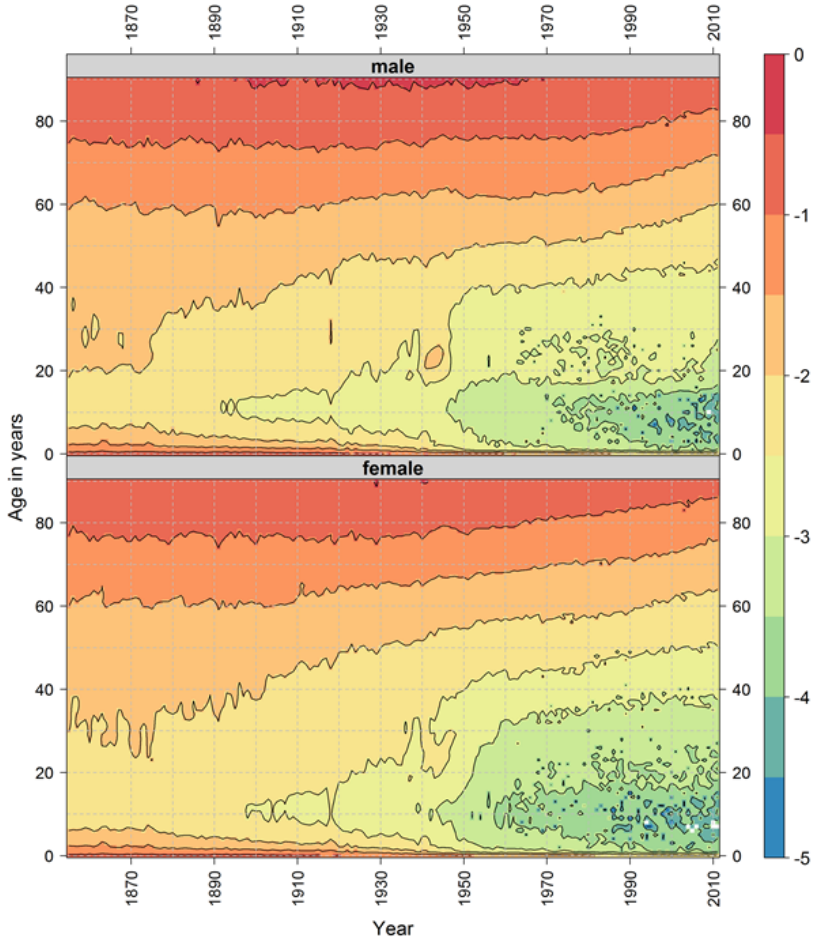
Maps as Gestalts



- Source: `https://geology.com/articles/satellite-photo-earth-at-night.shtml`

Map of place; map of age-time





- Source: `https://ije-blog.com/2016/06/27/lexis-cubes-1-from-maps-of-space-to-maps-of-time/comment-page-1/`

Maps and Mapping

Data visualisations

- Not all graphics are data visualisations
- Data visualisations require a consistent application of **mapping rules**

Mapping rules:

- Variable in data $->$ graphical feature
- Can be specified formally using the Grammer of Graphics

Examples of variables in data

- Age
- Year
- Gender
- Death rate
- Crime rate
- Fertility rate
- Health scores
- Political Attitudes

etc

Examples of Graphical features

- Position across horizontal axis
- Position across vertical axis
- Colour of marks
- Size of dot/width of line
- Transparency
- whether lines are solid or dashed
- Colour in filled areas between marks
- Angle

etc

Example

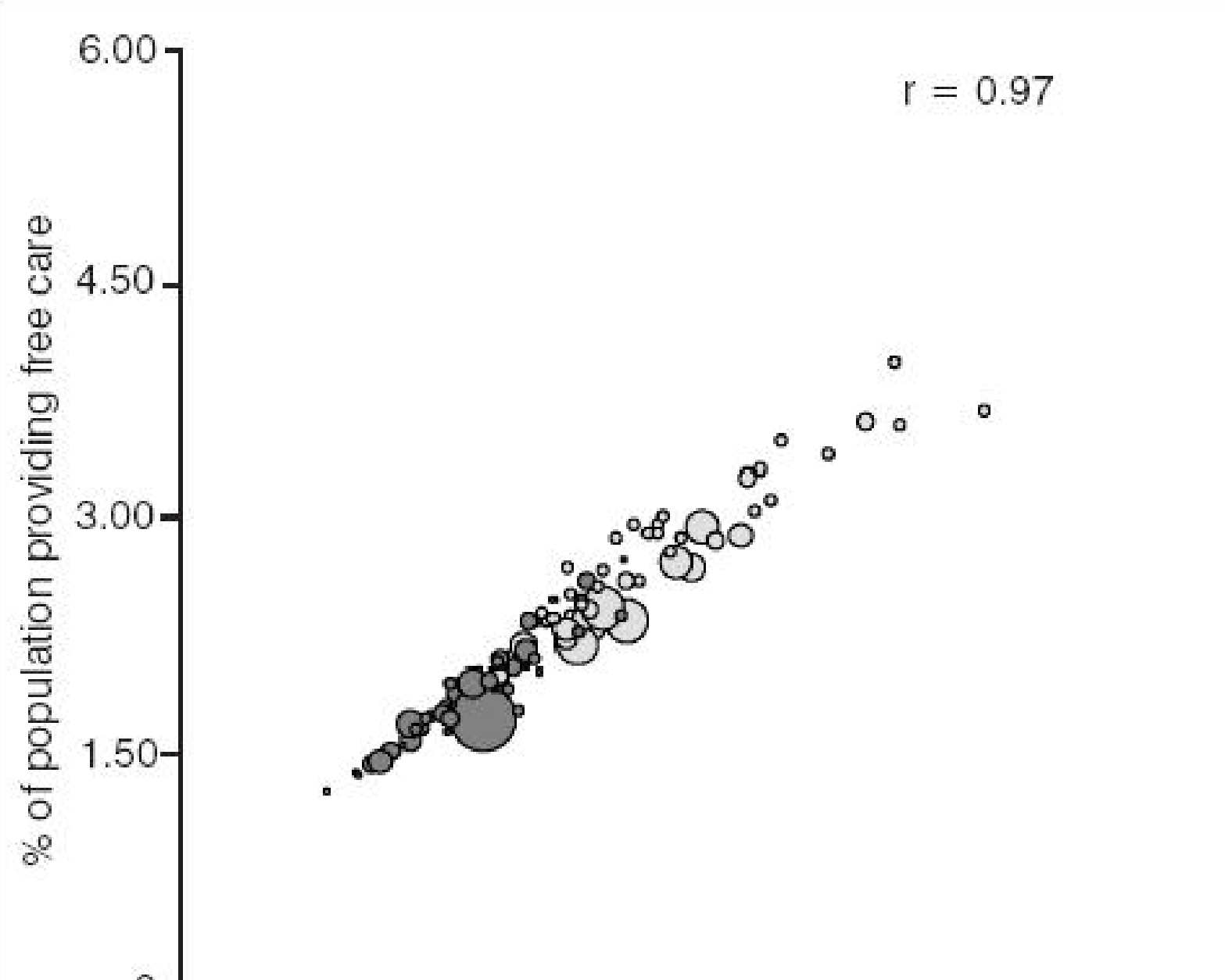
This set of mapping rules...

Variable in Dataset	Graphical Feature
% of population providing free care	Position along vertical axis
% of population with health needs	Position along horizontal axis
Size of population in areal unit	Size of circle
Whether in North or south of England	Colour of bubble

Applied to 2001 English/Welsh Census data...

