SBAM: An Algorithm for Pair Matching*

Peter Stephensen & Tobias Markeprand, DREAM†

October 31, 2013

Abstract

This paper introduces a new algorithm for pair matching. The method is called SBAM (Sparse Biproportionate Adjustment Matching) and can be characterized as either *cross-entropy minimizing* or *matrix balancing*. This implies that we use information efficiently according to the historic observations on pair matching. The advantage of the method is its efficient use of information and its reduced computational requirements. We compare the resulting matching pattern with the harmonic and ChooSiow matching functions and find that in important cases the SBAM and ChooSiow method change the couples pattern in the same way. We also compare the computational requirements of the SBAM with alternative methods used in microsimulation models. The method is demonstrated in the context of a new Danish microsimulation model that has been used for forecasting the housing demand.

*Financial support from the Knowledge Centre for Housing Economics, Realdania is gratefully acknowledged.

†The DREAM model. Amaliegade 44, 1256 Copenhagen K. www.dreammodel.dk

1 Introduction

Dynamic microsimulation finds increasing use in demographic and socioeconomic forecasting. A big advantage of the microsimulation approach is that it makes it possible to analyze family structure. In traditional population projections the goal is usually to forecast the population by age, gender and a few other characteristics (such as origin and/or geographical region) while the family structure is ignored. Thus children and their parents are unrelated in the model, as is the typical prerequest for parenthood: matching of couples. Introducing family structure into this approach is problematic, mainly because the size of the model increases tremendously when forecasting the population by family/household characteristics. These characteristics can alternatively be analyzed in a microsimulation model without loosing control over the size of the model.

Modelling the family structure requires some extra features compared to the traditional approach. To get the family composition right, at least two things are necessesary: Parity[1] must included in the fertility determination, and pair matching must be modelled. This paper deals with the latter subject and introduces a new algorithm for pair matching.

A fundamental issue in the pair matching methodology and the event of coupeling compared with most traditional demographic events modelled in microsimulations is the coordination nature. While mortality, emigration, fertility etc. are person or household specific events, coupeling needs the coincidence of events between two independent households/persons.

Usually two distinct methodologies are mentioned when it comes to matching: the stable marriage approach (previously used in the CORSIM and DYNACANE models) and the stochastic approach (used in the DYNASIM model). The method decribed in this paper cannot be categorized as either, which use a method known from other economic and/or statistical problems, known as the *balancing matrix method*.

In the stable marriage approach the matching is determined by a behavior derived from

[1] The number of children a woman already has.

(rational) preferences. A matching of a pool of individuals is stable in this context if you cannot form a new pair which improves both individuals well-being. Theoretically, it was shown by Gale & Shapley (1962) that such an stable matching always exists and they furthermore provides an algoritme to find such a stable matching. The advantage is obviously that if the preferences are known you obtain more correct observable matchings which are more stable. Further, changes in matching patterns are explainable by well-known concepts as demand and supply. The method has at least two drawbacks in the application to simulation: first of all, the correctness is conditional upon the estimation of preferences and that the actual matching occurs as if the algorithme determines the matching. Second, the matching algorithme is very computational demanding since one needs to match directly each pair, one by one. Whereas in genuine stable matching algorithms where each person state his/her ordering of mate, the matching in more applied approaches is based upon a "compatibility index", reflecting the likelihood that they are matched. The mechanism then is as follows: pair any potential couple and estimate the compatibility index and match the pair with the highest index value. Exclude these persons from the pool and match the remaining pairs with higest index value. Repeat until the pool of is empty.

In the stochastic approach one estimates the probability of a match by the difference in characteristics, like e.g. age and education. The matching procedure then proceed by pairing using the Monte Carlo method. The method is generally very coarse in its predictions of matching pairs and equally computational demanding a like the stable approach.

