

Acceleration, alignment and matching in multi-purpose household microsimulations

Richard Cumpston

This paper has been prepared for presentation at the 2nd general conference of the International Microsimulation Association, Ottawa, Canada, June 8-10 2009. Richard Cumpston is a director of JR Cumpston Pty Ltd, consulting actuaries, and a PhD student at the Australian National University. His address is 284 Albert Road, South Melbourne, Australia 3205, email richard.cumpston@gmail.com, telephone and fax 61396827916.

Abstract

The applications of longitudinal microsimulation models can be limited by slow run times. This paper describes techniques for reducing run times, including partial sampling with short time cycles, and the use of indexed arrays for alignment and matching.

Users often expect simulations to align with deterministic projections from other sources. The paper suggests the use of random sampling within each area and age group to give one-pass alignment of births, deaths, emigrants and immigrants. An iterative process is suggested for alignment of person types.

Aligning each of 8 person types, for each of 9 age groups in 8 areas, resulted in disturbingly high numbers of household changes. Type alignment may be better done by manually adjusting exit and entry assumptions to give approximate alignment.

Many different types of matches are needed in dynamic household microsimulations. This paper suggests the use of “best of n” matching for six different applications, and gives detailed results for partnership matching. “Best of n” matching is likely to be quicker than stochastic, and allow simulations with shorter time periods and smaller regions.

The paper is illustrated with examples from a microsimulation model for Australian households, capable of operation over wide ranges of spatial resolutions and sampling densities, with projection cycles as short as a day.

All of the techniques suggested here appear to be robust, capable of being scaled up for larger applications. Run times should be broadly in proportion to person numbers.

1. BACKGROUND

As an actuary, I have been asked to help with a variety of planning issues

- Kindergarten and school locations in areas with changing populations
- Locations of churches to best use available ministers
- Route planning for very fast trains
- Locations of wood to methanol plants
- Locations of state-licensed crematoriums
- Locations for resident-funded aged-care facilities.

Some of these needs were met by deterministic projections, using similar demographic assumptions to those in the national population projections prepared by the Australian Bureau of Statistics. But deterministic projections are limited in their ability to deal with household characteristics such as income and wealth, as the numbers of cross-classifications become too high.

These limitations suggested the need for a multi-purpose household microsimulation model, capable of operation over wide ranges of spatial resolutions, sampling densities and projection cycles.

Techniques developed for well-established national models such as Canada's DYNACAN have been extremely helpful, but the different context has created some novel problems. Some of the techniques suggested in this paper may have applications elsewhere. Quantitative tests have been provided.

2. ACCELERATION

2.1 Advantages of fast run times for microsimulation models

Modern computers are so fast that there is normally little need to worry about reducing run times. But household microsimulations can involve very large amounts of data, complex calculations and surprisingly long run times.

Faster run times may allow

- Finer geographical subdivisions
- Higher sampling densities
- Modelling of events at time scales as short as days
- Closer alignment to user-specified results
- Multiple runs to provide statistical distributions of possible outcomes
- Searches for assumptions better matching available data
- Faster responses to new requests
- A wider range of government and commercial applications.

Faster runs may reduce the times required to debug new code, and to choose appropriate assumptions. Staff are likely to be more efficient and accurate if they are able to work continuously on a project, and client deadlines are more likely to be met.

2.2 Techniques used to reduce run times

- Keeping data and intermediate results in memory
- Efficient matching
- Partial sampling when using short time periods
- Periodic restacking
- Suppression of event listings
- Indexed arrays for alignment and matching
- Common-sense coding techniques.

No form of parallel processing or multi-spool operation is considered here.

2.3 Keeping data and intermediate results in memory

Substantial delays are likely if data have to be moved in and out of memory frequently, or if significant use is made of virtual memory. Very large amounts of memory are now commercially available, although there is still a price jump at 4GB. Memory costs are likely to be negligible compared with the multi-million dollar development costs of national models.

The tests described here were done with a 32 bit laptop, with 3GB of memory and 2.2 GHz clock speed. The test results were derived from a unit record sample from the 2001 Australian census, including about 175,000 persons initially. Tests with more memory and more data are planned.

2.4 Efficient matching

Many different types of matches are needed in longitudinal household microsimulations. For example, males and females are matched to form partnerships, individuals are matched to form groups, and households are matched to dwellings.

In section 6 of this paper I suggest the use of “best of n” matching. This is likely to be quicker than the stochastic matching suggested by Easter & Vink (2000), but give broadly similar results.

2.5 Partial sampling when using short time cycles

Household microsimulations require many different types of event to be simulated. For example, a random number between 0 and 1 may be generated for each person once a year, and compared with the estimated probability of that person dying. Typically, this is done by moving through the person file sequentially. Problems arise about the order in which different types of event are simulated - for example, should births be simulated before or after immigration?

These order problems become much less significant if shorter time cycles than a year are used for projections. Short time cycles also allow more complex events to be simulated – for example, the protracted bidding and price review processes often occurring when dwellings are sold. A program option was created to allow any length of projection cycle between a day and a year.

When using short time cycles, it may be time-consuming and unnecessary to test each individual each cycle. For example, suppose that monthly time cycles are being used, and there are 240,000 persons in the simulation. Then 20,000 persons could be randomly selected each month, and tested for death using probabilities 12 times their probabilities of death in the month.