Example

. . . Produces the following



Population data

Population data are data where:

- **Something**
- Has been recorded consistently about **types of people**
- For different **ages**
- and at different **times**

Example of data in this format

From the Human Mortality Database

```
dta
```

```
FALSE # A tibble: 1,445,886 x 6
FALSE    country  year   age sex    death_count population_
FALSE    <chr>   <int> <int> <chr>        <dbl>
FALSE  1 AUS      1921     0 female      3842.
FALSE  2 AUS      1921     1 female       586.
FALSE  3 AUS      1921     2 female       390.
FALSE  4 AUS      1921     3 female       254.
FALSE  5 AUS      1921     4 female       176.
FALSE  6 AUS      1921     5 female       146.
FALSE  7 AUS      1921     6 female       128.
FALSE  8 AUS      1921     7 female       112.
FALSE  9 AUS      1921     8 female        97.0
FALSE 10 AUS      1921     9 female        83.8
FALSE # ... with 1,445,876 more rows
```

Decoding this

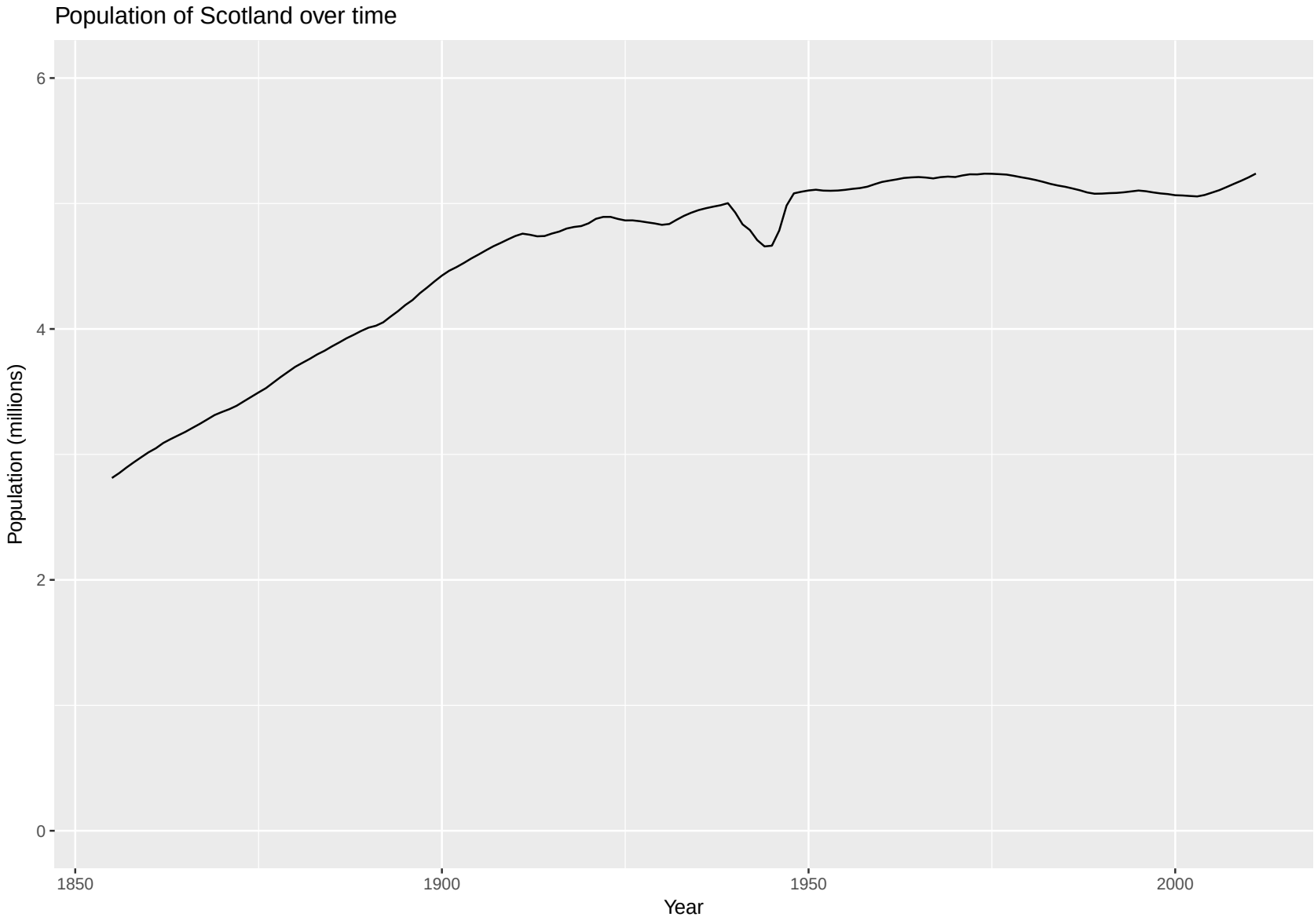
- country and sex are grouping variables (i.e. categorical not cardinal)
- year and age are continuous variables
- death_count and population_count are attributes that are specific to different combinations of country, sex, year, and age.

1.4 million rows!

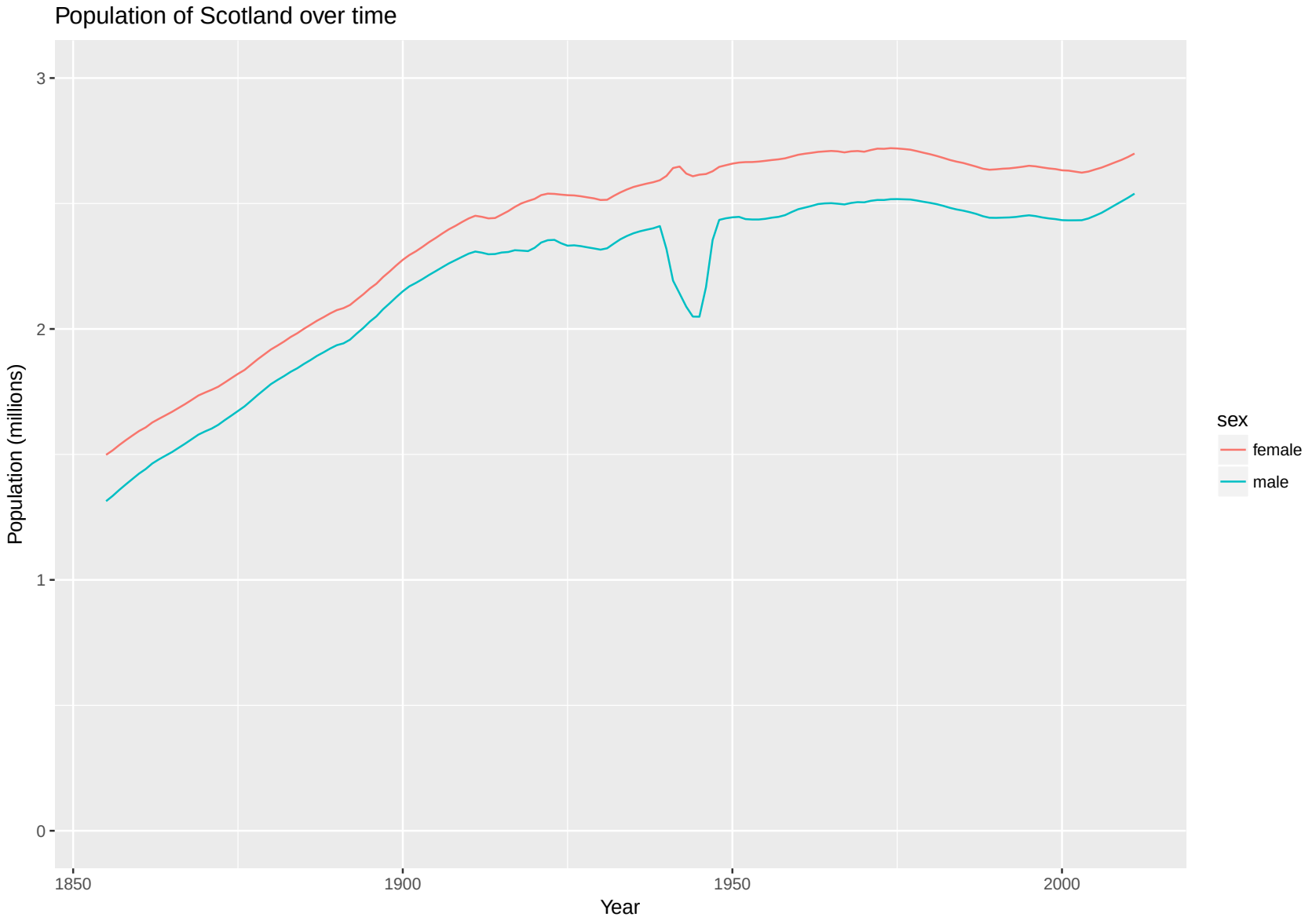
Standard ways of exploring

- Sweeping by year: Life expectancies, crude mortality rates
- Sweeping by age: 'Bathtub curves'
- Conditional sweeping by year: different age groups
- etc