The method that we introduce, called SBAM (Sparse Biproportionate Adjustment Matching), can be characterized as either *cross-entropy minimizing* or *matrix balancing* (defined below). The SBAM method is based on historical observations of pair mathings from one or more years, distributed on a set of types (age, gender, education, geographical region ect.). In a forecasted year, it is assumed that a *matching pool* of individuals has been formed. If the individuals in this pool are distributed on types as in the historical

data, the matching problem is easy to solve: We simply distribute the pairs as in the historical data. If this is not the case (which it typically is not), the pairs must be distributed in a new way. But how should the distribution be adjusted? and what principles should apply to these adjustments? One principle could be to distribute the pairs such that the distribution deviates as little as possible from the historical distribution. This can be interpreted as a so-called matrix balancing problem (Schneider & Zenios, 1990): Change the original data (defined as a matrix) such that the row and column sums are given by predefined values. A number of solutions exist to this kind of problem. One such solution is called biproportionate adjustment (or RAS adjustment). This method has at least two advantageous properties: It is relatively easy to implement, and it has a nice interpretation. Further, the matrix entries preserve the signs after the adjustment. Using biproportionate adjustment, the outcome can be interpreted as the result of a so-called cross-entropy minimization problem (McDougall, 1999). In other words, the adjusted matching changes the distribution of pairs relative to the original distribution, so that the information loss is as small as possible. The information loss is defined by Shannon's Information Theory (Shannon, 1948).

Section 2 decribes the methodology of the matching method. Section 3 analyze the matching function induced by the RAS method by comparing the properties to two other matching functions: the harmonic and Choo-Siow matching functions. The comparison is carried out by considering the changes in the distributions when the population characteristics change. Section 4 compare the calculation complexity of the SBAM method to other pratical matching processes used in dynamic microsimulation models. Section 5 shows an application of SBAM in a new Danish microsimulation model in which we compare historical observed matching distributions with the SBAM model's predictions. We also explain why the methodology of the Choo-Siow matching function is not suitable for the purpose of our model. Section 6 conclude on the main findings of this paper.

2 Methodology

There are assumed to be N individuals to be matched into pairs[2]. The individuals are divided into T types:

$$N = \sum_{j=1}^{T} N_j$$

A type could for example be defined on the basis of gender, age, origin and education. The number T can therefore be expected to be rather large[3].

The aim is to find real numbers $x_{i,j}$ $(i = 1, ..., T,\ j = 1, ..., T)$ such that

$$\sum_{j=1}^{T} x_{ij} = N_i,\ i = 1, ..., T \tag{1}$$

and

$$x_{ij} = x_{ji},\ i, j = 1, ..., T \tag{2}$$

The matching is defined by (1). The variable x_{ij} indicates the number of individuals of type i that are paired with an individual of type j. If an individual of type i is paired with a person of type j, then the opposite is also the case: An individual of type j is paired with a person of type i. This gives rise to the symmetry assumption (2).

2.1 Data

The algorithm is based on data from actual matchings. Let x^0_{ij} be the number of individuals of type i that according to data is matched with an individual of type j. As mentioned above, the data set x^0_{ij} is symmetric. This is ensured in the following way: When a pair

[2] N is assumed to be an even number.

[3] As an example, assume the types are defined on the basis of 2 genders, 50 ages (15-65), 5 education levels and 11 geographical regions. Then $T = 2 * 50 * 5 * 11 = 5.500$

of type (i, j) is added to data, it is done by following the procedure:

$$\begin{aligned} x_{ij}^0 &=: x_{ij}^0 + 1 \\ x_{ji}^0 &=: x_{ji}^0 + 1 \end{aligned}$$

where =: is an algorithmic equal sign[4]. In the data set, individuals are distributed on T types:

$$N_t^0 = \sum_{i=1}^{T} x_{it}^0 = \sum_{j=1}^{T} x_{tj}^0 \tag{3}$$

and the total number of individuals is given by

$$N_0 = \sum_{j=1}^{T} N_j^0$$

It is advantageous to describe the problem in matrix notation. The data set x_{ij}^0 can be described as a $T \times T$ matrix, X^0. Define the vector

$$\vec{N}^0 = \left(N_1^0, ..., N_T^0\right)$$

According to (3), both the row and column sums of X^0 should be given by $\vec{N}^0$.

