Developing a daily time step individual level demographic simulation model

Andy Turner

http://bit.ly/TStpJP

Presentation slides prepared for a CSAP research cluster meeting, School of Geography, University of Leeds, UK, 2012-12-11

Outline

- Why?
- What?
- How?
- Results
- Plans and Next Steps
- Feedback

Why?

- Generally...
  - Demographic data is used in a wide range of applications
    - Epidemiology for estimating prevalence and incidence rates
    - Service planning
    - Risk management
    - Commercial
  - Census data tend to be years old by the time outputs are made available
  - Contemporary populations are assumed to be highly and increasingly mobile and fertility in terms of live births is perhaps becoming more variable

- Demographic forecasting is important
  - Planning is a key to sustainability
    - Dependency ratios are increasing in many countries with increasing aged populations
    - Pensions
    - Welfare
    - Services and infrastructure
- Many countries (including the UK) do not have official residential registration data that tracks the location of people (between censuses)
- Our ability to track where everyone has lived improves continually, but we need the models and the data to forecast and provide the best estimates

- Why daily time steps?
  - People tend to be born, die and move residences on specific days
    - It intrinsically makes sense to model at this resolution
    - Modelling for multiple days either misses important events or becomes much more complicated
      - Consider
        - migrations of individuals within a time period
        - births (same mother) at different times within a year
  - Allows for linkage with models of daily activity that work on sub-daily time steps

- Allows for variation in mortality, fertility and migration rates over the year to be modelled
  - Mortality, Fertility, Miscarriage and Migration are seasonal
    - Student migration
    - Holidays and fertility
    - Winter mortality
  - Power cuts/flood events and birth spikes
- Allows for new and exciting aggregate data/statistics to be produced
  - Distribution of the total number of
    - births per month in a region
    - moves per person in a year
      - maximium, minimum, average, variance

- Why individual level?
  - Data can be linked with other individual level data
    - e.g. Disease data
  - It has the possibility that other individual data can be augmented or linked
    - Linkage and substitution with "real data"
  - Everybody is different
    - Individuals have their own mortality, fertility and migration probabilities and history
      - There is scope for specifying these in the model
        - In such a way as to keep overall counts of births and deaths at observed levels
      - Return migration

What?

- Stages of development
- The nature of the model
- Initialisation
- Daily Simulation
  - Death
  - Birth
  - Migration
- Results

Stages of Development

- Natural change Simulation Model
  - ESRC funded GENESIS Project
  - Leeds Output Area level results produced
- e-Infrastructure
  - JISC funded NeISS Project
  - Web Portal based User Interface
  - Simulation Models configured, run and results stored on e-Science resources
- Migration model component
  - Not externally funded
  - Developed since July 2012

The nature of the model

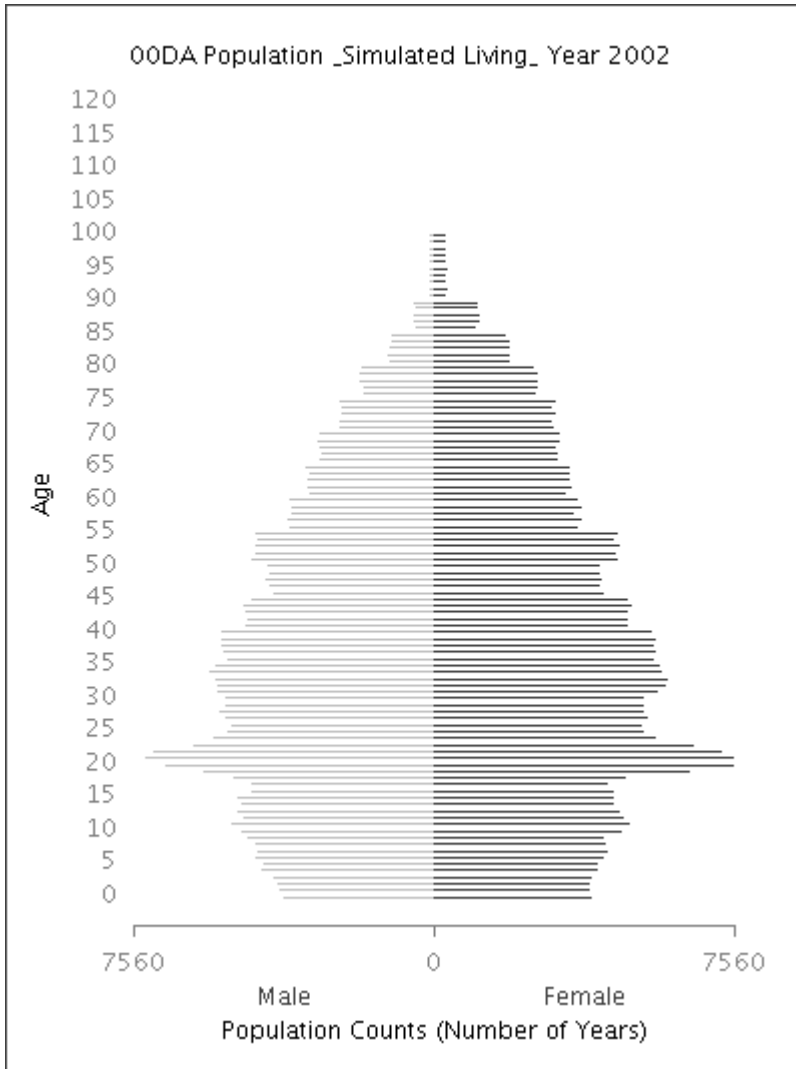
- Open Source
  - Development repositories
    - Sourceforge
      - https://sourceforge.net/p/neiss/code/328/tree/genesis/
    - University of St Andrews
      - https://e-research.cs.st-andrews.ac.uk/repos/sim/projects/genesis/
      - Thanks to Alex Voss
- Java
  - http://en.wikipedia.org/wiki/Java_(software_platform)

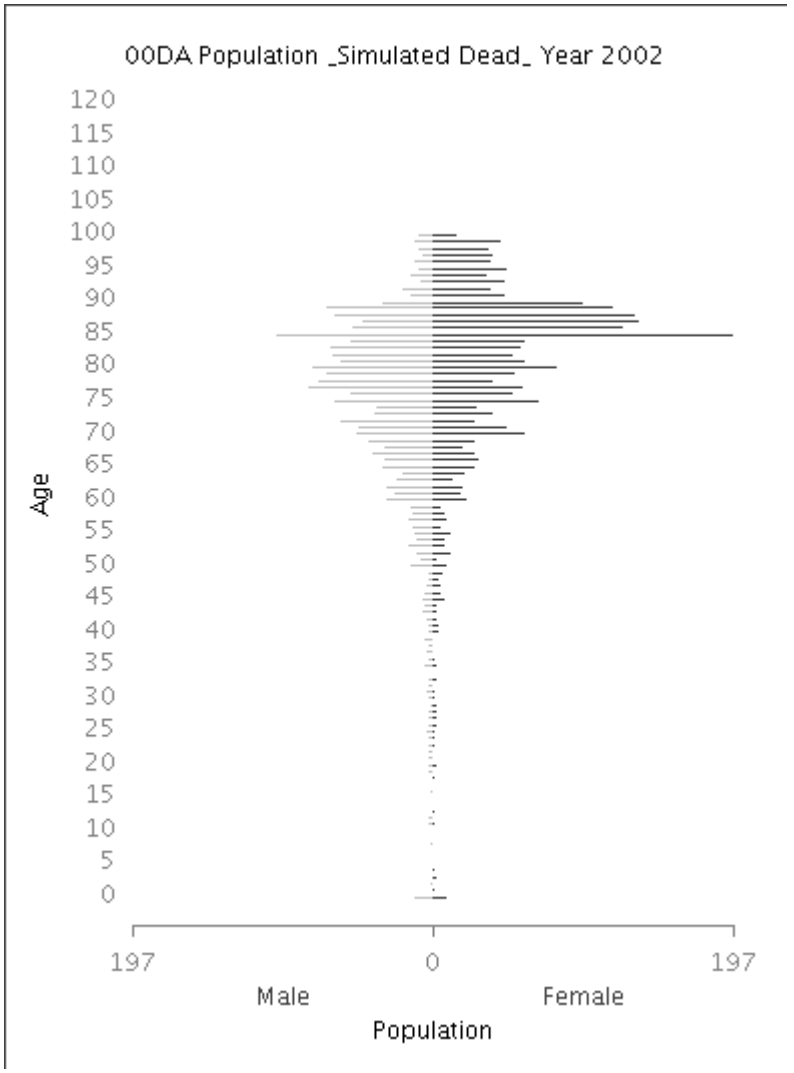
- Dependencies
  - Generic Library
  - MoSeS Code
    - For loading 2001 UK Population Census Data
- Grid enabled
  - Thanks to NeISS collaboration with Tom Doherty based at University of Glasgow
- Run for multiples of a year
- Individual representation
  - Males
  - Females