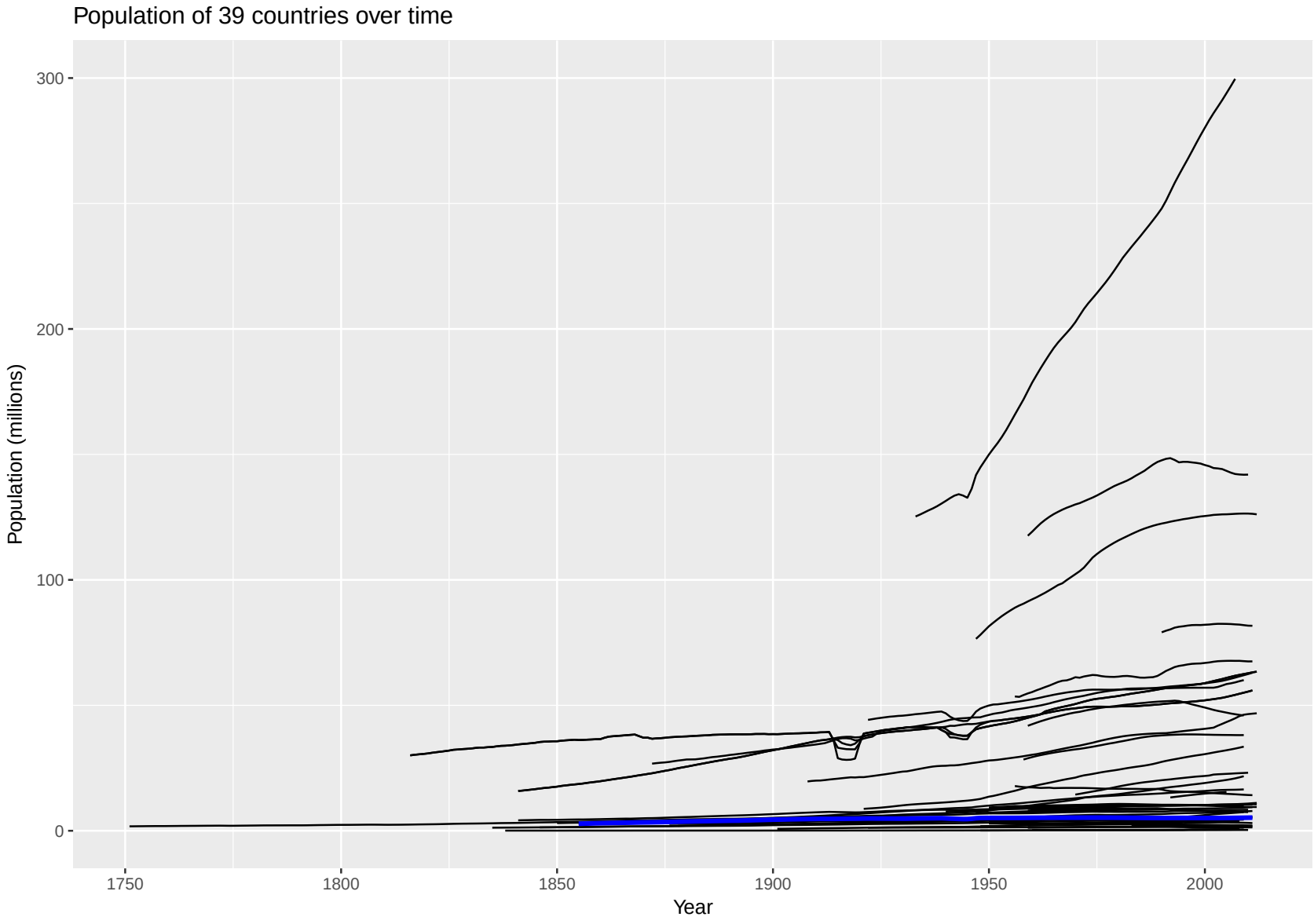
Example: Scotland, population size



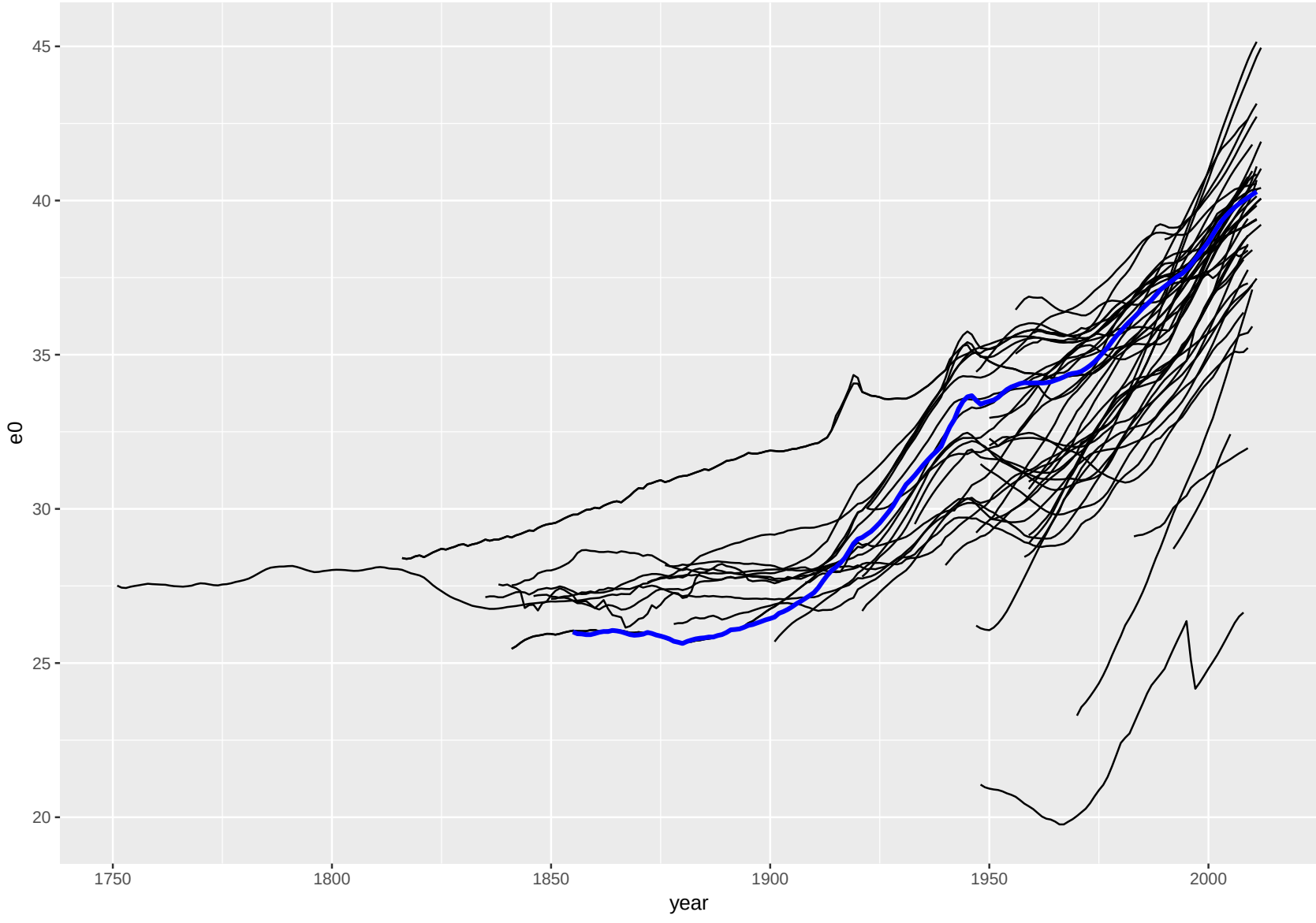
Example: Scotland, population size



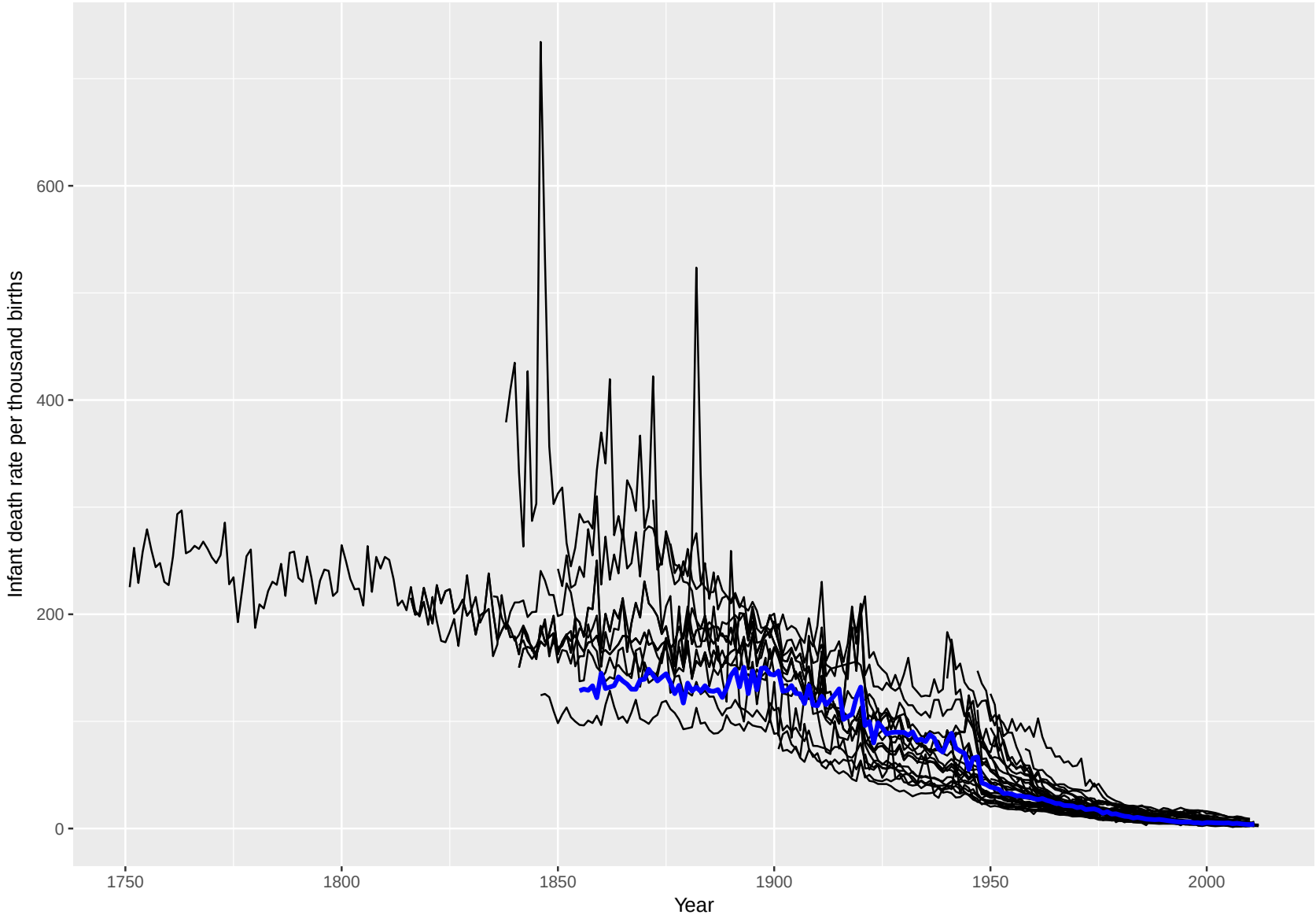
But this quickly becomes overwhelming



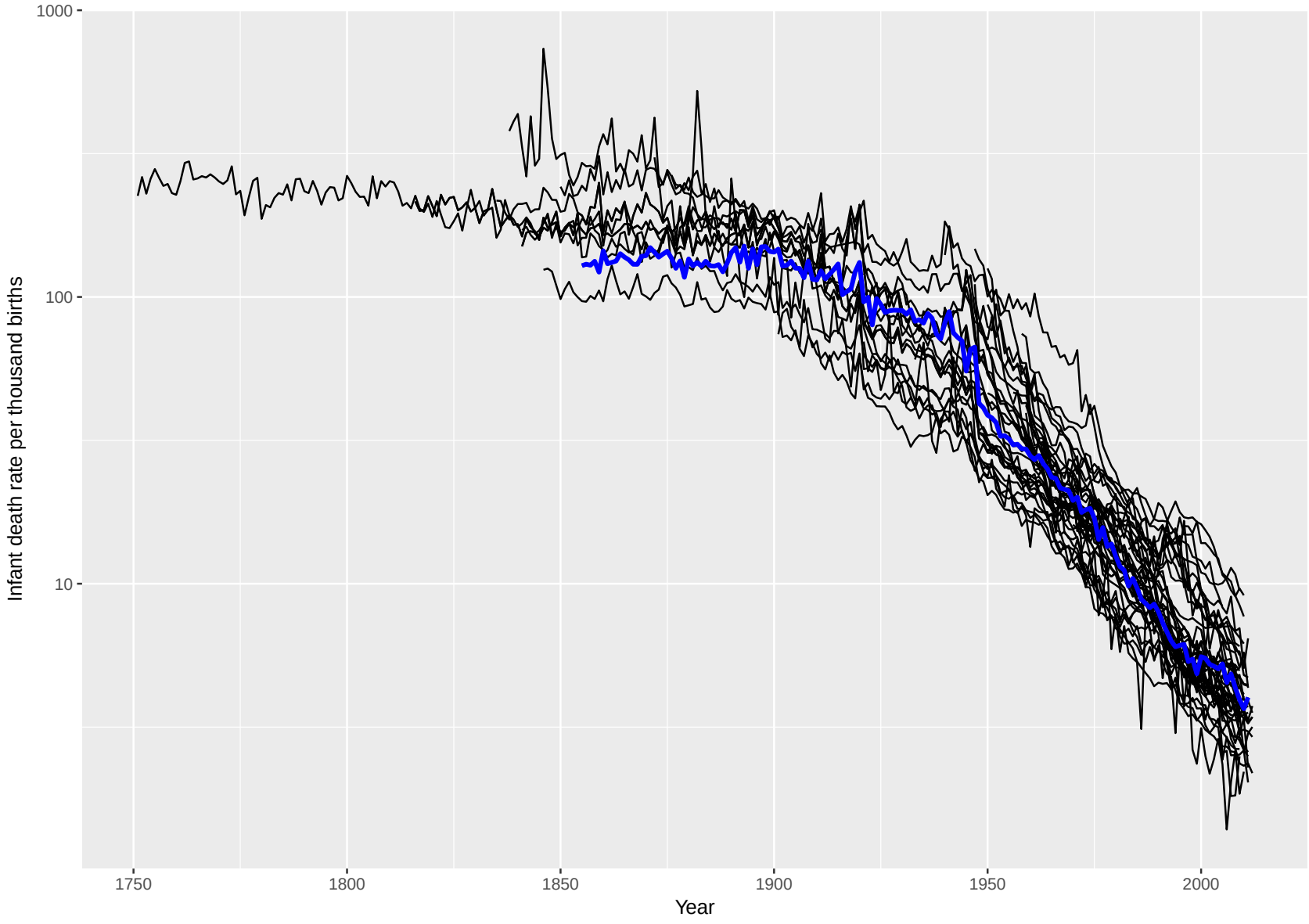
Life expectancies



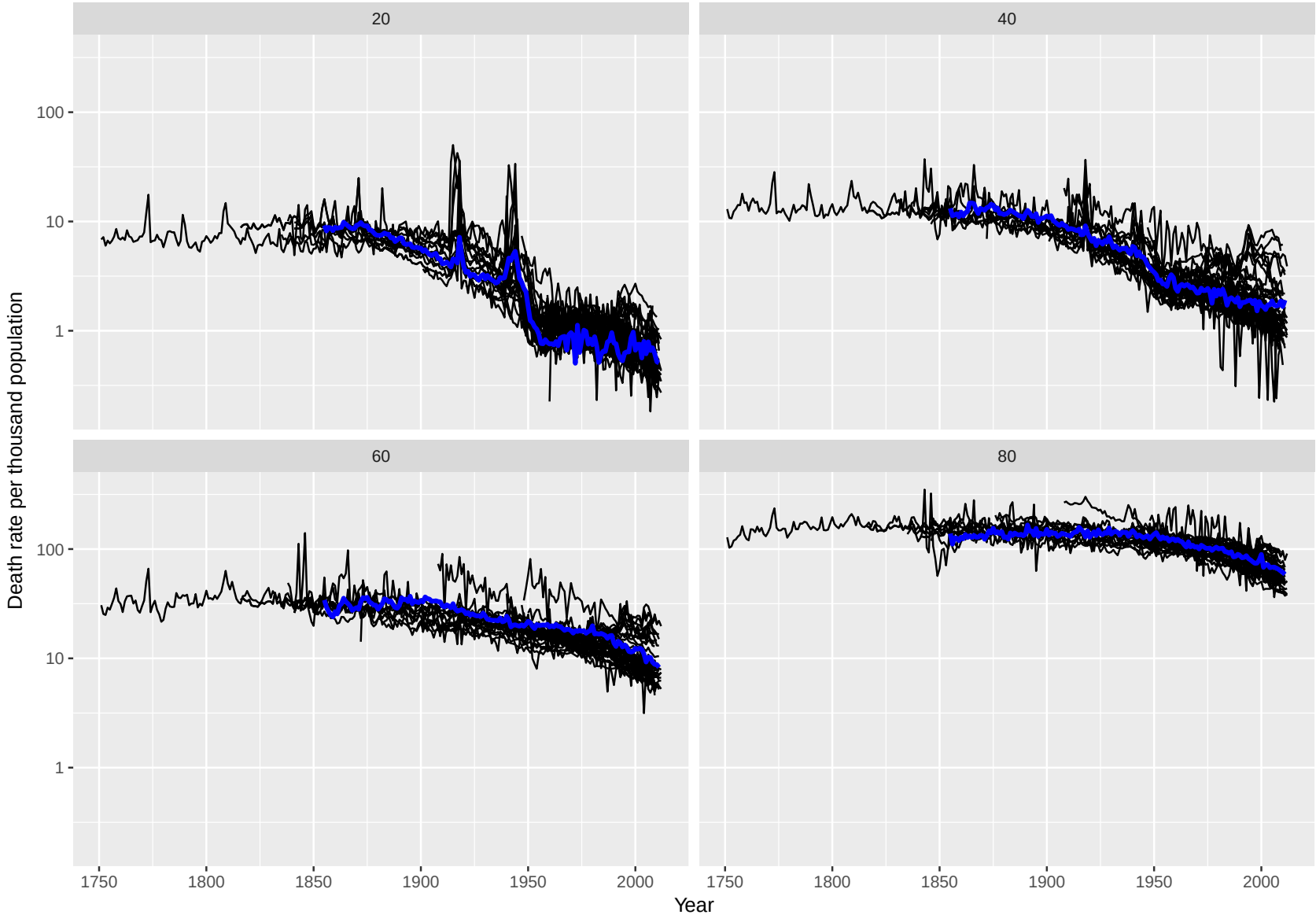
Infant mortality



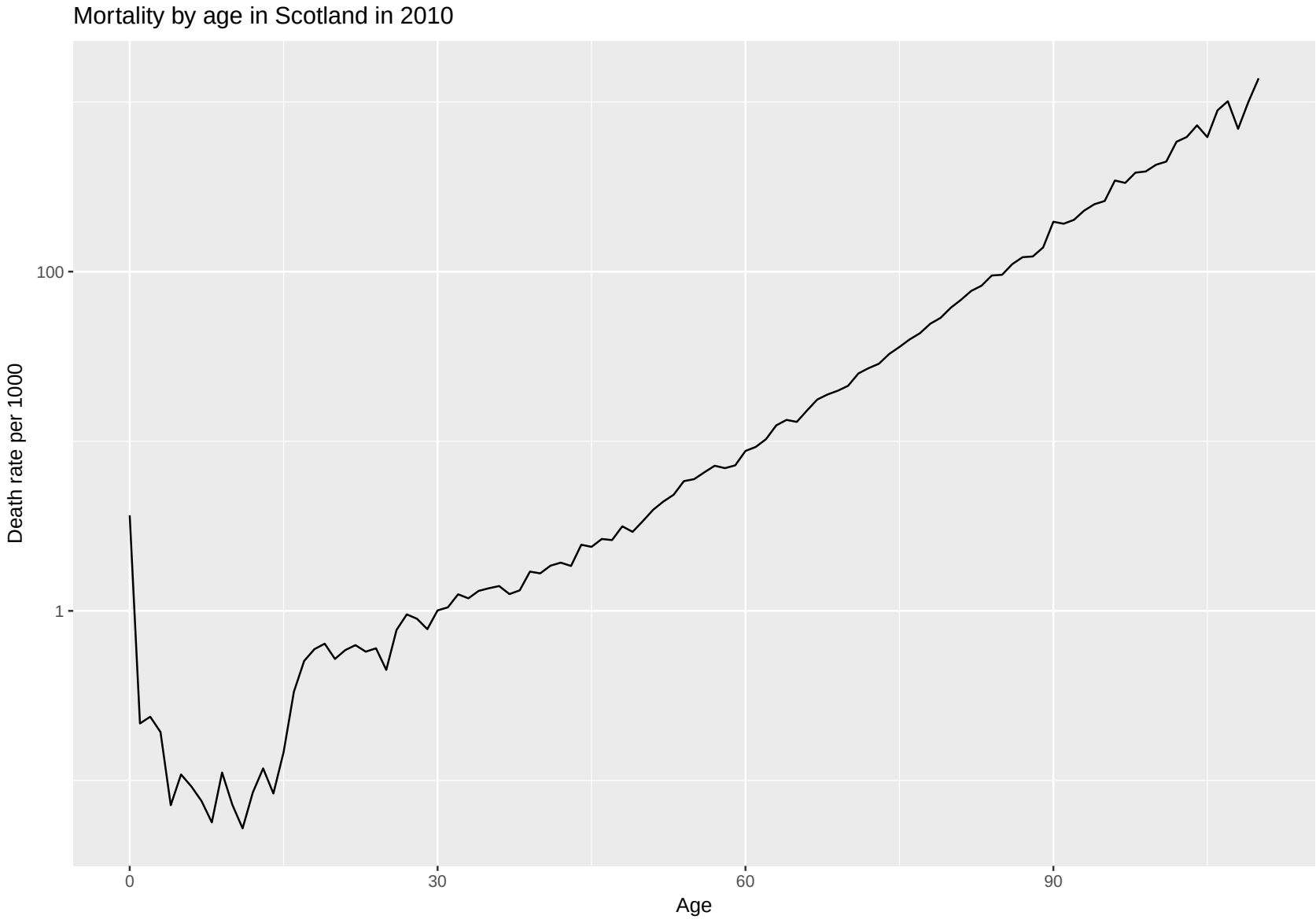
Infant mortality



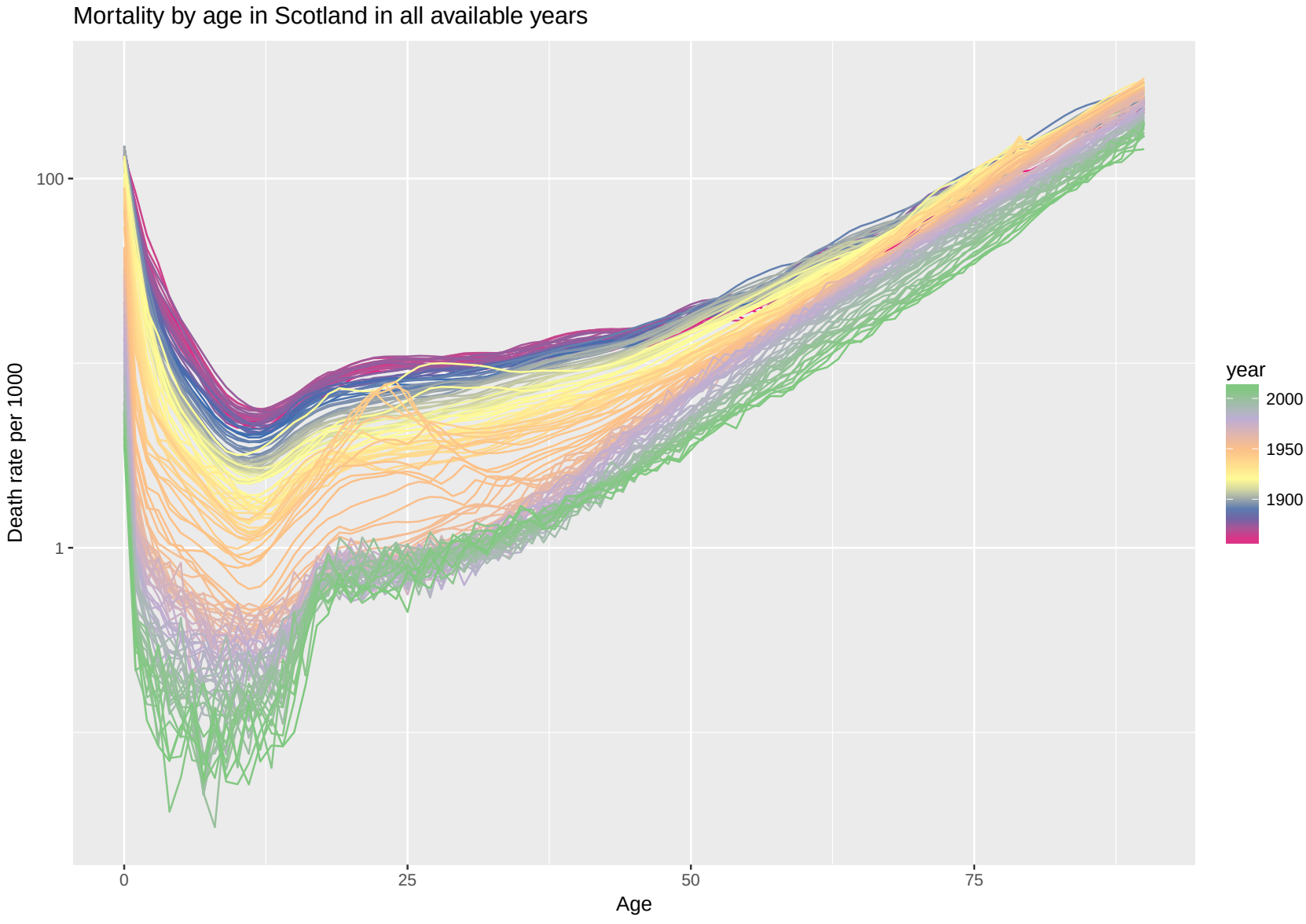
Other ages


20
40
60
80
100
10
1
Death rate per thousand population
1750
1800
1850
1900
1950
2000
Year

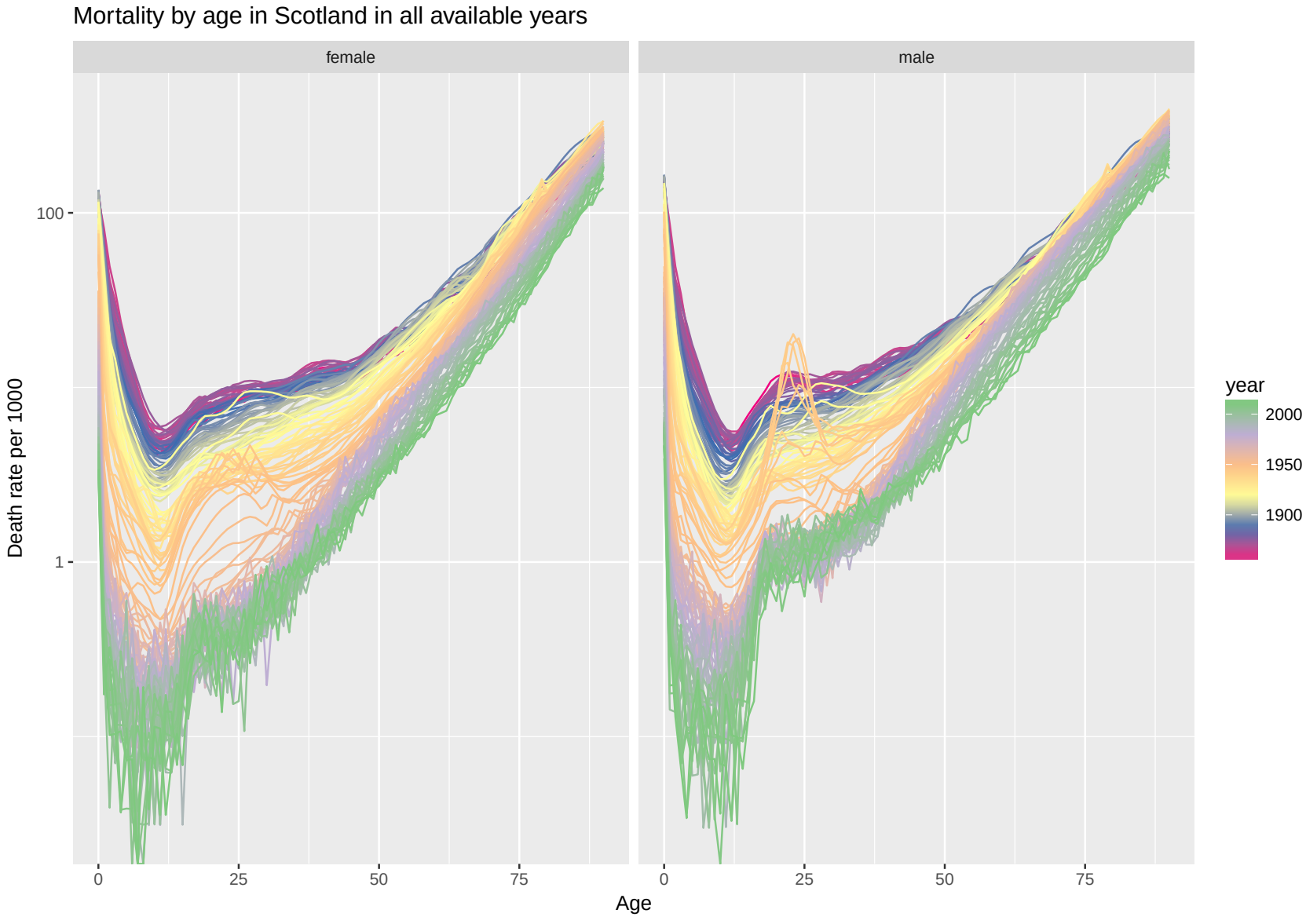
Relationship with age



Bathtub, Scotland, all years



Bathtub, Scotland, all years, by gender



A variable-based approach (most quantitative research)

- Simple linear regression: Regress one variable on one variable
- Multiple linear regression: Regress one variable on multiple variables
- Assume independence between explanatory variables