2.2 Biproportionate Adjustment

We are going to match N individuals, distributed on types according to $\vec{N} = (N_1, ..., N_T)$. We wish to find a $T \times T$ dimensional symmetric matrix X such that its row and column sums add to $\vec{N}$. This should be done so that X deviates as little as possible from the original data X^0. In other words, we would like our matching X to reflect as much as possible of the matching information in the original (real world) matching X^0. This can be interpreted as a classical matrix balancing problem: *Given a rectangular matrix A,*

[4] $x =: x + a$ means that x is increased with the value a.

determine a matrix X that is close to A and satisfies a given set of linear restrictions on its entities (Schneider & Zenios, 1990).

Algorithms for matrix balancing can be separated into two broad classes: scaling algorithms and optimization algorithms. Scaling algorithms multiply the rows and columns of the original matrix by positive constants until the matrix is balanced. Optimization algorithms minimize a penalty function that measures the deviation of a candidate balanced matrix from the original matrix. The balance conditions are constraints in the optimization model, so that the optimal solution is the balanced matrix closest to the original matrix.

We are going to use the scaling approach here. According to the biproportionate adjustment model (also called RAS adjustment), the balancing problem can be solved in the following iterative way: Start with the original matrix. Scale the rows such that the row sums are correct. Then scale the columns such that the column sums are correct. Repeat these two operations until a new stable matrix has emerged.

When using the optimization algorithms, it is obvious that the new matrix deviates as little as possible from the original matrix (that is part of the definition of the problem). This is less obvious when it comes to the scaling algorithms. However, it has been demonstrated that the biproportionate model is an entropy-theoretic model (see e.g. McDougall, 1999 or Bregman, 1967). The new matrix can be characterized as the solution to a cross-entropy minimization model. Entropy should here be understood in an information theoretical context (Shannon, 1948). By using the biproportionate model we are actually minimizing the loss of information when changing from the type distribution $\vec{N}^0$ to $\vec{N}$.

The balancing condition in this particular problem is

$$\sum_{i=1}^{T} x_{ij} = N_j$$

$$\sum_{j=1}^{T} x_{ij} = N_i$$

for every $i, j = 1, \ldots, T$. Note that this impose a symmetry in the balancing conditions whenever the row and column indice is the same.

The procedure for the RAS algorithm is as follows: for any k let

1. for $i = 1, \ldots, m$ let $\rho_i^k = \frac{N_i}{\sum_j x_{ij}^k}$ and let $y_{ij}^k = \rho_i^k x_{ij}^k$ for all $i = 1, \ldots, m; j = 1, \ldots, n$
2. for $j = 1, \ldots, n$ let $\sigma_j^k = \frac{N_j}{\sum_i y_{ij}^k}$ and let $x_{ij}^{k+1} = \sigma_j^k y_{ij}^k$ for all $i = 1, \ldots, m; j = 1, \ldots, n$

Letting $X^0 = X_0$ the RAS-solution is the limit $X^{RAS} = lim_{k\to\infty} X^k$. It is easy to see that we can write the solution in the terms of a matrix product: letting R be the $m \times m$ diagonal matrix with elements

$$r_{ii} = \prod_{k=1}^{\infty} \rho_i^k$$

and S the $n \times n$ diagonal matrix with elements

$$s_{jj} = \prod_{k=1}^{\infty} \sigma_j^k$$

Then we can write the solution to the RAS algorithm as $X^{RAS} = RX_0S$.

An important property of the RAS algorithm, in our context, is that it preserves symmetry of the initial distribution when the balancing conditions are symmetric. Symmetry implies that for each k we have that $x_{ij}^{k+1} = x_{ji}^{k+1}$ whenever $x_{ij}^k = x_{ji}^k$. But we have that

$$\begin{aligned} x_{ij}^{k+1} &= \sigma_j^k y_{ij}^k = \frac{N_j}{\sum_i y_{ij}^k} y_{ij}^k = \frac{N_j}{\sum_i \rho_i^k x_{ij}^k} \rho_i^k x_{ij}^k = \frac{N_j}{\sum_i \frac{N_i}{\sum_j x_{ij}^k} x_{ij}^k} \frac{N_i}{\sum_j x_{ij}^k} x_{ij}^k \\ &= \frac{N_i}{\sum_j \frac{N_j}{\sum_j x_{ij}^k} x_{ji}^k} \frac{N_j}{\sum_i x_{ij}^k} x_{ji}^k = x_{ji}^{k+1} \end{aligned}$$

which proves the symmtry-preservation property of the RAS solution. The symmetry of

the solution considerably reduce the computational requirements by reducing the computations by 50%. However, this may still be a considerable computational accomplishment depending on the number of types and the convergence of the iterative procedure.