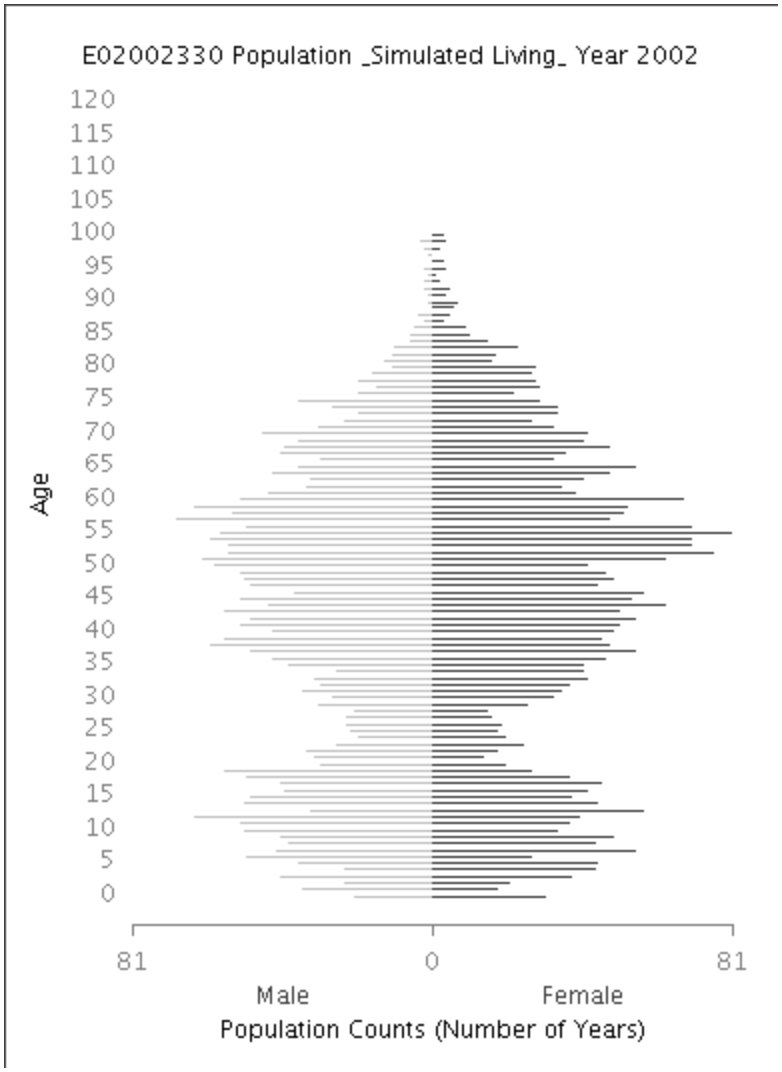
- Stochastic yet deterministic
  - Based on pseudo-random sequences
  - Results replicable
- Study Region
  - Comprised of regions and subregions
- 2 stages to modelling
  - Initialisation
  - Simulation
- Simulation proceeds for each subregion in turn, and for each individual in turn
- Synchronisation needed for each daily step

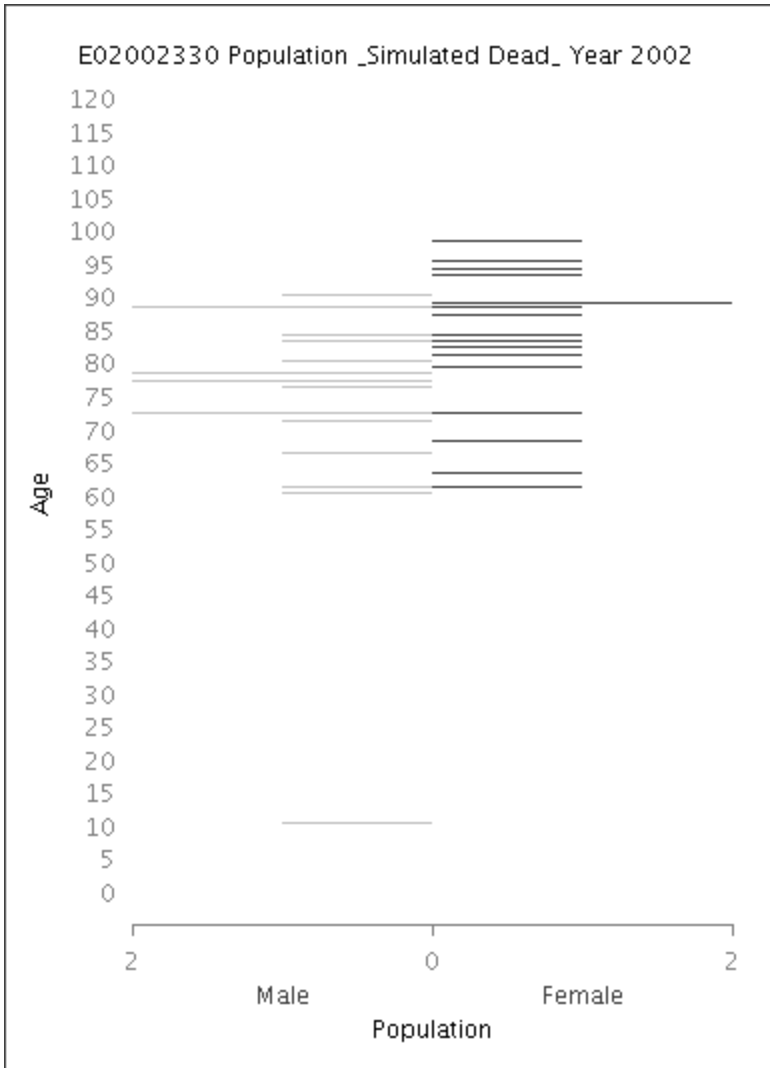
- There are many simplifying assumptions
  - Many things are assumed to be evenly distributed
  - Some things are not explicitly modelled
- There are interesting model details
  - Pregnancy and miscarriage
  - Multiple births
- Input data
  - Population count data by age and gender
  - Either birth and death counts or fertility and mortality probabilities
  - Migration data

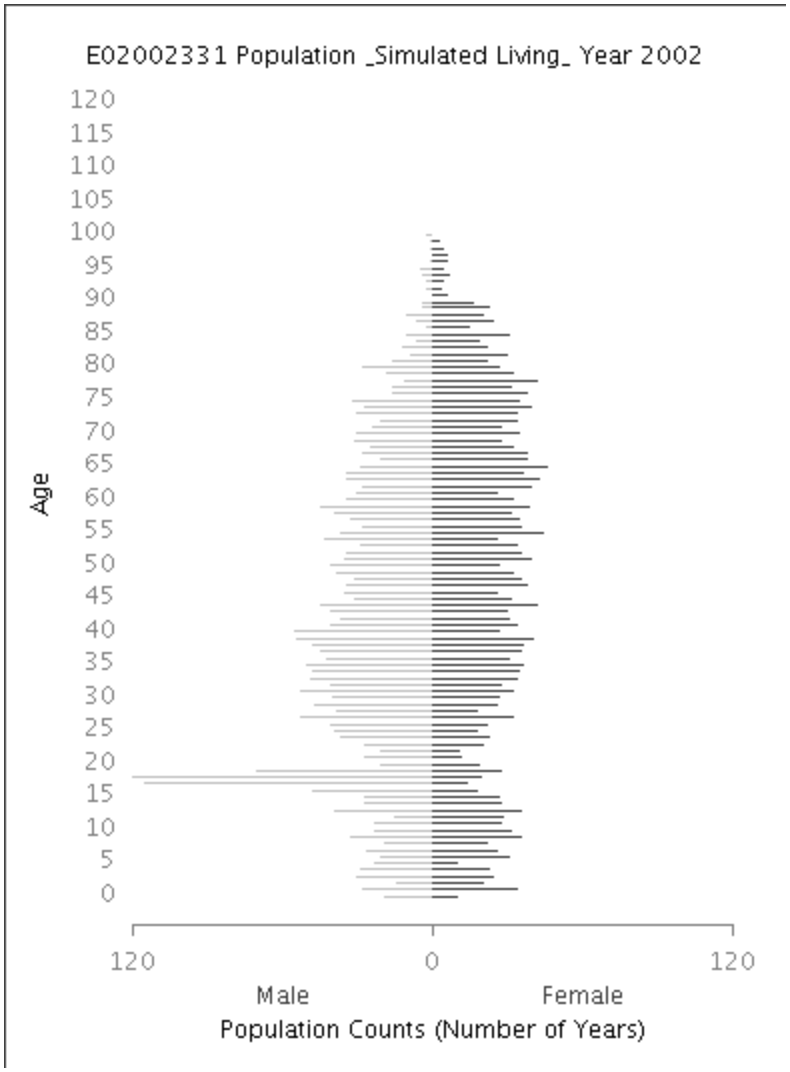
- Output data
  - Produced annually for study region, regions, subregions and aggregates of subregions
  - Includes raw ASCII data (XML,CSV), binary serialised Java object data, and images (PNG)
    - Population count estimates
    - Mortality and fertility estimates
    - Migration estimates
    - Comparisons with an annual time step model
      - Which uses mid year population estimation
    - An individual level population data set