2.6 Periodic restacking

Deaths and emigration result in inactive person records, and building demolitions result in inactive dwelling records. These inactive records may gradually increase the time required for each projection cycle, as well as increasing the memory requirements.

A facility was thus created to optionally restack the person and dwelling records at selected intervals, deleting the inactive records. This restacking changed the addresses of nearly all the person and dwelling records, making debugging of household changes more difficult. Restacking may help speed up production runs, where debugging is not an issue.

2.7 Suppression of event listings

For each type of event simulated, an output file was created giving sufficient details to verify the probability calculations and trace the consequences of each event. These files were large, and an option to suppress their creation gave significant time savings.

2.8 Indexed arrays for alignment and matching

The person array was indexed so that all the persons contributing to a particular alignment total could be readily identified. These indexes allowed persons in each of 8 areas (the Australian states), subdivided by 8 person types and 9 age groups, to be identified. These indices were very useful in selecting potential matches, as they helped ensure that each person selected was eligible for the match. Indexes also allowed dwellings to be identified by statistical division and occupancy status.

2.9 Commonsense coding techniques

- Choosing variable types giving only the required level of precision
- Carrying out calculations to the required level of precision
- Pre-calculating probabilities based on ordinal variables, and storing as arrays
- As far as possible, making repeated calculations outside nested loops.

2.10 Examples of run times

Table 1 : Examples of run times

Years	Periods per year	Aligned	Run times in seconds
50	1	No	239, 241, 252
50	1	Yes	219, 221
50	52	No	253

Although there is some random variation between run times, the times for the two runs with alignment are surprisingly low. Taking into account the additional household changes required for type alignment, run times might be expected to increase by about 15% when alignment is used. Using projection cycles much shorter than a year should increase run times marginally. These times exclude data input and input (about 7 seconds and 12 seconds), and output of movement data. Only demographic changes are modelled (births, deaths, emigration, immigration, household moves, exits from households, and moves to and from nonprivate dwellings).

3. GENERAL COMMENTS ON ALIGNMENT

3.1 Alignment needs for household microsimulations

Morrison (2006) gave an interesting history of the evolution of alignment over the three decades in which longitudinal microsimulation had served as a practical policy input. From his paper, there appear to be two broad needs for alignment

- Adjusting assumed probabilities of individual behavior to give overall results consistent with beliefs about the future
- Elimination of stochastic variation.

3.2 Consistency with beliefs about the future

Governments may have national population projections that underpin planning in many areas. Household microsimulations for policy purposes need to have assumptions consistent with those in the national projections, and give comparable results to them.

As Morrison notes, even when considerable care and expertise are used in estimating behavioural equations from past data, very rarely do the microsimulations match expectations about the future.

Harding (2007, p 12) asks whether aligning everything at a very disaggregated level may reduce the predictive usefulness of the dynamic model, by imposing upon the micro results predetermined macro outcomes.

3.3 Use of alignment to eliminate stochastic variation

Morrison (2006, page 16) comments that clients almost invariably prefer point estimates, and do not usually value information about likely distributions of results. As an actuary advising non-life insurers, I used to encounter a similar attitude. But regulatory authorities for insurers now want information on the range of possible results, including unlikely but costly disasters. Many investors want to understand the risks and potential returns from each investment, as well as the risk reductions from diversification.

The use of alignment to eliminate stochastic variation may thus be concealing valuable information from clients. Note however that stochastic variation during a projection may only be a minor source of overall uncertainty. Uncertainties inherent in the estimation of model coefficients, and changes over time in individual behavior, government policy and economic conditions may be much greater sources of uncertainty.

3.4 Obtaining reasonable projections without alignment

Morrison (2008, p20) notes that DYNACAN had been misinterpreting mortality coefficients inherited from CORSIM, and that alignment had masked the misinterpretation. This appears to be one of the risks of using alignment, where a major underlying error may lead to a variety of distortions. It seems desirable to obtain microsimulation results reasonably comparable with national projections, before applying alignment.

The results from one such comparison, for a 50 year projection of a 0.1% sample of the Australian population, are described by Cumpston (2007, p7). The small sample size meant that there were appreciable variations between runs, and the averages of 20 runs were used to compare total births, deaths, immigrants, emigrants and persons with national projections after 50 years. Detailed information about the assumptions underlying the national projections had to be obtained, and a number of program errors corrected, before close fits were obtained.

A wide range of problems can cause microsimulation results to differ from national projections. For example, low births may result from low partnering rates or wrong age assumptions about immigrants. Wherever possible, it seems better to correct such problems early, rather than during alignment.

3.5 Aligning to pool totals

Morrison (2006, pages 6-7) notes that a 90-year DYNACAN run involves more than 22,000 alignment pools, and more than 53 million prospective events. Assuming these are annual pools, there are about 250 pools a year, with about 2,400 prospective events per pool. The initial size of the DYNACAN sample is over 200,000 persons.

For a sample of 175,044 Australians, alignment was carried out separately for 8 areas (each state and the Australian Capital Territory). Births in each cycle were aligned for 4 age groups (15-24, 24-34,35-44,45-54), and deaths, emigration and immigration were aligned over 9 age groups (0-14,15-24, ...75-84,85+). The percentages of persons of each type at the end of each cycle were aligned over 8 person types (partner, lone parent, child, other related person, unrelated person, lone person, group member & non-private resident), 8 areas and 9 age groups.