As we have mentioned, the solution of the RAS method is equivalent to the solution of a well-defined minimization problem. It has been shown that the RAS solution is the solution to the problem of minimizing of the cross entropy

$$\min_{(\xi_{ij})_{i,j}} \sum_{i,j} \xi_{ij} \log \frac{\xi_{ij}}{\xi_{ij}^0}$$

subject to the balancing condition where $\xi_{ij} = \frac{x_{ij}}{N_i}$ and $\xi_{ij}^0 = \frac{x_{ij}^0}{N_i^0}$.

The RAS algorithm has a well-known graph structure associated with it, which is valuable for understanding the mathematical structure. A *graph* is a pair (V, E) where V is the set of *vertices* and E the set of *arces* a subset of $V \times V$. Given two vertices i and j we say that they are connected in the graph (V, E) if $(i, j) \in E$ or $(j, i) \in E$, that is if there exists an arc connecting the two vertices. A graph is *biparitite* if there exists adjoint subsets $V_1, V_2 \subset V$ that covers Vsuch that for every $i \in V_1$ there exists an $j \in V_2$ and $(i, j) \in E$. We can now define the *transportation graph* for the RAS algorithm: Given X^0 we denote by V_1 the set of rows and V_2 the set of columns and we let the transportation graph be the biparitite graph (V, E) given by

$$V = \{1, \ldots, m\} \cup \{1', \ldots, n'\} = V_1 \cup V_2$$

and

$$E = \{(i, j') \in V_1 \times V_2 | x_{ij'}^0 > 0\}$$

where $V_2 = \{1', \ldots, n'\}$ is just a relabelling of $\{1, \ldots, n\}$ to distinguish elements from the two sets V_1 and V_2. Hencefourth, we will refer to arces as (i, j) instead of (i, j') when no confusion is at risk. The biparitite property of the transportation graph stems from the

absence of a zero row (column). The matrix X^0 can then be identified by the positive map $x^0 : E \longrightarrow \mathbb{R}_+$. The RAS algorithm then cycles through the transportation graph scaling the map X^0 on the transportation graph such that the limit satisfies

$$\sum_{\{j|(i,j)\in E\}} x_{ij} = N_i$$

for every $i \in V_1$ and

$$\sum_{\{i|(i,j)\in E\}} x_{ij} = N_j$$

for every $j \in V_2$. It is important to note that there are computational differences between the matrix algorithm and the graph algorithm, since the former requires computations for all T^2 entries, while the latter only refer to the subset of these elements in which there are a strictly positive entry.

2.3 Sparse algorithm

As mentioned above, the number of types T can and will in applications often be very large. Therefore, a $T \times T$ matrix can easily become so large that it gives rise to computational problems. As there at the same time often will be many zeros in the X^0 matrix, it will have obvious advantages to introduce a sparse matrix method in which operations on zero-elements are ignored. The method is implemented in C# and is based on so-called *linked lists*[5]. A $T \times T$ matrix can be represented by a *SBAMMatrix*. A SBAMMatrix is a C# object that essentially contains $2T$ linked lists: T linked lists for the rows and T linked lists for the columns. Each element in the linked list contains a pointer to data and a reference to the next element in the list. In this way, data is actually represented twice: as rows and as columns. The motivation for this redundancy is that it makes biproportionate

[5] A linked list is a data structure consisting of a group of nodes that together represent a sequence. Each node is composed of data and a reference (a link) to the next node in the sequence. This structure is memory space saving and allows for efficient insertion or removal of elements from any position in the sequence.

scaling much easier. The SBAMMatrix actually implements the graph structure of the RAS algorithm, as described above.