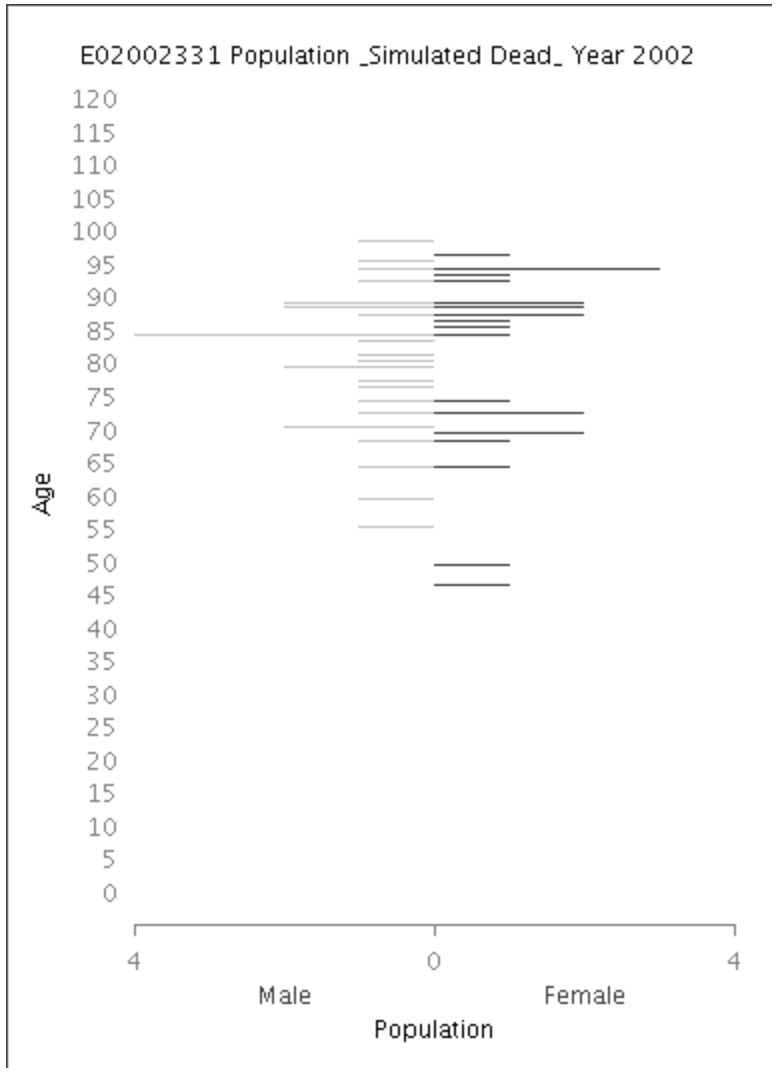

00DA Population _Simulated Living_ Year 2002
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
7560
0
7560
Male
Female
Population Counts (Number of Years)

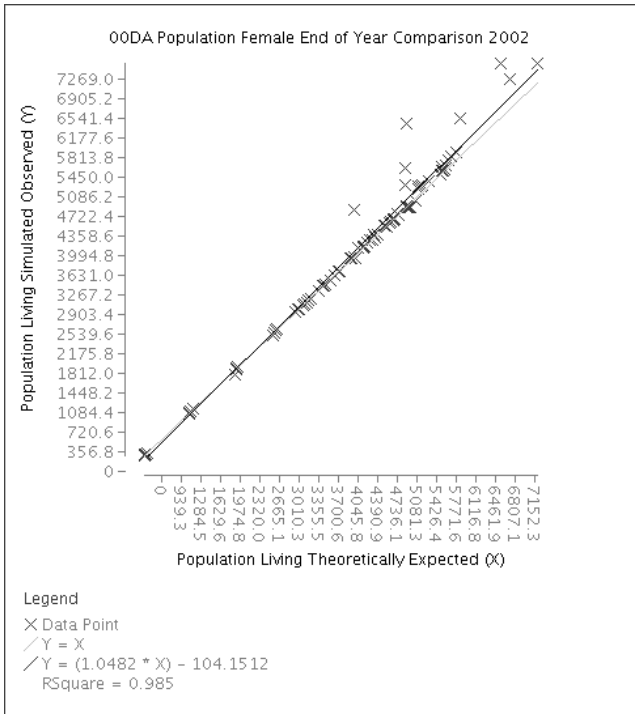

00DA Population _Simulated Dead_ Year 2002
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
197
0
197
Male
Female
Population

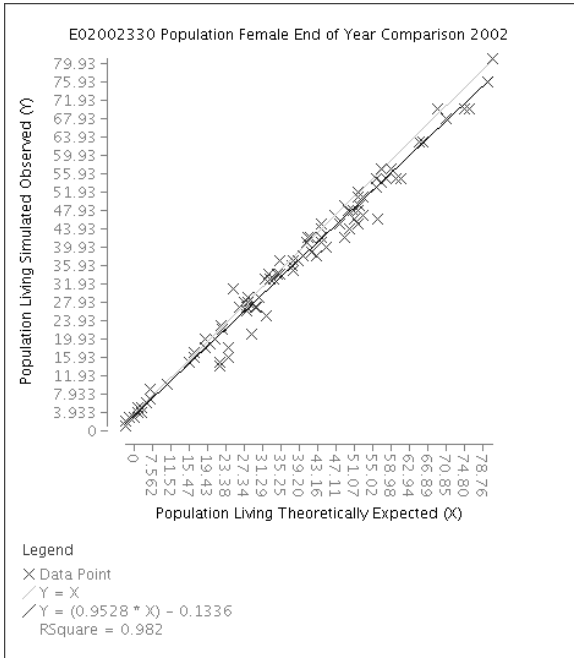

E02002330 Population _Simulated Living_ Year 2002
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
81
0
81
Male
Female
Population Counts (Number of Years)

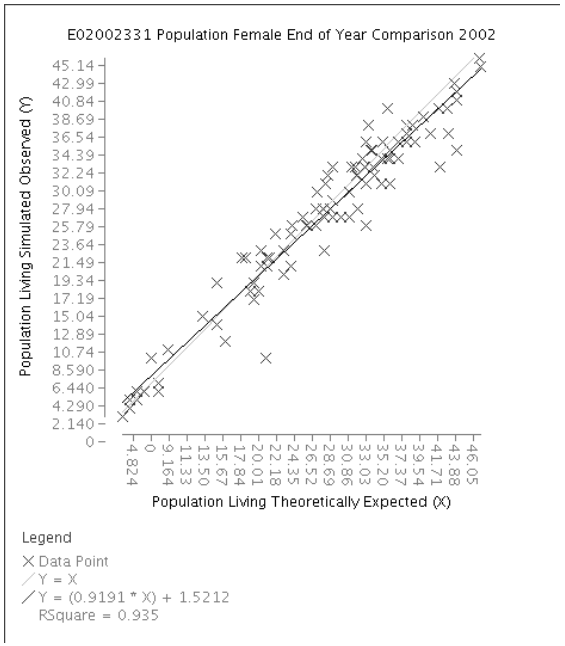

E02002330 Population _Simulated Dead_ Year 2002
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
2
0
2
Male
Female
Population

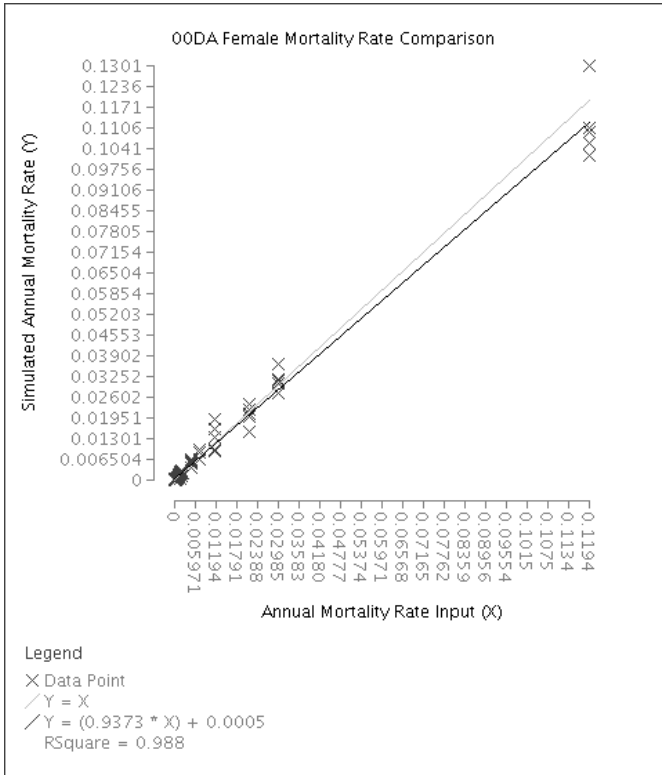

E02002331 Population _Simulated Living_ Year 2002
Age
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
120
0
120
Male
Female
Population Counts (Number of Years)