Table 2 : Alignment pools for Australia

Description	Number areas	Number ages	Number types	Number pools
births	8	4	1	32
deaths	8	9	1	72
emigrants	8	9	1	72
immigrants	8	9	7	504
person types	8	9	8	576
Total				1256

3.6 Difficulties in obtaining alignment totals

Numbers of persons, and births, deaths, emigrants and immigrants were aligned against national projections (Australian Bureau of Statistics 2003). As the necessary details by age group and area were not published, they were obtained from a deterministic projection program very closely replicating the national population projections. These approximate alignment totals led to some additional emigrants and immigrants, and moves between areas, in order to achieve exact alignment.

3.7 Derivation of initial population sample

The Australian Bureau of Statistics estimated the resident population of Australia at 30 June 2001 as 19.482m (after taking into account the results of the August 2001 census). Their 1% sample of the census returns gave unit records for 188,013 persons, living in 75,451 households. As the sample was based on location at census night, rather than usual location, there were some incomplete households, with missing partners, or children without adults. Omitting clearly incomplete families left 175,044 persons.

4. EVENT ALIGNMENT

4.1 Sampling by sorting, and alignment by sorting

In 1995 Baekgaard suggested a “sampling by sorting” alignment method where events are randomly simulated, and any departure from the alignment total corrected by reversing the outcomes of those events where the generated random number is closest to the probability of occurrence (see Baekgaard 2002 p12).

In 2001 Johnson proposed “alignment by sorting”, a method also involving reversing some of the simulated events. For each prospective event i , a value v_i is calculated, where

$$V_i = f(r_i) - f(p_i)$$

$f(x)$ has the form $-\ln((1-x)/x)$
 p_i is the prospective event's probability
 r_i is a random number drawn from a (0,1) distribution.

The prospective events with the lowest v_i values are then implemented to match the alignment total. In 2006, this was DYNACAN's primary alignment method (Morrison 2006 page 6).

4.2 Alignment of single-outcome events using random selection

If random selection is being used to select persons for event testing, then alignment can be included in the process

- Obtain the desired event total for the projection cycle (for example, the desired total for a month might be one-twelfth of the expected number for the year)
- Randomly select an individual to be tested for the event
- Compare a random number with the event probability for that individual, and decide if the event occurs

- Stop the process when the desired event total is reached.

Table 1 shows that random selection gives similar results to alignment by sorting. It avoids any form of sorting or probability transformation. Random selection is used here to align births, deaths, emigrants and emigrants, but lack of alignment totals for person type changes made it unusable for type alignment.

4.3 Tests on three different event alignment methods

Table 3 : Average ages at death

Alignment method	Not aligned	Aligned to 63% of mean	Aligned to 100% of mean	Aligned to 127% of mean
Sampling by sorting	77.57	86.88	76.93	69.16
Alignment by sorting	77.57	77.82	77.53	77.21
Random selection	77.24	77.22	77.27	77.06

Table 1 shows the average ages obtained at death, initially without alignment, and then aligning deaths to 63%, 100% and 127% of mean deaths. For the first two methods, 20 different simulations were made of deaths in a year from 20,000 persons, giving about 3,200 deaths for both methods before alignment. For the third method, 3 simulations were made of the deaths in a year from 175,044 persons, giving about 3,900 deaths before alignment. The persons tested were Australian population samples, of all ages.

Sampling by sorting only works reasonably for alignment totals close to expected. Because alignment is done by reversing the outcomes of those events where the generated random number is closest to the probability of occurrence, most reversals occur at the young ages. An alignment total below the mean thus leaves most of the old deaths unchanged, and gets rid of many of the younger deaths, giving a high average age on death.

The probability transformation used in alignment by sorting means that reversals occur where the difference between the log of the probability and the log of the random number is small. Reversals should thus be shared fairly evenly between young and old deaths, leaving the average age at death stable.

Alignment by random selection does not involve any event reversals, and should not alter average ages. Trials showed no correlation between the between the ages at death of consecutively simulated deaths, or between their file positions.

4.4 Pre-generation of immigrant families

Immigrants to Australia often come as family groups, rather than as individuals. Simulating immigrant individuals in a short time cycle for a small area, and then trying to form these into families, may produce implausible families. Instead, immigrant families for each area are pre-generated for the next year, matching assumed numbers of persons of each age group and type for each area, and randomly sampled without replacement during each cycle in the year.

5. ALIGNMENT OF PERSON NUMBERS, AGES & TYPES

5.1 Multi-stage alignment of numbers, ages and types of persons in each area

Where alignment by number, age and type was needed for each area, it was found easier to do this in a three-stage process

- Alignment of the total numbers of persons in each age group, for Australia as a whole
- Alignment of the total numbers of persons in each age group, for each area
- Alignment of the total numbers of persons of each type in each age group, for each area.

5.2 Numbers of household changes without and with alignment

Table 4 : Household changes without & with alignment

Year	Normal exits (un-aligned)	Normal exits (aligned)	Immigrants to align	State exits to align	Type exits to align
0	0	0	892	1327	12949
1	7588	7546	346	1134	9649
2	7766	7905	530	1178	9002
3	7792	7970	484	860	8945
4	8014	8213	403	1009	9137
5	8190	8505	563	1082	9764
Average	7870	8028	465	1053	9299

Most of the entries in Table 3 relate to exits, where an “exit” is a departure from a household of a person, possibly followed by one or more of the other household members. By contrast, a “move” is where the whole household moves to another dwelling. A “normal” exit is one simulated using the assumed probabilities of exit, rather than an exit artificially created for alignment purposes. Normal exits are about 5% of the initial sample population of 175,044 persons. The numbers of normal exits grow broadly in line with population increases, and are only slightly higher when alignment is being used.