3 Matching functions

A well-known characterization of reduced form matching models is through the use of matching functions. In terms of the model in this paper, a matching function is a map $\mu_{ij}(N) \in \mathbb{R}$ which gives the number of matchings between type i and type j individuals given a population N. This section provides a comparison of the resulting steady state matching pattern of the SBAM method with two other important methods: the harmonic mean method and the Choo-Siow method.

Traditionally, the definition has been termed a bit differently which is contained in our formulation. To arrive at the traditional formulation one can separat the set $N = M \cup F$ where M is the set of males to be matched and F is the set of females to be matched. Then $\mu_{ij}(M, F)$ is the number of males of type i to be married to females of type $j$dependent on the number of all males and females to be matched. In the present model we allow persons of the same gender to form relationships, which the traditional formulation does not allow.

The matching function usually satisfies some properties (beyond the balancing conditions):

- Zero spillover: letting N_i be the number of type i, then $\mu_{ij}(N) = \mu_{ij}(N_i, N_j)$
- Homogenity of one: for any scalar $\lambda > 0$ we have that $\mu_{ij}(\lambda N) = \lambda \mu_{ij}(N)$

The zero spillover property implies that the matching of pairs of given types only depend on the number of each type. This implies that no *substitutioneffect* is possible: "if I cannot be with the one I prefer most, I do not want to be with anyone at all".

An example of this is Harmonic Mean matching function (see Schoen (1988))

$$\mu_{ij}^{HM}(M,F) = \alpha_{ij}\frac{m_i f_j}{m_i + f_j} = \mu_{ij}^{HM}(m_i, f_j)$$

where $\alpha_{ij} > 0$, $\sum_i \alpha_{ij} \leq 1$ and $\sum_j \alpha_{ij} \leq 1$. Note that the sum of each type

$$\sum_j \mu_{ij}(M,F) = m_i \sum_j \frac{\alpha_{ij} f_j}{m_i + f_j} \leq m_i \sum_j \frac{f_j}{m_i + f_j} \leq m_i$$

such that the total number of matches assigned to type i-males is less than the total number of type i-males. And likewise for j-type females. Thus, the proportions of married men of type i is then given by

$$\rho_i(M,F) = \frac{\sum_j \mu_{ij}(M,F)}{m_i} = \sum_j \alpha_{ij}\frac{f_j}{m_i + f_j}$$

while the proportion of married females $\eta_j(M,F)$ is defined likewise. The remainning males/females is not matched, and is given by

$$\mu_{i0} = (1 - \rho_i(M,F))\, m_i = \left(\sum_j \frac{m_i + (1-\alpha_{ij}) f_j}{m_i + f_j}\right) m_i$$

and

$$\mu_{0j} = (1 - \eta_j(M,F))\, f_j = \left(\sum_i \frac{f_j + (1-\alpha_{ij}) m_i}{f_j + m_i}\right) f_j$$

which we note are all strictly positive and less than one. We note that these rates are all dependent on, and thus changes whenever, the number of male/females changes. The change in the individual matching parameters are

$$\begin{bmatrix} \frac{\partial \mu_{ij}}{\partial m_i} & \frac{\partial \mu_{ij}}{\partial f_j} \\ \frac{\partial \rho_i}{\partial m_i} & \frac{\partial \rho_i}{\partial f_j} \\ \frac{\partial \eta_i}{\partial m_i} & \frac{\partial \eta_i}{\partial f_j} \end{bmatrix} = \begin{bmatrix} \alpha_{ij}\left(\frac{f_j}{m_i+f_j}\right)^2 & \alpha_{ij}\left(\frac{m_i}{m_i+f_j}\right)^2 \\ -\sum_j \alpha_{ij}\frac{f_j}{(m_i+f_j)^2} & \alpha_{ij}\frac{m_i}{(m_i+f_j)^2} \\ \alpha_{ij}\frac{f_j}{(m_i+f_j)^2} & -\sum_i \alpha_{ij}\frac{m_i}{(m_i+f_j)^2} \end{bmatrix}$$