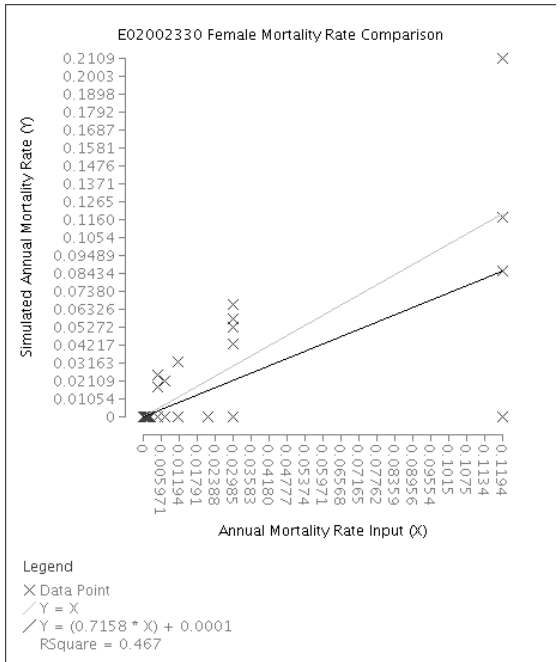

E02002331 Population _Simulated Dead_ Year 2002
Age
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
4
0
4
Male
Female
Population

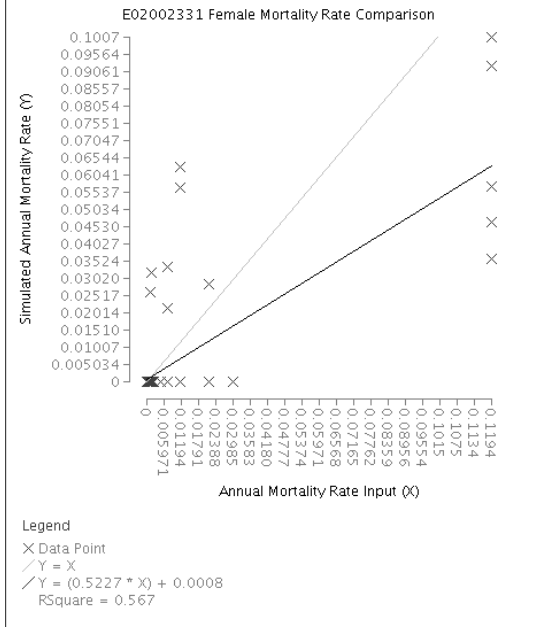

00DA Population Female End of Year Comparison 2002
Population Living Simulated Observed (Y)
7269.0
6905.2
6541.4
6177.6
5813.8
5450.0
5086.2
4722.4
4358.6
3994.8
3631.0
3267.2
2903.4
2539.6
2175.8
1812.0
1448.2
1084.4
720.6
356.8
0
0
939.3
1284.5
1629.6
1974.8
2320.0
2665.1
3010.3
3355.5
3700.6
4045.8
4390.9
4736.1
5081.3
5426.4
5771.6
6116.8
6461.9
6807.1
7152.3
Population Living Theoretically Expected (X)
Legend
Data Point
Y = X
Y = (1.0482 * X) - 104.1512
RSquare = 0.985

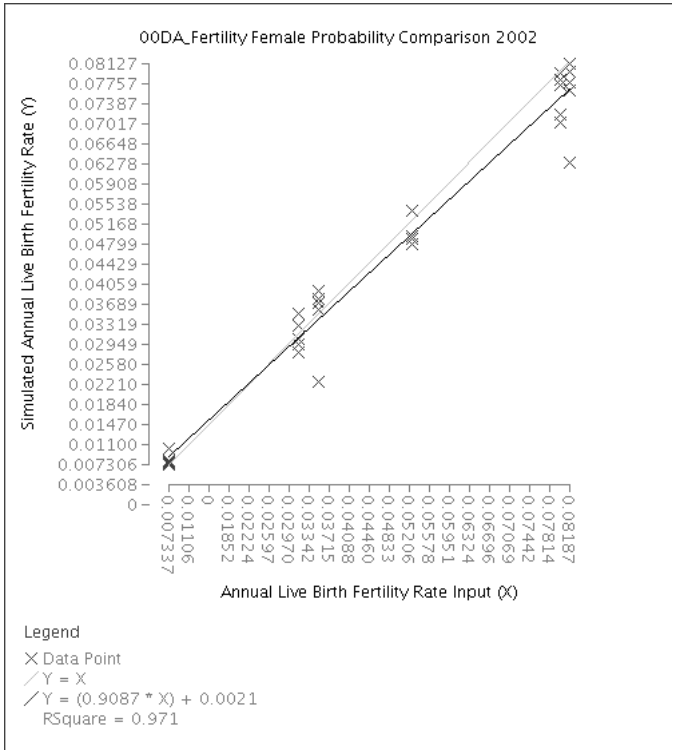

E02002330 Population Female End of Year Comparison 2002
Population Living Simulated Observed (Y)
79.93
75.93
71.93
67.93
63.93
59.93
55.93
51.93
47.93
43.93
39.93
35.93
31.93
27.93
23.93
19.93
15.93
11.93
7.933
3.933
0
0
7.562
11.52
15.47
19.43
23.38
27.34
31.29
35.25
39.20
43.16
47.11
51.07
55.02
58.98
62.94
66.89
70.85
74.80
78.76
Population Living Theoretically Expected (X)
Legend
Data Point
Y = X
Y = (0.9528 * X) - 0.1336
RSquare = 0.982

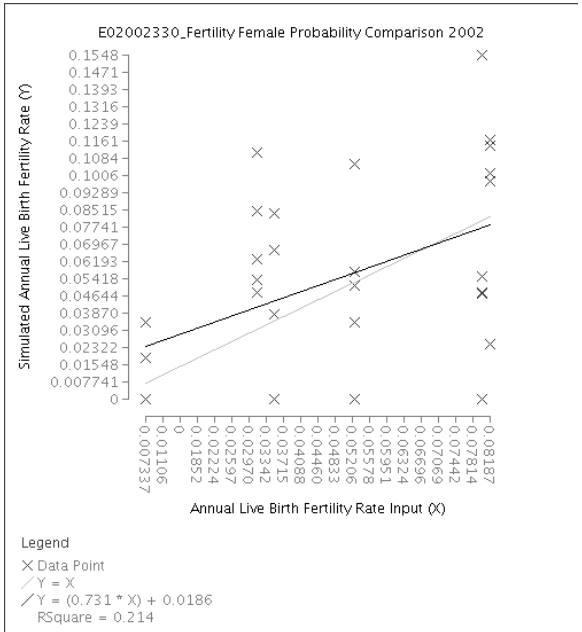

E02002331 Population Female End of Year Comparison 2002
Population Living Simulated Observed (Y)
45.14
42.99
40.84
38.69
36.54
34.39
32.24
30.09
27.94
25.79
23.64
21.49
19.34
17.19
15.04
12.89
10.74
8.590
6.440
4.290
2.140
0
4.824
0
9.164
11.33
13.50
15.67
17.84
20.01
22.18
24.35
26.52
28.69
30.86
33.03
35.20
37.37
39.54
41.71
43.88
46.05
Population Living Theoretically Expected (X)
Legend
Data Point
Y = X
Y = (0.9191 * X) + 1.5212
RSquare = 0.935

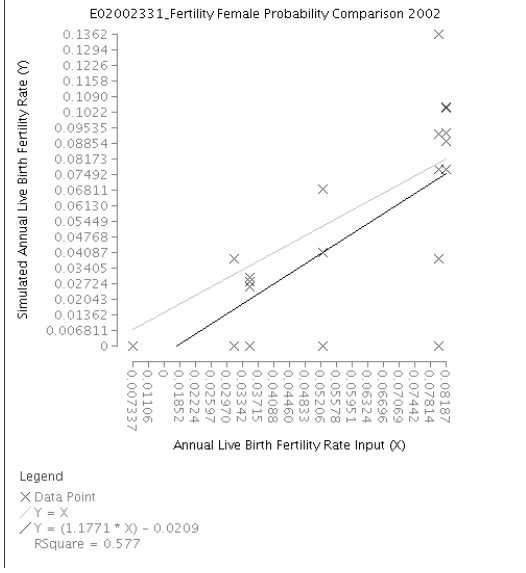

00DA Female Mortality Rate Comparison
Simulated Annual Mortality Rate (Y)
0.1301
0.1236
0.1171
0.1106
0.1041
0.09756
0.09106
0.08455
0.07805
0.07154
0.06504
0.05854
0.05203
0.04553
0.03902
0.03252
0.02602
0.01951
0.01301
0.006504
0
0
0.005971
0.01194
0.01791
0.02388
0.02985
0.03583
0.04180
0.04777
0.05374
0.05971
0.06568
0.07165
0.07762
0.08359
0.08956
0.09554
0.1015
0.1075
0.1134
0.1194
Annual Mortality Rate Input (X)
Legend
Data Point
Y = X
Y = (0.9373 * X) + 0.0005
RSquare = 0.988


E02002330 Female Mortality Rate Comparison
Simulated Annual Mortality Rate (Y)
0.2109
0.2003
0.1898
0.1792
0.1687
0.1581
0.1476
0.1371
0.1265
0.1160
0.1054
0.09489
0.08434
0.07380
0.06326
0.05272
0.04217
0.03163
0.02109
0.01054
0
0
0.005971
0.01194
0.01791
0.02388
0.02985
0.03583
0.04180
0.04777
0.05374
0.05971
0.06568
0.07165
0.07762
0.08359
0.08956
0.09554
0.1015
0.1075
0.1134
0.1194
Annual Mortality Rate Input (X)
Legend
Data Point
Y = X
Y = (0.7158 * X) + 0.0001
RSquare = 0.467