5.3 Initial alignments to correct census under-reporting

Table 3 shows large numbers of alignments at “year 0”, ie at the start of the projection. These alignments are needed to correct the under-reporting inherent in the 1% census sample used as the projection starting point. The Australian Bureau of Statistics does detailed post-census surveys to estimate the extent of census under-reporting, and uses these to publish “estimated resident populations”, the starting points for their population projections.

Three types of alignment were used at the start of the projection

- Alignment of the total numbers of persons in each of 9 age groups, done by simulating emigration by persons in over-represented ages, and immigration by those in under-represented ages (892 emigrants and 892 immigrants were needed)
- Alignment of the total numbers of persons in each age group in each of the 8 areas (done by 1,327 exits of individuals between areas)
- Alignment of the total numbers of persons of each of 8 person types, for each age group in each area (done by 12,949 exits of individuals within areas).

5.4 Country-wide alignments at the end of each projection period

When using alignment, the numbers of births, deaths, emigrants and immigrants in each age group were exactly constrained to the numbers of these events derived from Australian national population projections. In theory, the correct number of persons in each age group at the end of the year should have automatically resulted. In practice, it was not possible to fully replicate the event numbers in the national projections, and small numbers of emigrants and emigrants were simulated to correct over or under-represented age groups (generally about 400 of each).

5.5 Alignments of persons of each age group in each area

Movements between areas were simulated without constraint, using independently derived assumptions. At the end of each projection period exits between areas were simulated so as to bring the numbers in each age group into line with national projections. In each case, movements of exit “leaders” were simulated from areas where their age group was in surplus, to areas were in deficit. As exit leaders were sometimes followed by family members of varying ages, an iterative process was needed to reach equilibrium. Given the independent assumptions about movements between areas, it was pleasing that only about 1,000 simulated exits a year were needed to align area numbers. Given more knowledge of the movement assumptions underlying the national projections, fewer alignment exits might have been needed.

5.6 Alignments of persons by type

Numbers of partnership formations and breakups were not available from national projections. Alignments of persons of each type were done against the percentages of persons of each type and age group at the end of each year, which were available from Australian 25-year projections (Australian Bureau of Statistics 2004).

Alignment was done in the following sequence

- Persons in nonprivate dwellings
- Lone parents
- Partners
- Children
- Related persons in family households
- Unrelated persons in family households
- Group households
- Lone persons.

For example, if the number of partners in an area and age group was less than the alignment total, persons of the area and age group, not in non-private dwellings, and not lone parents or partners, were randomly selected, and partners found for them by “best of n” matching. If the numbers of partners were too high, partners of the area and age group were selected, and their departure from partnership simulated. Restrictions were placed on the simulations used to align partner numbers, to avoid altering the numbers of lone parents. Some exits involved multiple persons, and could alter previously aligned totals in other age groups. An iterative process was used to reach equilibrium.

5.7 Average numbers of exits needed to align misalignments

Table 5 : Average number of exits per misalignment

Year	Area misalignments	Exits to align areas	Exits per misalignment	Type misalignments	Exits to align types	Exits per misalignment
0	2362	1327	0.56	11194	12949	1.16
1	1884	1134	0.60	5278	9649	1.83
2	2046	1178	0.58	5238	9002	1.72
3	1520	860	0.57	4950	8945	1.81
4	1770	1009	0.57	5314	9137	1.72
5	1822	1082	0.59	5788	9764	1.69
Average	1808	1053	0.58	5314	9299	1.75

A "misalignment" was a difference between an actual and a target number of persons in any pool, whether positive or negative. On average, about 0.58 exits were needed to correct an area misalignment. This low number is because most of the simulated exits were of persons without followers, and each of these exits corrected a misalignment in the source area and the new area.

By contrast, about 1.75 exits were needed to correct a type misalignment. No attempt was made to direct persons from one type in excess to another type in deficit, so that something over 1 exit per misalignment was expected (after allowing for the effects of followers).

5.8 Why so many exits needed to correct type misalignments?

In part, these high numbers of type misalignments were due to the use of 576 separate alignment pools (8 states, 9 ages and 8 types). The average number of normal exits in the first 5 years was about 8,000, which was about 14 per pool. Each exit is both a departure from a pool and an entry into another, so that on average each pool had about 28 exit changes a year. The net movement of persons in or out of the pool may be distributed about $N(0,5.3)$, and the average number of misalignments may be about 4.2. This suggests that about 2,400 misalignments a year might occur if there were 576 equal alignment pools. Table 4 shows an average of 5,314 misalignments a year, and this higher number may reflect the considerable disparity in the numbers of expected exits from each of the alignment pools. The more alignment pools, the larger the numbers of likely type misalignments.

5.9 Derivation and adjustment of exit assumptions

The initial exit assumptions were derived from the first 6 waves of the Household, Income and Labour Dynamics in Australia Survey (HILDA). This longitudinal survey covers 19,914 initial persons, together with any children subsequently born or adopted. All new entrants to a household who have a child with an original member of the survey are also covered (Watson 2008 pages 2-3).