where we see that there is a crowding out effect: as there becomes more of one type of males, this implies that the proportion of married males of this type decreases. An important proporty of the harmonic mean matching function is that it is homogenous of degree one, that is the ratios of marriage propensities is affected by a scaling of the number of men and females:

$$\mu_{ij}(\lambda M, \lambda F) = \mu_{ij}(\lambda m_i, \lambda f_j) = \lambda \mu_{ij}(m_i, f_j)$$

for any $\lambda > 0$. The interpretation of this proporty is that the searching/matching process is made easier by the presence of a larger set of supply of possible matches. Note, however, that the proportion of married men/female of a given type is not affected by a equi-proportional increase in the men and female of the type, i.e.

$$\rho_i(\lambda M, \lambda F) = \rho_i(M, F)$$

A second example is the Pollard/Hohn matching function given by

$$\mu_{ij} = \frac{a_i m_i b_j f_j}{\frac{1}{2}\sum_k (h_{kj} a_k m_k + h_{ik} b_j f_j)}$$

where a_i, b_j and h_{ij} are parameters/weights to be estimated, m_i is the number of men of type i and f_j is the number of females of type j.

A third example is the Choo/Siow matching function given by

$$\mu_{ij} = \pi_{ij}\sqrt{\mu_{i0}\mu_{0j}}$$

where π_{ij} is a parameter to be estimated which represents the sum of gross benefits of type i and j of a matching of the two types in excess of being single, μ_{i0} is the number of unmarried persons (males) of type i and μ_{j0} is the number of unmarried persons (females)

of type j, both derived residually. Choo & Siow (1993) derives the matching function from a transferable utility model of matching model. Since, by definition, the number of married men and the number of unmarried men exhaust the males of a given type i

$$\mu_{i0} + \sum_{j=1}^{J} \mu_{ij} = m_i$$

and likewise for females

$$\mu_{0j} + \sum_{i=1}^{I} \mu_{ij} = f_j$$

we have that a solution for $\left((\mu_{i0})_i , (\mu_{0j})_j \right)$ must satisfy

$$m_i - \mu_{i0} \quad = \quad \sum_{j=1}^{J} \pi_{ij} \sqrt{\mu_{i0}\mu_{0j}}$$

and

$$f_j - \mu_{0j} = \sum_{i=1}^{I} \pi_{ij} \sqrt{\mu_{i0}\mu_{0j}}$$

which is basically a quadratic form.[6] Changes in the parameters $((m_i), (f_j), \prod)$ can determine the changes in the unemployment rates $(\mu_{i0}, \mu_{j0})_{ij}$ which can then be used for calculating the change in the distribution of marriages.

It is easy to see that our derived matching function violates the zero-spillover property but satisfies the homogenity of one property. All the above examples also satisfies the homogenity property, which is a property that implies that the number of pairs matched does change with a general increase in the population, and thus that the population density does not affect the ability of matching. The SBAM contains a very general spillover effect, in that, not only does the matching of say a type i to a type j depend on any other match of type i with any other type j', but also on the matching of type $i' \neq i$ with any

[6] Make the transformation $\tilde{\mu}_{ij} = \mu_{ij}^2$ then the equation for males of type i becomes $m_i - \tilde{\mu}_{i0}^2 = \sum_j \pi_{ij} \tilde{\mu}_{i0} \tilde{\mu}_{0j}$ which has as a highest number of exponent two.

other type j'. A second important aspect in which our matching function differ from the mentioned, is that the proportion of matched pairs ρ and η are independent of the number of males/females of the different types.