- (Usually) assess statistical significance of regression coefficients
  - "The sizeless stare"
  - Conflates statistical with substantive significance

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5977 -1.1075 -0.1401  1.2590  3.2585
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -3.60459    0.02012 -179.148  < 2e-16 ***
## sexmale      0.22777    0.02845    8.005 1.35e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 1.326 on 8682 degrees of freedo
## Multiple R-squared:  0.007326,   Adjusted R-squared:  0.
## F-statistic: 64.07 on 1 and 8682 DF,  p-value: 1.353e-15
```

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex + years_since_first,
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62526 -0.38316  0.01859  0.40022  1.70713
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.9755770  0.0246242   39.62   <2e-16
## sexmale           0.2277717  0.0120983   18.83   <2e-16
## years_since_first -0.0234263  0.0001181 -198.36   <2e-16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.5637 on 8681 degrees of freed
## Multiple R-squared:  0.8206, Adjusted R-squared:  0.8205
```

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex + poly(years_since_fi
##     2), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63102 -0.25707 -0.03739  0.22951  1.37089
##
## Coefficients:
##                                  Estimate Std. Error t valu
## (Intercept)                    -3.605e+00  5.590e-03 -644.8
## sexmale                         2.278e-01  7.905e-03   28.8
## poly(years_since_first, 2)1 -1.118e+02  3.683e-01 -303.5
## poly(years_since_first, 2)2 -3.976e+01  3.683e-01 -107.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex + poly(years_since_fi
##     3), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53729 -0.24378 -0.03294  0.22219  1.31336
##
## Coefficients:
##                                  Estimate Std. Error t valu
## (Intercept)                    -3.605e+00  5.488e-03 -656.7
## sexmale                         2.278e-01  7.762e-03   29.3
## poly(years_since_first, 3)1 -1.118e+02  3.616e-01 -309.1
## poly(years_since_first, 3)2 -3.976e+01  3.616e-01 -109.9
## poly(years_since_first, 3)3 -6.509e+00  3.616e-01  -18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex * poly(years_since_f
##     2), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62939 -0.25635 -0.03894  0.22798  1.37746
##
## Coefficients:
##                                          Estimate Std. Erro
## (Intercept)                             -3.605e+00  5.587e-0
## sexmale                                  2.278e-01  7.901e-0
## poly(years_since_first, 2)1             -1.127e+02  5.207e-0
## poly(years_since_first, 2)2             -3.907e+01  5.207e-0
## sexmale:poly(years_since_first, 2)1      1.769e+00  7.363e-0
## sexmale:poly(years_since_first, 2)2 -1.385e+00  7.363e-0
##                                         Pr(>|t|)
```

An example