During the first 6 waves of HILDA, 5,396 persons were apparently lost due to non-response, and another 4,368 were identified as exits from a household where at least one person remained in the survey. Due to the confidentiality provisions of the survey, no current data could be obtained on persons who were in a household with no remaining persons in the survey. In particular, this meant that no information could be obtained on the exit rates of lone persons. The larger numbers of non-respondents, compared with the identified exits, cast considerable doubt on the reliability of the derived exit assumptions.

The projections cited above used exit assumptions obtained after a repeated process of making projections without alignment, comparing with the type alignment numbers, and approximately adjusting the assumed exit assumptions. This process gave type proportions that agreed closely with the person type distributions for the country as a whole, but had some significant departures for particular type/age combinations. Without this adjustment process, more alignments would have been needed to correct systematic biases as well as random deviations.

5.10 Conclusions on alignment

- Event alignment can be readily done by random selection, without distortions
- Initial misalignments are inevitable with census samples, and a systematic process for correcting them is needed (see 5.3)
- Alignment of person types during projections can produce large numbers of household changes, exaggerating household instability
- The larger the numbers of alignment pools for person types, the more household changes will be needed for alignment (see 5.8)
- If exact person type alignment is required, then alignment totals for each type of household change should be obtained, and event alignment used
- If approximate type alignment will suffice, then this can be done by adjusting exit assumptions until the desired person type distributions are approximately obtained.

6. MATCHING

6.1 Historical background

Orcutt, Greenberger, Korbel and Rivlin (1961) described the selection of equal numbers of men and women to be married in a simulation year, and a matching process to pair them off. Their microsimulation model led to the creation of the DYNASIM model.

As described by Perese (2002, p17), one version of DYNASIM involved the identification of a male to be married, and a probability-based selection of a partner from amongst 10 available females. For each female in turn, a random number was drawn, and a match made if the number was less than the calculated probability. If no match occurred, the highest-probability match was made.

Bouffard, Easther, Johnson, Morrison and Vink (2001, p6) note that the CORSIM, DYNACAN, POLSIM and SVIERGE models all used something similar to CORSIM's original "stable marriage algorithm" to form specific couples from pools of prospective marriage partners. In spite of its theoretical elegance, this was found to produce too many close partnerships, and too many distant ones.

A version of stochastic matching, proposed by Easther and Vink in 2000, calculates the relative probability of each potential pairing from two equal pools of males and females, and randomly selects a pairing on a probability-weighted basis. This process proceeds sequentially until pairing is complete. This process was implemented in CORSIM and DYNACAN.

Perese (2002, p17) proposed a sequential matching process for male and female pools, where the probabilities of marriage to a male are calculated for each remaining female, and normalized so that the highest probability is 1. For each female in turn, a random number is drawn, and a match made if the number is less than the calculated probability. As a result of the normalization, a match is always found. This process was implemented in CBOLT.

6.2 Questions about matching techniques

- Can stochastic matching be applied to select a partner for a given person, rather than to match pools of males and females?
- How should probability or compatibility be measured?
- What measures of matching success should be used?
- How many potential partners are needed to give realistic results?
- How similar is CBOLT matching to stochastic?
- Can similar results be obtained by picking the potential partner with the highest comparability, rather than using probability-weighting?
- Can the same matching techniques be used for other applications, such as matching families to dwellings?
- When should we stop creating couples?
- What do we do if there are not enough potential matches?
- Can the same techniques be used for data synthesis for small regions?

6.3 Selecting partners using "best of n" matching

The limitations of early computers may have led to matching pools of males and females, and this is still widely used. But DYNASIM and SESIM selected single individuals to be married, and then chose from 10 randomly selected potential partners. This individual approach is better suited to simulations with short time periods or small geographic regions.

Easther and Vink's stochastic matching for pools can be readily adapted to matching individuals. A single individual is simulated as partnering, a number of potential partners randomly selected, a compatibility index calculated for each of the possible pairings, and the partner selected stochastically on a probability-weighted basis. Results from applications of this approach are described in this paper, and compared with "best of n".

6.4 Probability and compatibility measures

DYNASIM apparently calculated the probability of a male and female marrying as

$$\exp(-0.5 * ((\text{age difference})^2 + (\text{education difference})^2))^{.5}$$

This expression appears to have been based on the observed tendency for partners to be similar in age and education, rather than on formal data analysis.

CORSIM and DYNACAN use compatibility measures derived from logistic estimation using US census data. Variables used include age difference and its square, difference in years of education, number of children the women has, race, labour force participation, earnings difference and some interaction effects (Bouffard et al, 2001 p5). The compatibility index is estimated using logistic regression on a potential pairs data file of recent marriages. The potential pairs may include each possible pairing of the persons recently married, as well as the pairs which did marry (Perese 2002 p9).

Logistic regression is widely used to estimate probabilities of events occurring, and the data needs to have some cases where the event occurs, and others where it does not. The addition of potential pairs is thus a device to make the regression feasible, and there is no need to add any particular number of potential pairs.

This paper describes logistic regressions fitted to five Australian data sets, where the data for each regression consists of the actual pairs, plus an equal number of records obtained by randomizing the actual pairs. The number of potential pairs affects the constant derived from the regression, but not the other parameter estimates. Log likelihood reductions and standard errors decrease with larger numbers of randomized pairs, providing only relative information. The more potential pairs used, the more spuriously significant parameter estimates will result.