3.1 Comparing matching functions

To see the potential effect of considering the joint distribution when marginal distributions pertubates, consider the initial distribution

$$\begin{bmatrix} 0 & 0 & 0 & 10 & 3 & 1 \\ 0 & 0 & 0 & 5 & 10 & 5 \\ 0 & 0 & 0 & 1 & 3 & 10 \\ 10 & 5 & 1 & 0 & 0 & 0 \\ 3 & 10 & 3 & 0 & 0 & 0 \\ 1 & 5 & 10 & 0 & 0 & 0 \end{bmatrix}$$

where the first three rows (columns) are men (with 3 types) and the last three rows (columns) are female (with 3 types). Thus the initial row/column sums are $N^0 = (14, 20, 14, 16, 16, 16)$. The population vector is initially given by $L^0 = (20, 25, 17, 20, 20, 20)$ such that the proportion of married of each type is $(\rho, \eta) = (0.7, 0.8, 0.82, 0.8, 0.8, 0.8)$. Assume that 5 males of the first type is added to the total population, such that $L = (25, 25, 17, 20, 20, 20)$, how does this change the matching distribution? The unmarried proportions are changed to

$$\left(\rho^{HM}, \eta^{HM}\right) = \left((0.62, 0.8, 0.82), (0.86, 0.82, 0.81)\right)$$

for the harmonic mean matching function, and

$$\left(\rho^{CS}, \eta^{CS}\right) = \left((0.63, 0.78, 0.81), (0.83, 0.81, 0.81)\right)$$

which implies that the number of couples increase by 1.3 in the HM-case and 1.6 in the CS-case. The resulting pair-matching of the two matching functions are

$$\mu^{HM} = \begin{bmatrix} 0 & 0 & 0 & 11.1 & 3.3 & 1.1 \\ 0 & 0 & 0 & 5 & 10 & 5 \\ 0 & 0 & 0 & 1 & 3 & 10 \\ 11.1 & 5 & 1 & 0 & 0 & 0 \\ 3.3 & 10 & 3 & 0 & 0 & 0 \\ 1.1 & 5 & 10 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 0 & 0 & 0 & 11.1 & 3.5 & 1.2 \\ 0 & 0 & 0 & 4.7 & 9.9 & 5 \\ 0 & 0 & 0 & 0.9 & 2.9 & 10 \\ 11.1 & 4.7 & 0.9 & 0 & 0 & 0 \\ 3.5 & 9.9 & 2.9 & 0 & 0 & 0 \\ 1.2 & 5 & 10 & 0 & 0 & 0 \end{bmatrix}$$

from which we see that the main difference between the two types of matching functions is the cross-type effect which is zero in the HM-case and non-zero in the CS-case. The interpretation is that a greater number of males increases the competition on the male-side, which increases the demand for females and the supply of males. Like any traditional economic mechanism would predict this would lower the relative price of males and symmetrically increase the relative price of females. This would force males to increase their utility-transfer to females and thus more females would be attracted to move from being single to engage in marriage, while spur relatively more males to be singles.

Consider now the resulting change in a SBAM procedure. Unfortunately, the SBAM and the CooSiow (and also the Harmonic Mean) matching procedures are in there most basic form in some sense incomparable. To see why, consider an increase in the male population of, say, type 1. For a given number of females, of each female-type, the number of couples could not increase if the number of females are not increased as well. In the CS- matching this is accomblished by an increase in the benefit of each female and thus new females are attracted to the market for matches. However, in the SBAM the number of females is unchanged, since the proportion of unmarried is unchanged. However, the procedure requires that any married individual as predicted by the fixed marriage proportion assumption is matched to a partner. The solution in the SBAM case is to allow for

same-sex matchings. Recall, however, that matchings do not necessarily imply a sexual relationship, but instead implies a common household. So in order to make the two (three) models comparable we now extend the harmonic mean and Choo-Siow models to allow for the possibility of same-sex matchings. Further, to be able to compare the HM- and CS-functions with the SBAM we also need to assume that in each period all couples are dissolved and the entire population must be rematched. The need for this assumption is discussed in section 5 and has to do with stock and flow models.