E02002331 Female Mortality Rate Comparison
Simulated Annual Mortality Rate (Y)
0.1007
0.09564
0.09061
0.08557
0.08054
0.07551
0.07047
0.06544
0.06041
0.05537
0.05034
0.04530
0.04027
0.03524
0.03020
0.02517
0.02014
0.01510
0.01007
0.005034
0
0
0.005971
0.01194
0.01791
0.02388
0.02985
0.03583
0.04180
0.04777
0.05374
0.05971
0.06568
0.07165
0.07762
0.08359
0.08956
0.09554
0.1015
0.1075
0.1134
0.1194
Annual Mortality Rate Input (X)
Legend
Data Point
Y = X
Y = (0.5227 * X) + 0.0008
RSquare = 0.567


00DA_Fertility Female Probability Comparison 2002
Simulated Annual Live Birth Fertility Rate (Y)
0.08127
0.07757
0.07387
0.07017
0.06648
0.06278
0.05908
0.05538
0.05168
0.04799
0.04429
0.04059
0.03689
0.03319
0.02949
0.02580
0.02210
0.01840
0.01470
0.01100
0.007306
0.003608
0
0.007337
0.01106
0
0.01852
0.02224
0.02597
0.02970
0.03342
0.03715
0.04088
0.04460
0.04833
0.05206
0.05578
0.05951
0.06324
0.06696
0.07069
0.07442
0.07814
0.08187
Annual Live Birth Fertility Rate Input (X)
Legend
Data Point
Y = X
Y = (0.9087 * X) + 0.0021
RSquare = 0.971


E02002330_Fertility Female Probability Comparison 2002
Simulated Annual Live Birth Fertility Rate (Y)
0.1548
0.1471
0.1393
0.1316
0.1239
0.1161
0.1084
0.1006
0.09289
0.08515
0.07741
0.06967
0.06193
0.05418
0.04644
0.03870
0.03096
0.02322
0.01548
0.007741
0
0.007337
0.01106
0
0.01852
0.02224
0.02597
0.02970
0.03342
0.03715
0.04088
0.04460
0.04833
0.05206
0.05578
0.05951
0.06324
0.06696
0.07069
0.07442
0.07814
0.08187
Annual Live Birth Fertility Rate Input (X)
Legend
Data Point
Y = X
Y = (0.731 * X) + 0.0186
RSquare = 0.214


E02002331_Fertility Female Probability Comparison 2002
Simulated Annual Live Birth Fertility Rate (Y)
0.1362
0.1294
0.1226
0.1158
0.1090
0.1022
0.09535
0.08854
0.08173
0.07492
0.06811
0.06130
0.05449
0.04768
0.04087
0.03405
0.02724
0.02043
0.01362
0.006811
0
0.007337
0.01106
0
0.01852
0.02224
0.02597
0.02970
0.03342
0.03715
0.04088
0.04460
0.04833
0.05206
0.05578
0.05951
0.06324
0.06696
0.07069
0.07442
0.07814
0.08187
Annual Live Birth Fertility Rate Input (X)
Legend
Data Point
Y = X
Y = (1.1771 * X) - 0.0209
RSquare = 0.577

Initialisation

- For each region
  - Daily survival probabilities are calculated for each age and gender
    - Death rate assumed to be even throughout the year
  - Daily pregnancy probabilities are calculated for each age of potential mother
    - Annual Live Birth Fertility Rates are factored for multiple births, miscarriage and death of mother
    - Pregnancy rate and miscarriage rate assumed to be even throughout the year

- Daily migration
  - Assumes migration evenly distributed throughout the year
  - General migration probability calculated
  - Internal migration rates are calculated for migration within the region
  - In migration rates are calculated for people moving from all regions not in the study region
- Cumulative sums of migration are calculated to help determine
  - The region destination for each out migration
  - The subregion destination for each in migration

- Each person is initialised
  - Assigned a date of birth
  - Assigned to a subregion as usually resident
  - Females are assigned pregnancies and due dates

Daily Simulation

- For each person
  - Do they die?
  - Is it their birthday?
    - If so update population statistics
  - Do they migrate?
    - If yes, find out where they move to
- For each female
  - If pregnant do they have a miscarriage
  - If due give birth
  - If not pregnant, determine if they become pregnant

- Having gone through the population for all regions in the study region
  - Migrate those migrating out of the study region
  - Migrate those migrating within the study region
  - For migration into the study region from outside of the study region
    - Create individuals
      - Assigning date of birth
      - Record migration origin location
      - Assign subregion usual resident location

How?

- Designed for (scalability) simulating large populations with large numbers of regions and subregions
  - Individual level data stored in collections which are swapped to and from slower access storage as required
  - Numerical indexes are stored in mapped collections
- Computational demands are considerable
  - Consider simulating a single region, population ~1 million, with ~10 thousand subregions
    - Can all the data be stored in the available fast access memory?

- For a simple model, a 10 year simulation might take many days with only one CPU
  - Each individual in the population is updated ~3650 times
- The amount of persistent data produced and that we want to store is in the order of tens of GigaBytes
- For a UK Simulation there are in the order of 60 million individuals and 200 thousand subregions

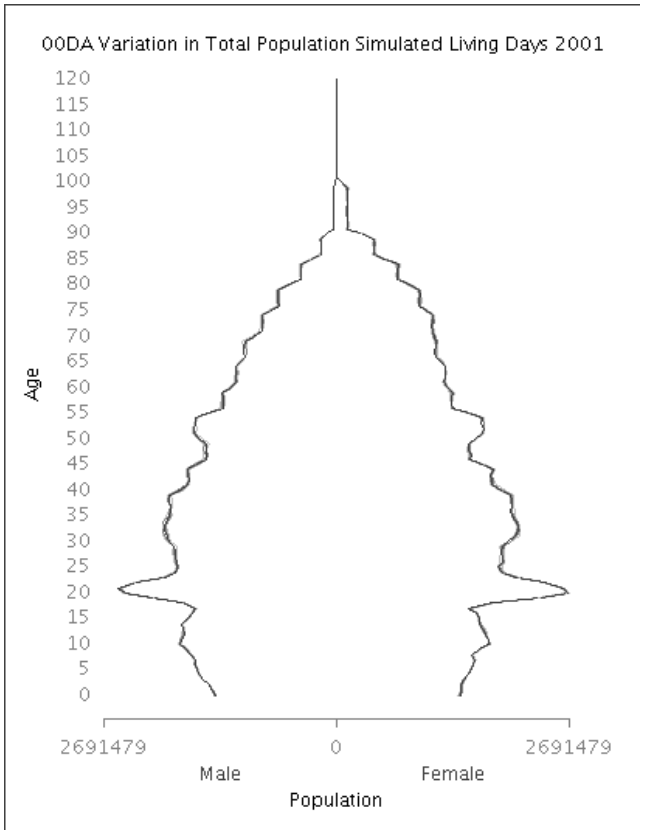
- Grid enabled
- Parallelisation
- Numerical precision
  - Java BigDecimal