```
##
## Call:
## lm(formula = log(death_rate) ~ sex + is_scotland + poly(
##     2), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63147 -0.25585 -0.03673  0.22849  1.36892
##
## Coefficients:
##                                  Estimate Std. Error  t val
## (Intercept)                    -3.603e+00  5.641e-03 -638.7
## sexmale                         2.278e-01  7.903e-03   28.8
## is_scotlandTRUE                -4.947e-02  2.124e-02   -2.3
## poly(years_since_first, 2)1 -1.119e+02  3.687e-01 -303.3
## poly(years_since_first, 2)2 -3.982e+01  3.690e-01 -107.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

An example

```
## Analysis of Variance Table
##
## Model 1: log(death_rate) ~ sex
## Model 2: log(death_rate) ~ sex + years_since_first
## Model 3: log(death_rate) ~ sex + poly(years_since_first,
## Model 4: log(death_rate) ~ sex + poly(years_since_first,
## Model 5: log(death_rate) ~ sex * poly(years_since_first,
## Model 6: log(death_rate) ~ sex + is_scotland + poly(year
##     2)
##   Res.Df     RSS Df Sum of Sq          F  Pr(>F)
## 1   8682 15261.4
## 2   8681  2758.5  1   12502.9 92243.6938 < 2e-16 ***
## 3   8680  1177.5  1    1581.1 11664.6875 < 2e-16 ***
## 4   8679  1135.1  1      42.4   312.5759 < 2e-16 ***
## 5   8678  1176.2  1     -41.1
## 6   8679  1176.8 -1      -0.5     3.8829 0.04881 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

A case-based approach (focus of this workshop)