Logistic regression fits probability models of the form

$$\begin{aligned} \text{score} &= \beta_0 + \sum \beta_i x_i \\ \text{probability} &= \exp(\text{score}) / (1 + \exp(\text{score})) \end{aligned}$$

where β_i are the fitted parameters, and x_i the explanatory variables.

If the potential partner with the highest probability is to be selected, then comparisons can be made between regression scores, and the value of the constant β_0 is immaterial. But if partners are to be selected on a probability-weighted basis, as in stochastic matching, then the value of β_0 will matter. The number of potential pairs added for the regression analysis will thus have some effect on the results of stochastic matching.

6.5 Measures of matching success

Bouffard et al (2001) used graphs of the percentages of couples with each age differences, as well as percentile graphs of compatibility levels. These graphs allow visual comparisons between the pairing results and population values. The authors noted (p4)

“...the synthetic marriages should resemble actual new marriages in both their central tendencies and their dispersion”

This paper compares the means and standard deviations of regression scores, the mean regression scores for couples falling in each score decile, and the means and standard deviations of differences in age, education and qualification.

6.6 Logistic model of relative probability of two persons being partners

Table 6 : Regression estimates for partnerships

Parameter	Estimate	SE	LLR
agediff^2	-0.0131	0.0001	0.3031
agediff	0.0836	0.0017	0.3373
hscdiff^2	-0.2253	0.0055	0.3591
qalldiff^2	-0.0372	0.0010	0.3721
constant	1.5688	0.0148	

The above parameters were estimated from a data file made up of 38,263 records of opposite sex partners from a 1% sample of the Australian 2001 census, plus the same number of unmatched pairs created by randomization. The LLR values are the cumulative log likelihood reductions obtained by successively adding parameters. Note that the data are for all partnerships, not just recent ones.

A logistic model was used to estimate the relative probability of a male and a female, both known to be partners, of being partners with each other. The independent variables found to give useful reductions in the magnitude of the log likelihood were

- agediff, the age of the male less the age of the female
- agediff^2, the square of agediff
- hscdiff^2, the square of the difference in the codes for the highest level of schooling completed (these codes ranged from 1 for still at school to 5 for not stated)
- qalldiff^2, the square of the difference in the codes for non-school qualification (these codes ranged from 1 for postgraduate to 8 for not stated)

The agediff and agediff^2 parameter estimates suggest that the highest probabilities of being in partnership occur when the male is about 3.2 years older than the female.

Regression scores were calculated from the fitted model as

$$1.5688 + 0.0836*\text{agediff} - 0.0131*\text{agediff}^2 - 0.2253*\text{hscdiff}^2 - 0.0372*\text{qalldiff}^2$$

Relative probabilities of matching were calculated from these scores as

$$\exp(\text{regression score}) / (1 + \exp(\text{regression score}))$$

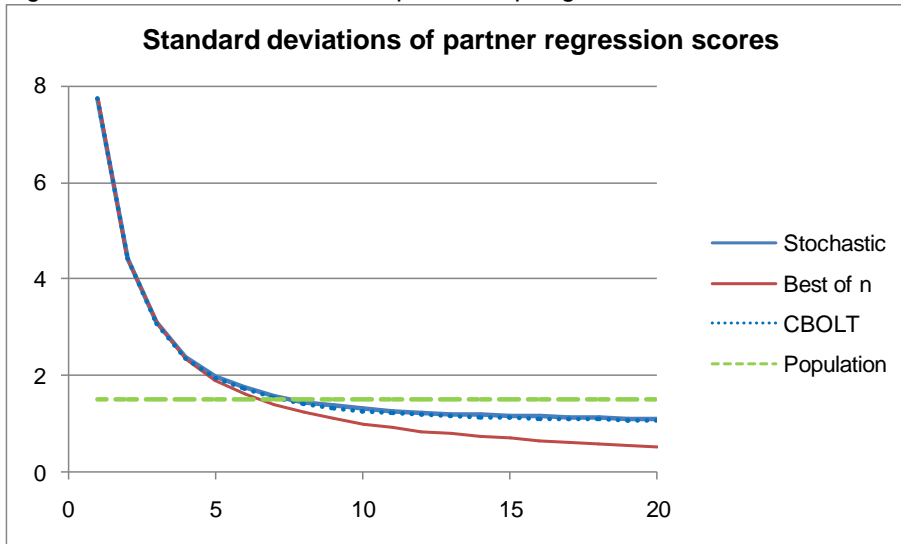
These relative probabilities were used in trials of stochastic matching. For “best of n” trials, the regression scores were used without exponentiation.

Figure 1 : Means of partnership regression scores



Even with 20 selections of potential partners, stochastic and CBOLT matching gave mean regression scores below the population mean. By contrast, “best of n” matching reproduced the population mean with about 8 selections, and continued to rise with n.