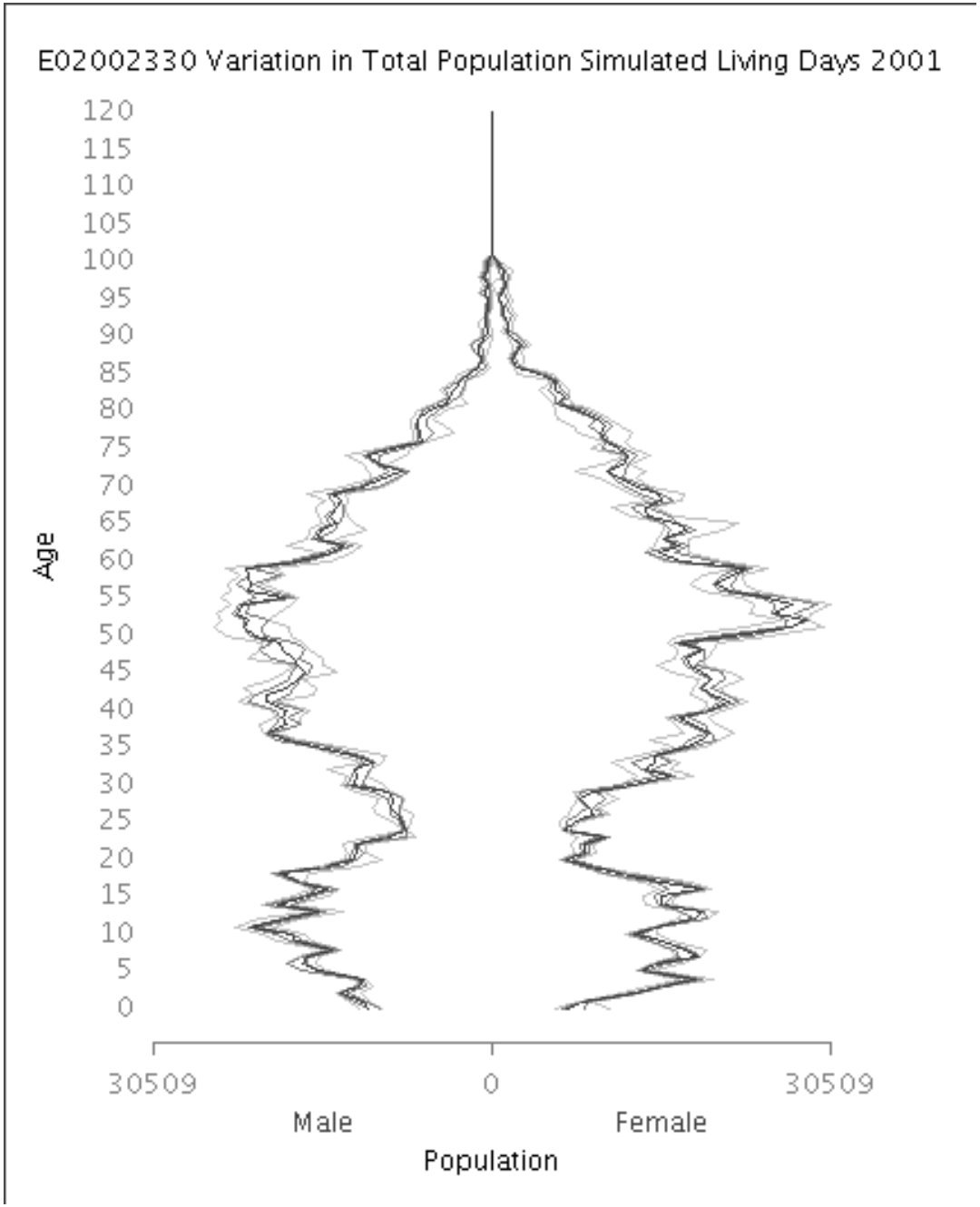
Results

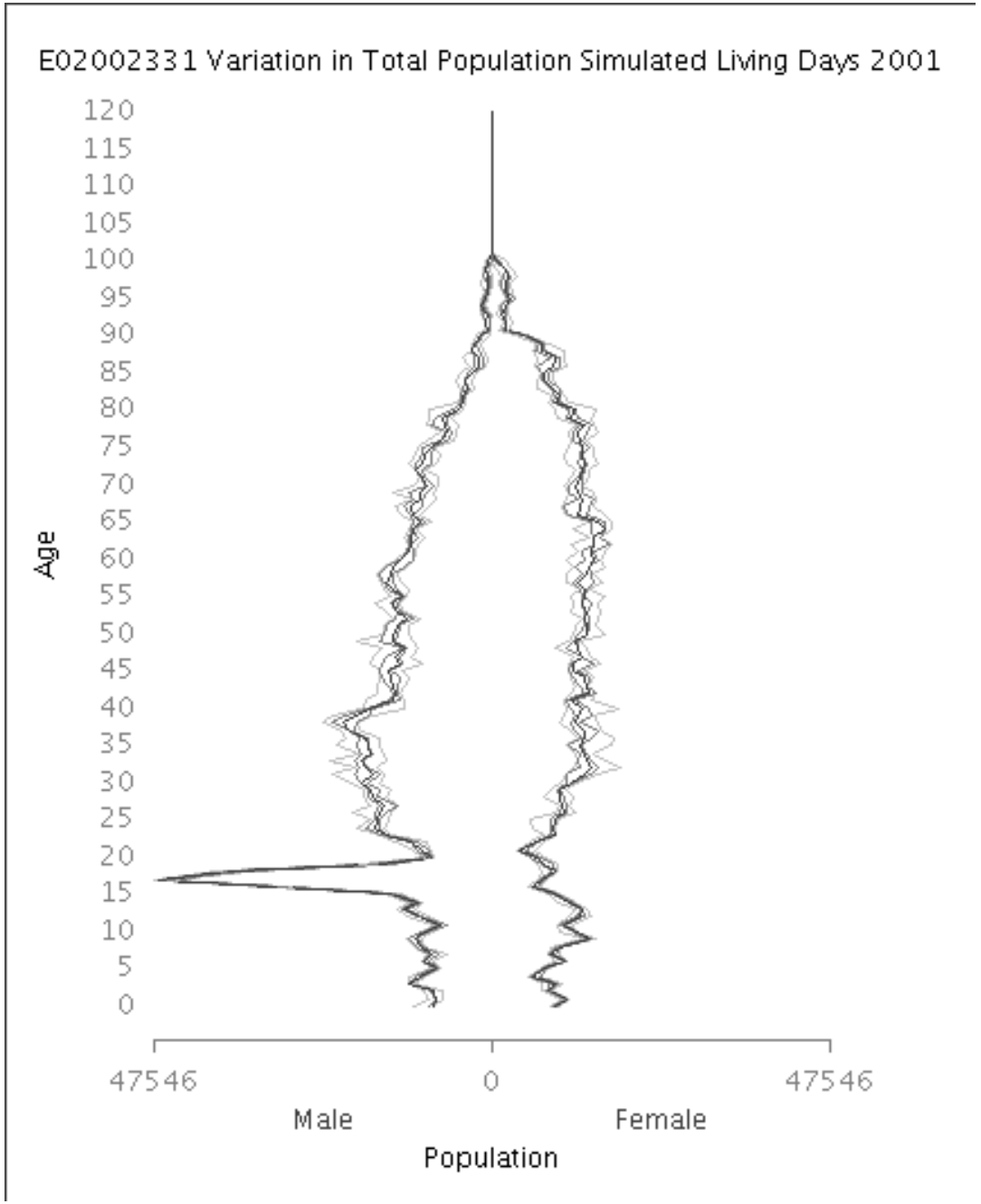
- Results for simulations without migration
  - Provide confidence in daily probability calculations for natural processes
    - The expected amounts of deaths, pregnancies, miscarriages and births result at a regional level
  - Variation
    - At sub-regional level can be large
    - At regional level are generally small
    - At aggregated sub-regional level are intermediate
    - For less frequently occurring events is greater

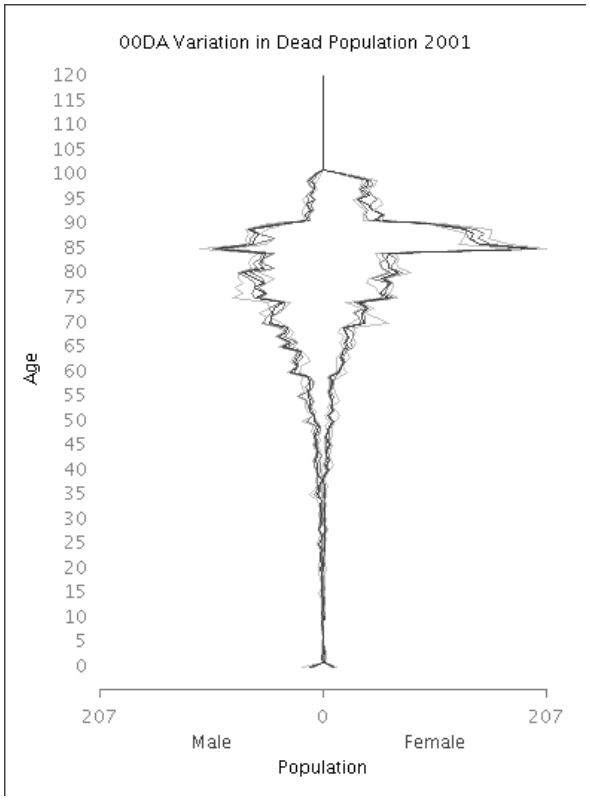
Variation in results

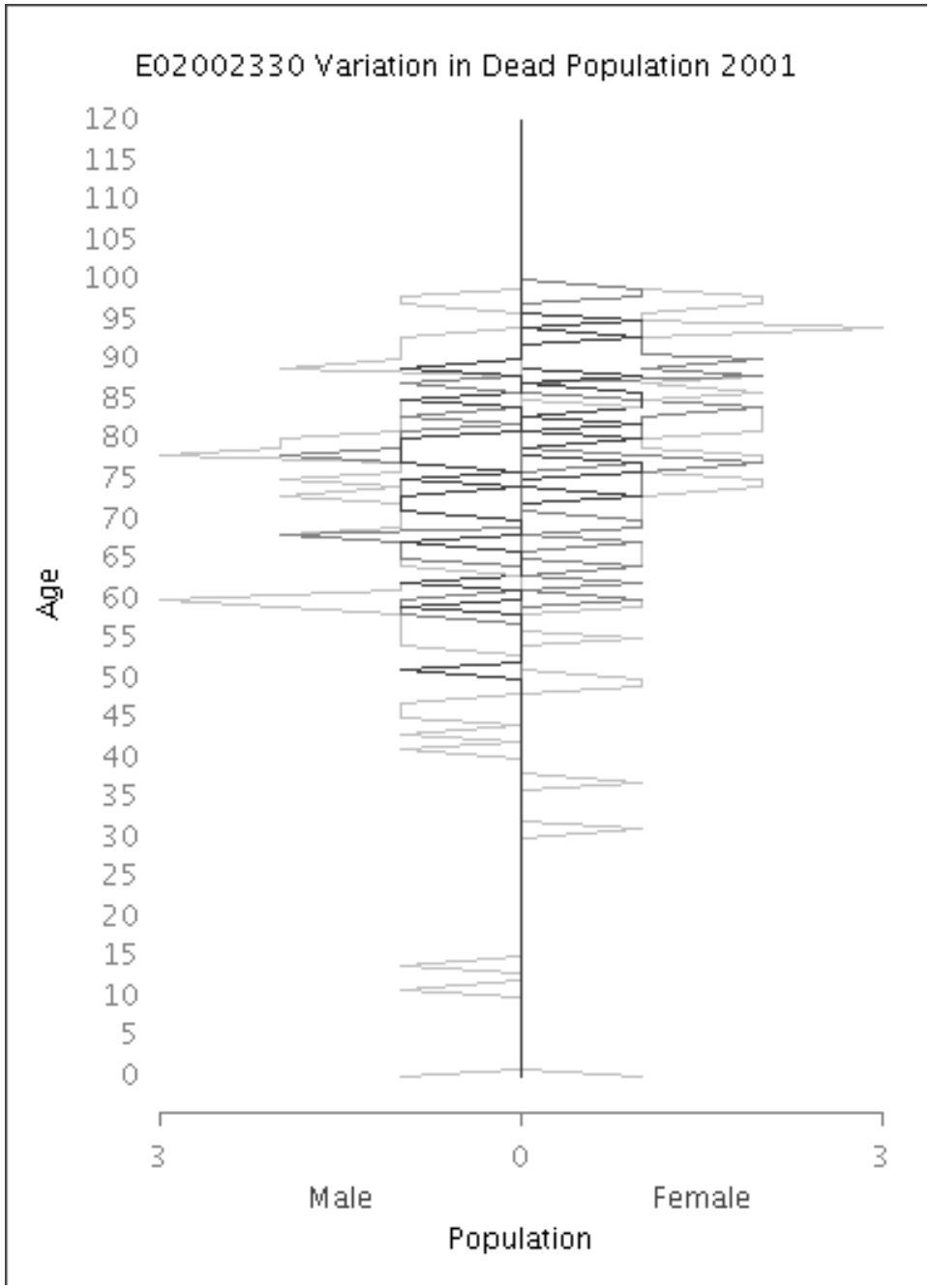
- 10 runs
  - Everything the same except the pseudo-random seed start point