- **Embrace inherent complexity**
- Interactions between factors norm not the exception?
- Imagine tens of thousands of values not as hived off into distinct variables (age, year)...
- But forming a complex *surface* of values over age-time

How? *By representing the surface on a map*

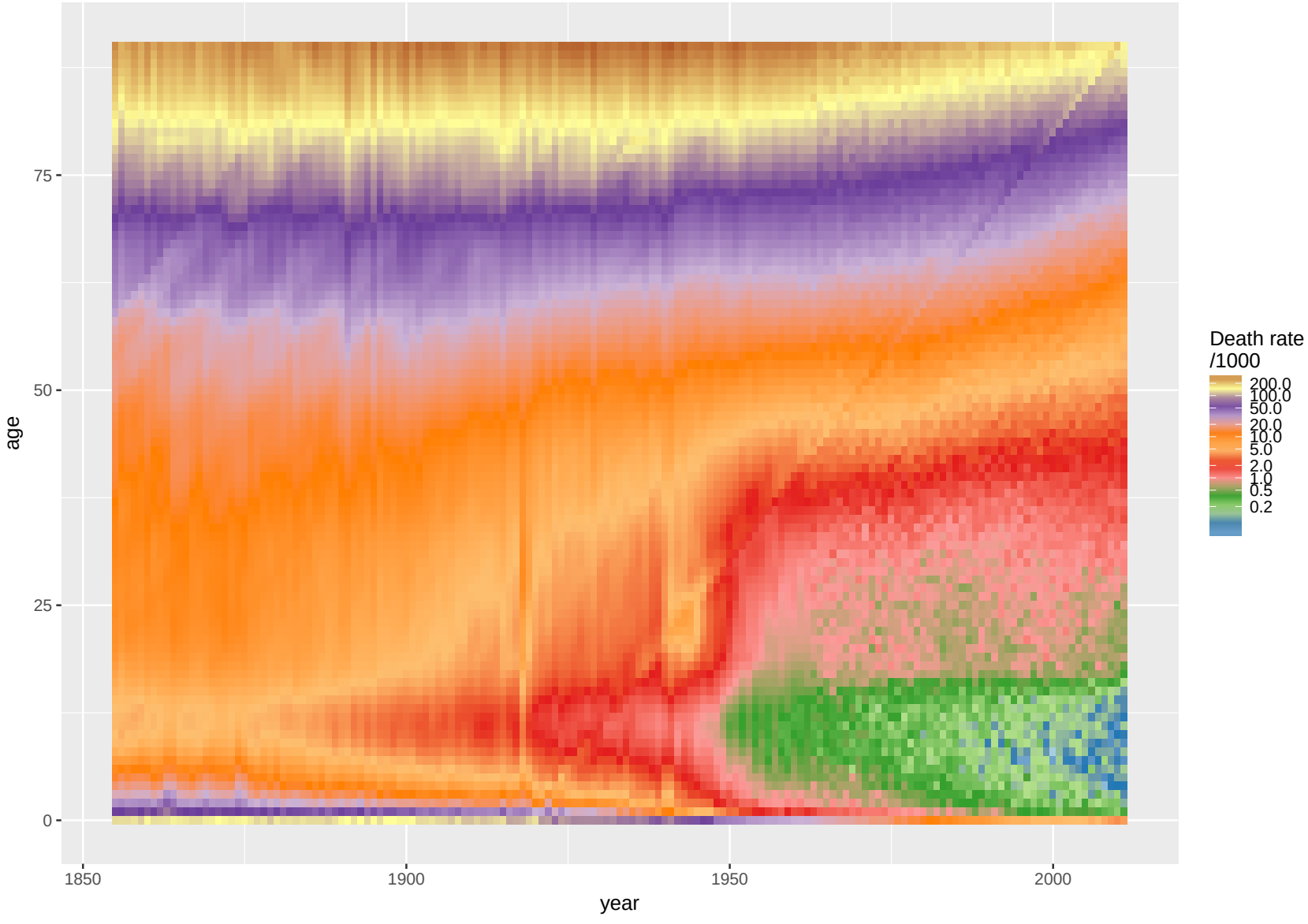
Mapping a spatial map

Data Variable	Aesthetic
Latitude	Horizontal position
Longitude	Vertical position
Elevation	Colour/shade/contour lines

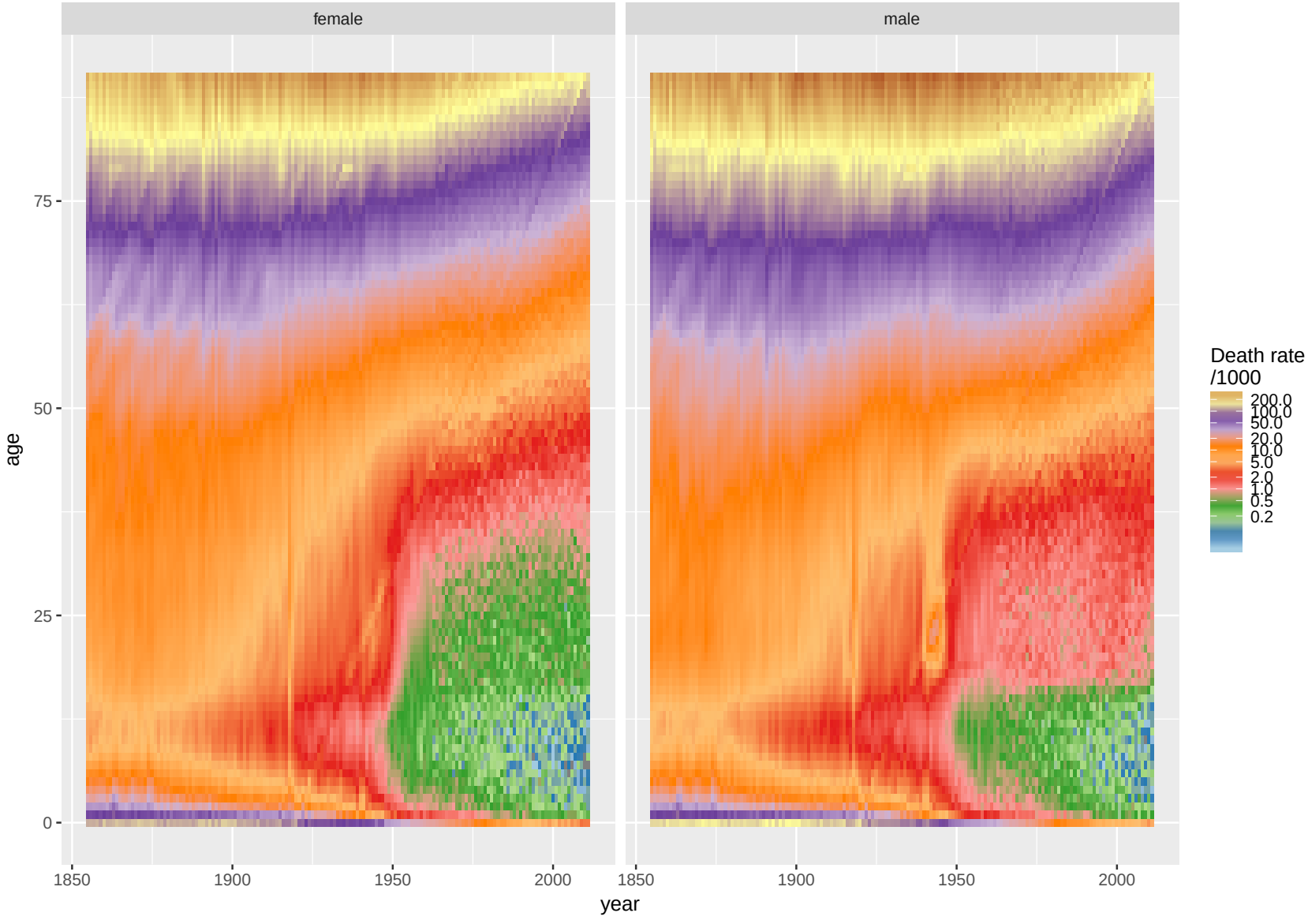
Mapping an age-time map (Lexis surface)

Data Variable	Aesthetic
Year	Horizontal position
Age	Vertical position
Mortality rate	Colour/shade/contour lines

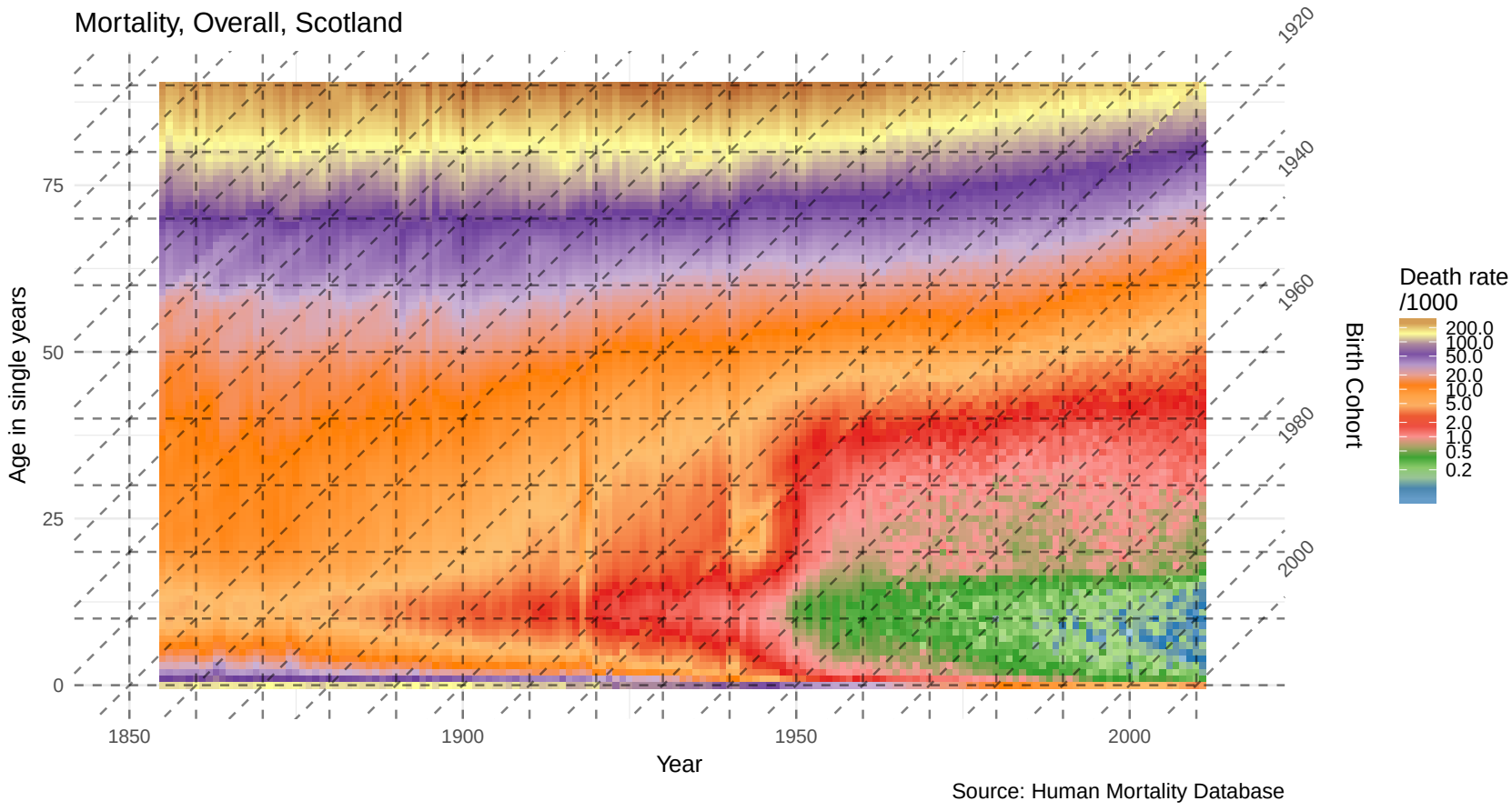
Lexis surface for Scotland



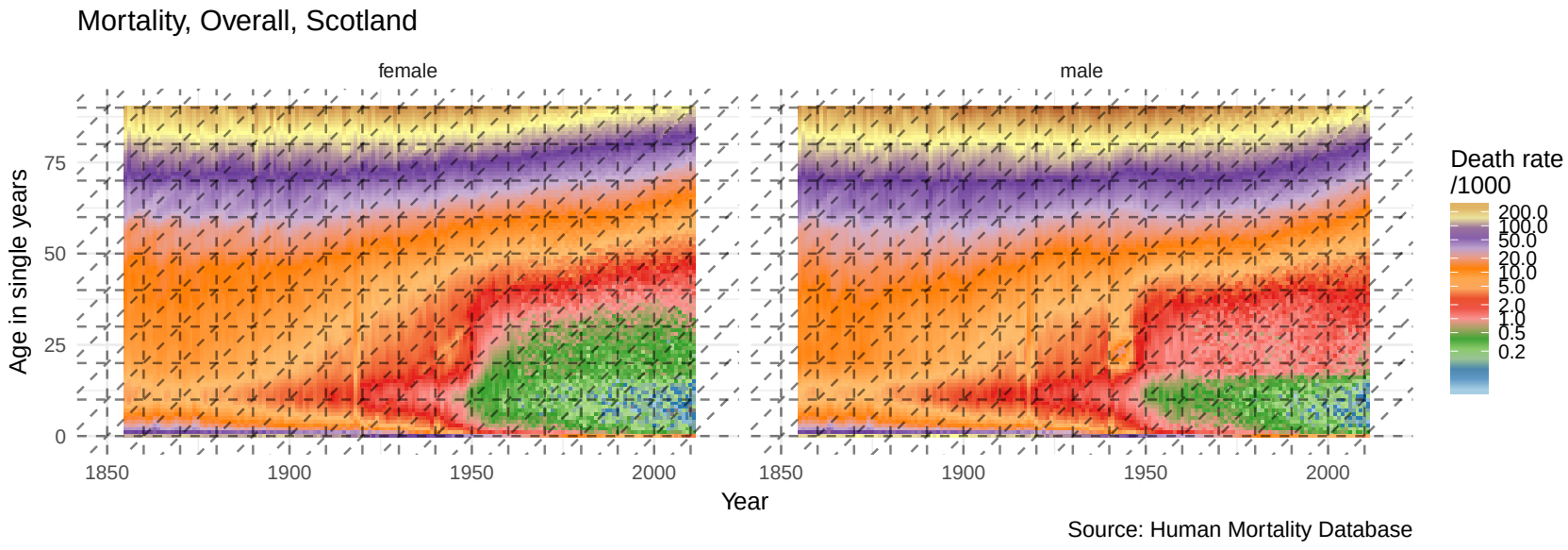
Lexis surface for Scotland



Additional modifications



By gender


Mortality, Overall, Scotland
female
male
Age in single years
0
25
50
75
1850
1900
1950
2000
Year
Death rate
/1000
200.0
100.0
50.0
20.0
10.0
5.0
2.0
1.0
0.5
0.2
Source: Human Mortality Database

Aim of morning session

- Explore country/gender-specific Lexis surfaces
- Identify, discuss and interpret features in cases
  - Common to many cases
  - Specific to one or two cases
- Produce group presentations on chosen case(s)
- Present!

Bonus!

Two additional applications

1. Comparative Mortality in Scotland
2. Cause-specific mortality in the USA