Figure 2 : Standard deviations of partnership regression scores



Stochastic and CBOLT matching gave score standard deviations matching the population with 8 selections, then stayed a little below the population value with increasing numbers of selections. “Best of n” matching replicated the population

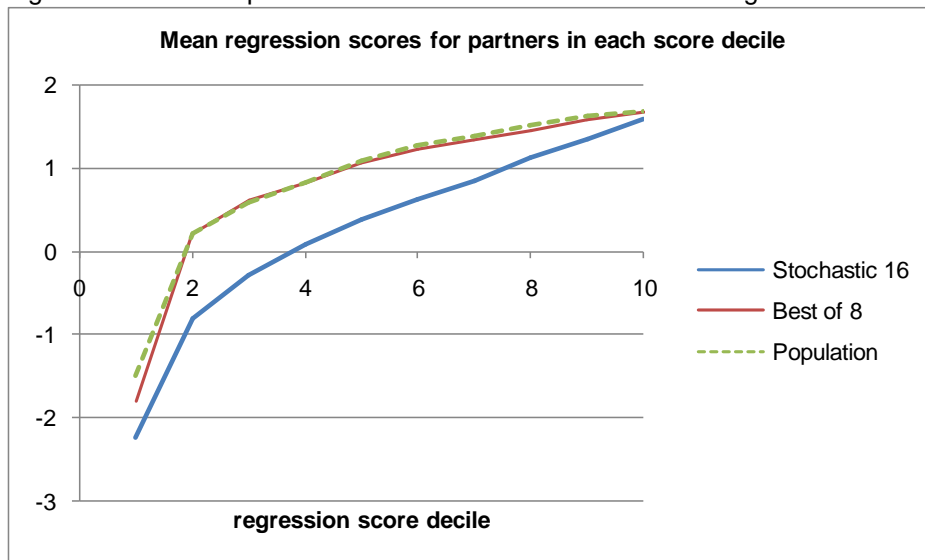
standard deviation after 7 selections, then fell well below with increasing numbers of selections.

The standard deviations with “best of n” matching can be replicated closely from the recursive relationship

$$SD(n) = SD(n-1) * [1 - 0.5 * SD(n-1)^2 / SD(1)]$$

where SD(n) is the standard deviation with n trials.

Figure 3 : Partnership score distributions with different matching methods



The above graph shows the average regression score for each decile of 40,000 synthetic partnerships formed by three different matching methods. Stochastic matching with 16 trials per match gave average scores for each decile that were below population values for all but the top decile. By contrast, best of 8 matching gave a good fit to all but the bottom decile.

6.7 Replication of component means by “best of n” partner matching

Table 7 : Component means and standard deviations for “best of n” partner matching

Number of selections	Mean of age difference	SD of age difference	Mean of schooling difference	SD of schooling difference	Mean of qualification difference	SD of qualification difference	Mean of score	SD of score
1	2.79	21.07	-0.01	1.43	-0.53	2.92	-4.92	7.71
2	2.92	14.17	-0.02	1.34	-0.52	2.78	-1.64	4.42
3	3.00	11.05	-0.02	1.25	-0.50	2.65	-0.53	3.05
4	3.07	9.26	-0.01	1.18	-0.48	2.53	0.02	2.34
5	3.04	8.07	-0.02	1.13	-0.47	2.42	0.33	1.91
6	3.03	7.25	-0.02	1.07	-0.46	2.32	0.55	1.62
7	3.04	6.61	-0.02	1.02	-0.45	2.24	0.70	1.39
8	3.06	6.15	-0.02	0.97	-0.43	2.16	0.81	1.24
9	3.07	5.77	-0.02	0.93	-0.42	2.10	0.90	1.12
10	3.07	5.46	-0.02	0.90	-0.40	2.04	0.97	1.04
Population	2.74	5.72	-0.02	0.97	-0.49	2.17	0.87	1.39

The simulated values for 1 to 10 selections were obtained by making 40,000 random choices of persons known to be partners, randomly selecting a number of potential partners for them, choosing the simulated partnership with the highest regression score, and calculating mean and standard deviations over the 40,000 trials.

The bottom line of the table shows the means and standard deviations obtained from the 38,263 partnerships for which details were available from the 1% census sample file. These results show that making eight random selections of potential partners, and choosing the one giving the highest regression score, gave means and standard deviations approximately replicating population values for partnerships. This replication occurred for age difference, education difference and qualification difference, as well as for regression scores.

There were some differences in the numbers of trials needed to replicate the standard deviation of population values

- Standard deviations of age differences were replicated with 9 trials
- Schooling and post-school qualification levels were replicated with 8 trials
- Regression scores were replicated with 7 trials.

None of the trials correctly reproduced the population mean age difference, generally giving means about 0.3 years higher. Depending on the application, this and other failures to match precisely may not matter.

6.8 Results for 6 matching applications using Australian census data

The following table shows the type of model used for each allocation, the log likelihood reduction (LLR) obtained from the model, the “score ratio”, and the number of selections considered appropriate with “best of n” matching.

Table 7 : Matching applications using Australian data

Allocation	Type of Model	LLR	Score ratio	Number selections
Partners	logistic	0.372	5.1	8
Children	logistic	0.114	5.1	5
Other	logistic	0.039	1.3	2
Unrelated	logistic	0.012	1.6	2
Groups	logistic	0.086	2.6	4
Dwellings	lognormal	0.137	1.4	2

The “score ratio” is the standard deviation of the regression score for randomly chosen non-matches, divided by the standard deviation of the regression score for population matches. For a logistic regression, the regression score is the linear combination of independent variables, before exponentiation. Similarly, for a lognormal regression of the value of a dwelling, the regression score is the linear combination before exponentiation.

The number of selections is the number required with “best of n” matching to approximately reproduce the observed mean and standard deviation of regression scores for actual matches. As a rule of thumb, the number of selections should be at least the score ratio.

6.9 Relative speeds of stochastic and “best of n” matching

“Best of n” matching can be done using regression scores, rather than the exponentials of these scores needed for stochastic matching. As a result of its asymptotic convergence, stochastic matching is likely to need more selections than “best of n”. Overall, “best of n” may be appreciably quicker.