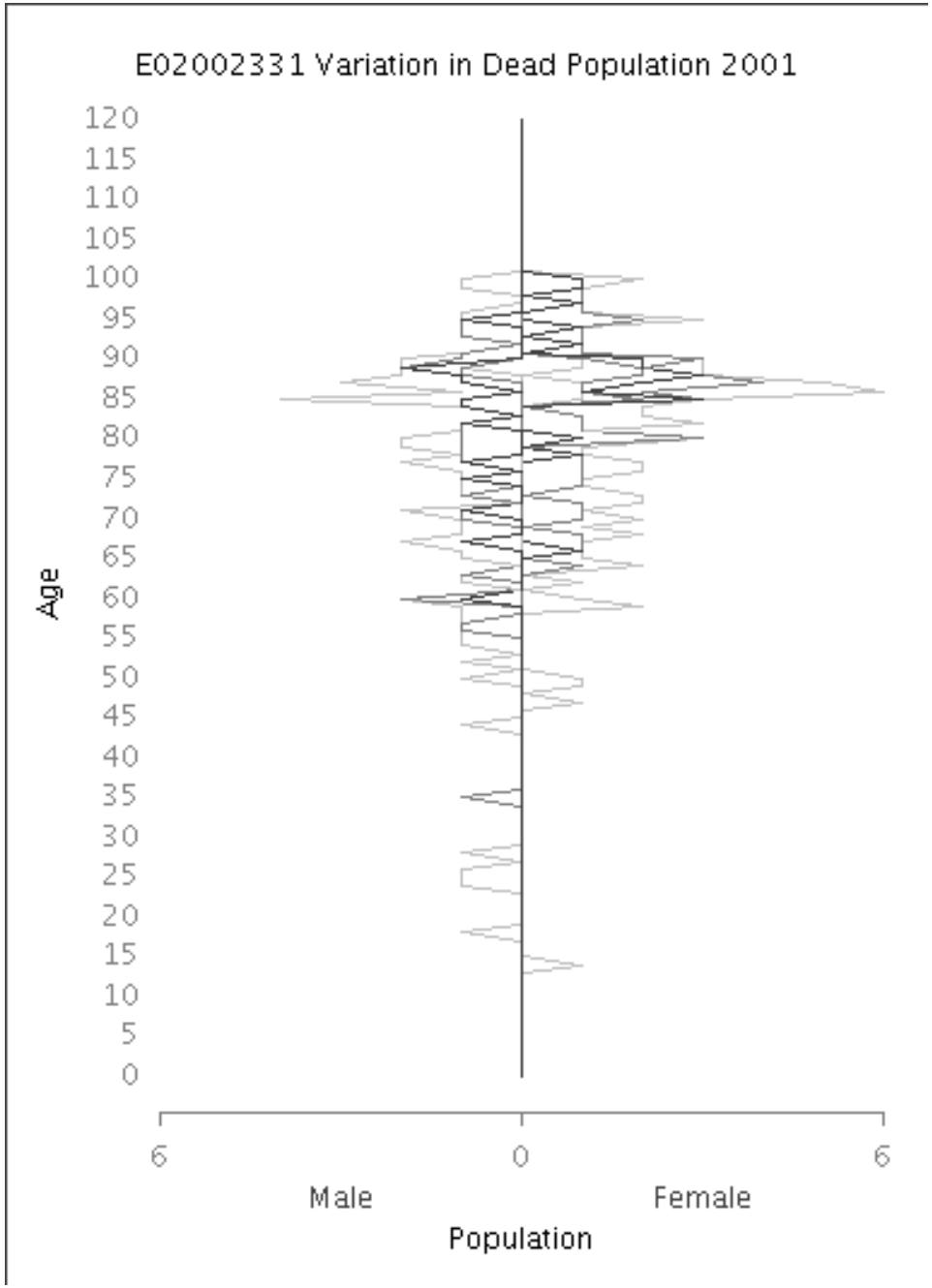

00DA Variation in Total Population Simulated Living Days 2001
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
2691479
0
2691479
Male
Female
Population

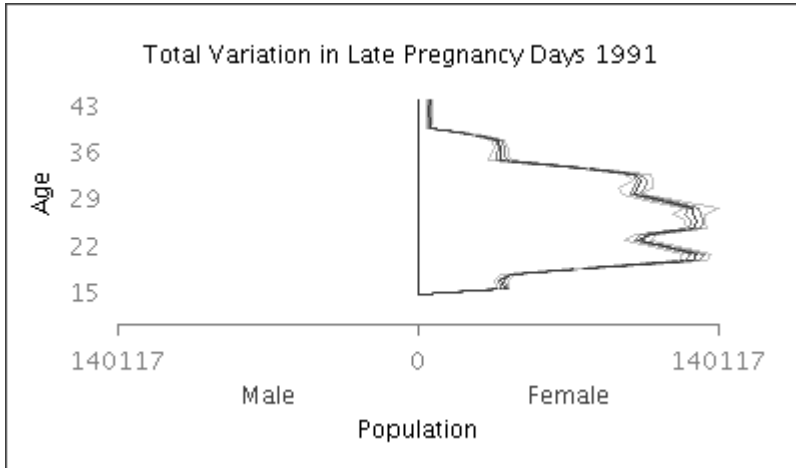

E02002330 Variation in Total Population Simulated Living Days 2001
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
30509
0
30509
Male
Female
Population

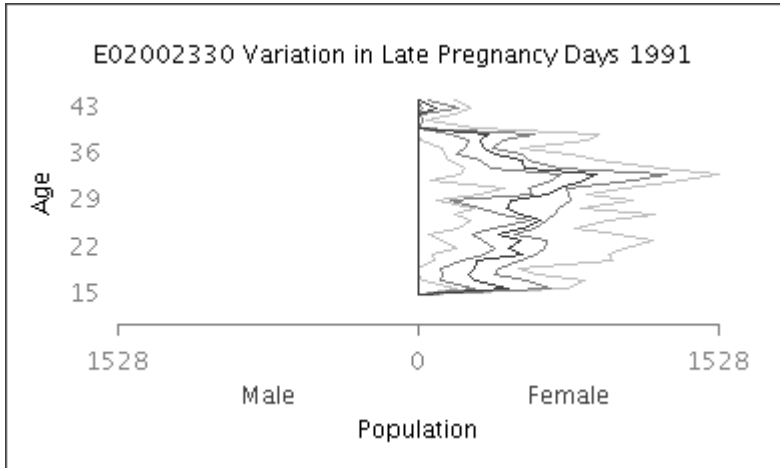

E02002331 Variation in Total Population Simulated Living Days 2001
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
47546
0
47546
Male
Female
Population

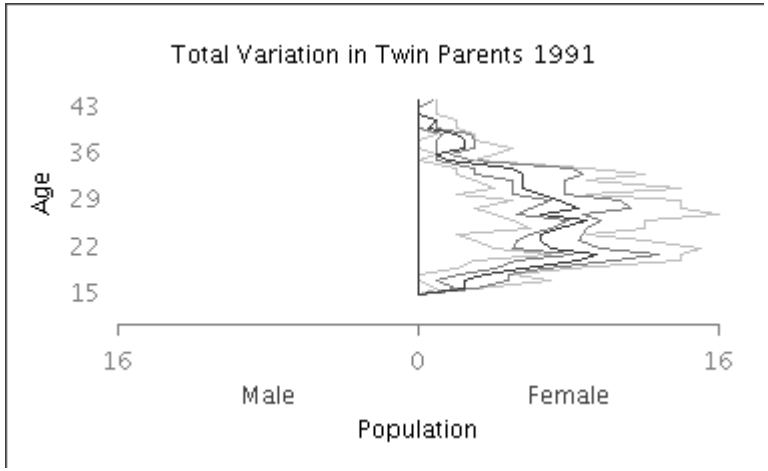

00DA Variation in Dead Population 2001
Age
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
207
0
207
Male
Female
Population


E02002330 Variation in Dead Population 2001
Age
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
3
0
3
Male
Female
Population

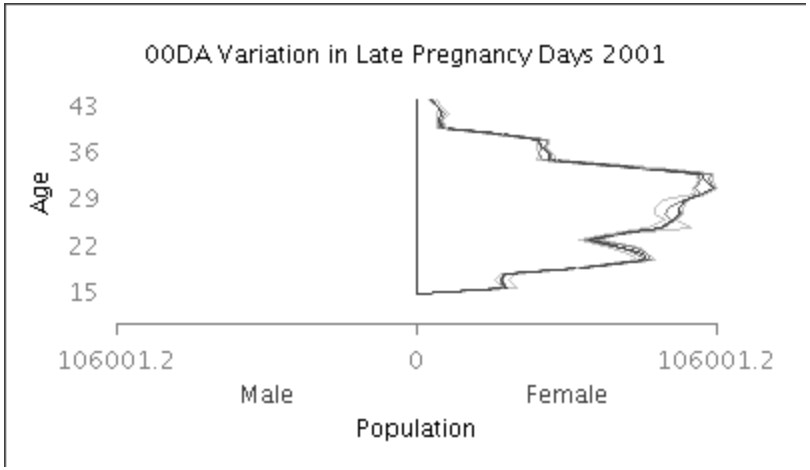

E02002331 Variation in Dead Population 2001
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Age
6
0
6
Male
Female
Population

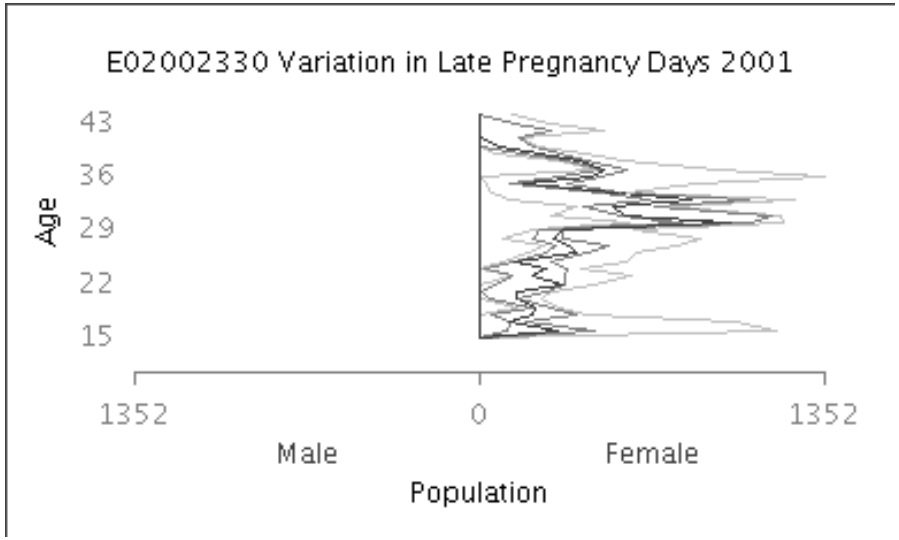

Total Variation in Late Pregnancy Days 1991
Age
43
36
29
22
15
140117
0
140117
Male
Female
Population

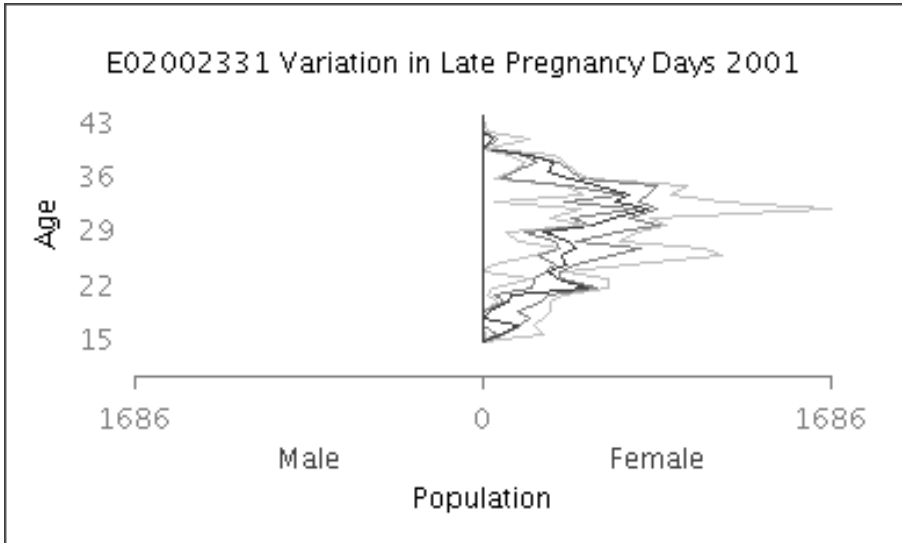

E02002330 Variation in Late Pregnancy Days 1991
Age
43
36
29
22
15
1528
0
1528
Male
Female
Population

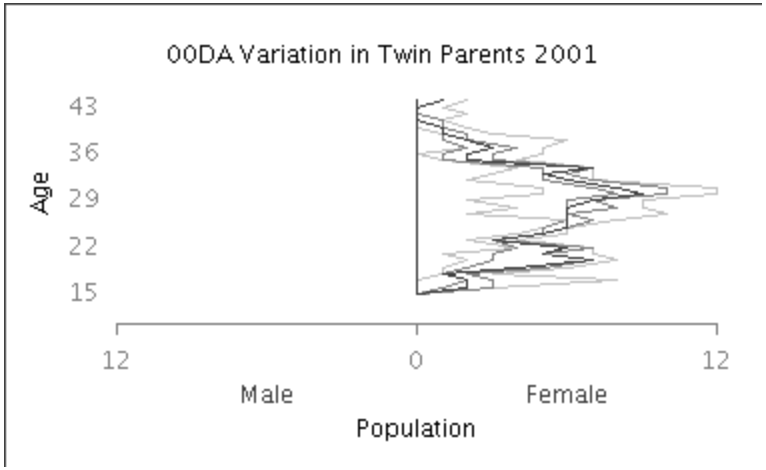

Total Variation in Twin Parents 1991
Age
43
36
29
22
15
16
0
16
Male
Female
Population


E02002330 Variation in Twin Parents 1991
Age
43
36
29
22
15
1
0
1
Male
Female
Population


00DA Variation in Late Pregnancy Days 2001
43
36
29
22
15
Age
106001.2
0
106001.2
Male
Female
Population


E02002330 Variation in Late Pregnancy Days 2001
43
36
29
22
15
Age
1352
0
1352
Male
Female
Population


E02002331 Variation in Late Pregnancy Days 2001
43
36
29
22
15
Age
1686
0
1686
Male
Female
Population


00DA Variation in Twin Parents 2001
43
36
29
22
15
Age
12
0
12
Male
Female
Population


E02002330 Variation in Twin Parents 2001
43
36
29
22
15
Age
1
0
1
Male
Female
Population


E02002331 Variation in Twin Parents 2001
43
36
29
22
15
Age
1
0
1
Male
Female
Population

Migration

- Types of migration modelled
  - Immigration
  - In migration to Study Region
  - Out migration from Study Region
  - Internal migration within Study Region
- Input data
  - 2001 UK Population Special Migration Statistics
    - LAD to LAD flows by age and gender
    - OA to OA flows by age and gender
- Region (LAD to LAD) flows are primarily used

- Subregion (OA to OA) flows are used to assign individuals to subregions with each region
- A migration factor and a minimum flow

Plans and Next Steps

- Add emigration to the model
- Detailed results statistics for migration
- Simulate population change in West Yorkshire from 2001 to 2011
  - Vary migration factor and minimum flow
  - Present results at an appropriate event
  - Publish a paper on the demographic simulation model and the results for West Yorkshire
- Simulate population change for all of England from 2001 to 2011
  - Compare results with 2011 census data
  - More publication

- Further modelling
  - Use Nik Lomax's estimated migration flows for 2001 to 2011
  - Constrain migration using subregion area classifications
  - Allow for variations in mortality, pregnancy, miscarriage and migration rates over the year
    - Student migration
  - Migrating groups (families/households)
  - Fathers

- Seek data for more detailed simulations
  - Annual and regional miscarriage data
- Seek collaboration with statistical offices
- Seek further funding
  - Secondment to UK ONS funded by ESRC?

Feedback

- Much can be done to improve this work
- What has emerged is something like the simplest demographic model
  - There is much detail to add...
- Anyone interested in writing this up or collaborating in anyway?
- Any questions?

Thank You

http://bit.ly/TStpJP