6.10 When should we stop creating matches?

Bouffard et al (2001, p4) commented

“...if we do not stop creating new couples from the pool at a reasonable point, then some of the couples created may have inappropriately unlikely characteristics. This danger is especially great when the numbers of persons is small in the pools from which we must draw partners”.

Easter & Vink (2000) showed analytically that two versions of stochastic matching, when applied to preselected equal batches of males and females, produced marriage probabilities that are proportional to the compatibility measures. At least for stochastic matching, there should be no need to stop until all the persons are paired. “Best of n” matching is not intended for batch use.

6.11 What do we do if there are not enough potential matches?

As ordinary people do in this situation, we may have to widen our area of search. For example, a person resident in a rural district may have to look for a partner in nearby

cities, as well as in the local region. If we drop the person from the search, then we may be excluding that person, and other similar persons, from our partnering process.

A higher minimum pool is likely to be needed when using stochastic matching for individuals. This is because stochastic selection asymptotes to a stable mean score, while “best of n” gives increasingly high mean values as n increases. With stochastic selection, we can safely use more than the minimum.

Note that selecting the best of a pool of candidates, where the pool is smaller than the optimal selection number, is likely to give score means below the population mean, and score standard deviations above the population.

6.12 Data synthesis for small regions

Stochastic matching appears suitable for data synthesis for small regions, where marginal totals but not unit records are available. A sequential allocation process was described by Cumpston (2007), but this is likely to work better with logistic probabilities and stochastic matching.

6.13 Further work

More detailed information might allow better models to be fitted, giving higher ratios of score standard deviations for unmatched and matched members. For example, use of information on the earnings of partnered persons might give a better partnership model, and require more selections in a “best of n” matching process. A better model for matching households to dwellings is clearly needed.

All three matching processes tested here apply the same regression relationship to all the possible pairings. Better results may be obtained, however, if the regression parameters are adjusted to reflect the particular individuals being matched.

All six applications considered here have involved matching of persons already known to be in particular relationships. For example, in a regional data synthesis we may know the numbers and ages of partnered males and females in the region, and we want to pair them off in plausible partnerships. But in simulating future partnerships we want to both simulate persons changing to partner status, and find partners for them.

7. LARGER DATA SETS

All of the simulation, alignment and matching techniques described here use random sampling, rather than sorting or permutations. Run times should thus be broadly in proportion to person times, so long as memory limits are not reached.

For the examples in this paper, all households were located in one of 57 statistical divisions, within 8 states. With larger data sets, it may be desirable to add a third geographic level.

8. ACKNOWLEDGMENTS

Papers published by the DYNACAN, CORSIM, Congressional Budget Office, NATSEM and SVERIGE modelling teams have been valuable. I am very grateful for the advice I have received from Cathal O'Donoghue, Richard Easter and Rick Morrison. The first general conference of the International Microsimulation Association was stimulating, and a source of useful contacts. David Service, as chair of my PhD supervisory panel, has been very helpful throughout.

REFERENCES

- Australian Bureau of Statistics (2003) "Population projections Australia 2002-2101", catalog no 3222.0, Canberra, September 3, iii + 186 pages (www.abs.gov.au)
- Australian Bureau of Statistics (2004) "Household and family projections Australia 2001 to 2026", catalog no 3236.0, Canberra, June 18, iii + 138 pages (www.abs.gov.au)
- Baekgaard H (2002) "Micro-macro linkage and the alignment of transition processes – some issues, techniques and examples," National Centre for Social and Economic Modelling, University of Canberra, Technical Paper No 25, June, v + 29 pages
- Bouffard N, Easter R, Johnson T, Morrison RJ & Vink J (2001) "Matchmaker, matchmaker, make me a match", Joint CORSIM and DYNACAN Project Document, September, 19 pages (also in Brazilian Electronic Journal of Economics, vol 4, no 2)
- Cumpston JR (2007) "Dynamic microsimulations of Australian individuals and households at varying geographic scales", paper presented to the first general conference of the International Microsimulation Association, Vienna, August, 7 pages
- Easter R & Vink J (2000) "A stochastic marriage market for CORSIM", Strategic Forecasting Technical Report, October, 8 pages
- Harding A (2007) "Challenges and opportunities of dynamic microsimulation modelling", plenary paper to the first general conference of the International Microsimulation Association, Vienna, August 21, 23 pages
- Johnson T (2001) "Nonlinear alignment by sorting", CORSIM Working Paper, February (cited by Morrison 2006)
- Morrison R (2006) "Make it so: event alignment in dynamic microsimulation", DYNACAN Team, 7th floor, Naron Building, 360 Laurier Avenue West, Ottawa, Ontario, May, 21 pages
- Morrison R (2008) "Validation of longitudinal microsimulation models", National Centre for Social and Economic Modelling, University of Canberra, Working Paper No 8, March, 41 pages
- Orcutt GH, Greenberger M, Korbel J & Rivlin A (1961) "Microanalysis of socioeconomic systems - a simulation study", Harper & Brothers, New York, vxiii + 425

Perese K (2002) "Mate matching for microsimulation models", Technical Paper 2002-3, Long-term Modelling Group, Congressional Budget Office, Washington DC, 32 pages

Watson L (2008) "HILDA user manual – release 6", University of Melbourne, March 13, 165 pages