The extension is straight forward and present no mentionable technical difficulties. The observed matchings are now altered to be

$$\begin{bmatrix} 1 & 0 & 0 & 10 & 2 & 1 \\ 0 & 1 & 0 & 4 & 10 & 5 \\ 0 & 0 & 1 & 1 & 2 & 10 \\ 10 & 4 & 1 & 0 & 0 & 0 \\ 2 & 10 & 2 & 0 & 0 & 0 \\ 1 & 5 & 10 & 0 & 0 & 0 \end{bmatrix}$$

such that the marginal propensities of 'marriage' are unchanged, given by

$$(\rho, \eta) = ((0.70, 0.80, 0.82), (0.75, 0.70, 0.80))$$

. The effect of an increase in the number of type 1-males by 5 individuals on the propensity to marriage is then

$$\begin{aligned} \left(\rho^{HM}, \eta^{HM}\right) & \quad ((0.63, 0.80, 0.82), (0.81, 0.71, 0.81)) \\ \left(\rho^{CS}, \eta^{CS}\right) & = ((0.65, 0.79, 0.82), (0.79, 0.72, 0.81)) \\ \left(\rho^{SBAM}, \eta^{SBAM}\right) & \quad ((0.70, 0.80, 0.82), (0.75, 0.70, 0.80)) \end{aligned}$$

such that the total number of matchings are changed by 1.7, 1.8 and 3.5, respectly, and

the distributions are

$$\mu^{HM} = \begin{bmatrix} 1.3 & 0 & 0 & 11.1 & 2.2 & 1.1 \\ 0 & 1.0 & 0 & 4 & 10 & 5 \\ 0 & 0 & 1.0 & 1 & 2 & 10 \\ 11.1 & 4 & 1 & 0 & 0 & 0 \\ 2.2 & 10 & 2 & 0 & 0 & 0 \\ 1.1 & 5 & 10 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 1.5 & 0 & 0 & 11.1 & 2.4 & 1.2 \\ 0 & 1.0 & 0 & 3.7 & 10.0 & 5 \\ 0 & 0 & 1.0 & 0.9 & 2.0 & 10 \\ 11.1 & 3.7 & 0.9 & 0 & 0 & 0 \\ 2.4 & 10.0 & 2.0 & 0 & 0 & 0 \\ 1.2 & 5 & 10 & 0 & 0 & 0 \end{bmatrix},$$

$$\mu^{SBAM} = \begin{bmatrix} 3.0 & 0 & 0 & 10.8 & 2.5 & 1.3 \\ 0 & 1.8 & 0 & 3.4 & 9.7 & 5.1 \\ 0 & 0 & 1.7 & 0.8 & 1.8 & 9.7 \\ 10.8 & 4.7 & 0.9 & 0 & 0 & 0 \\ 2.5 & 9.7 & 1.8 & 0 & 0 & 0 \\ 1.3 & 5.1 & 9.7 & 0 & 0 & 0 \end{bmatrix}$$

To give a more proportional feeling of the magitudes of the change, we can compute the percentage change in the distributions of matchings

$$\mu^{HM} = \begin{bmatrix} 0.25 & 0 & 0 & 0.11 & 0.11 & 0.11 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 0.47 & 0 & 0 & 0.11 & 0.18 & 0.19 \\ 0 & 0.05 & 0 & -0.06 & 0.00 & 0.00 \\ 0 & 0 & 0.03 & -0.07 & -0.01 & 0.00 \\ 0.11 & -0.06 & -0.07 & 0 & 0 & 0 \\ 0.18 & 0.00 & -0.01 & 0 & 0 & 0 \\ 0.19 & 0.00 & 0.00 & 0 & 0 & 0 \end{bmatrix}$$

$$\mu^{SBAM} = \begin{bmatrix} 1.97 & 0 & 0 & 0.08 & 0.23 & 0.28 \\ 0 & 0.85 & 0 & -0.15 & -0.03 & 0.01 \\ 0 & 0 & 0.68 & -0.19 & -0.08 & -0.03 \\ 0.08 & -0.15 & -0.19 & 0 & 0 & 0 \\ 0.23 & -0.03 & -0.08 & 0 & 0 & 0 \\ 0.28 & 0.01 & -0.03 & 0 & 0 & 0 \end{bmatrix}$$

We note that in general the SBAM result in much greater sensitivities towards the number and distribution of partnerships than the other matching functions. This greater sensitivity mainly comes from the constantness of the proportions of marriage, which translate into a considerable increase in the same-sex 'marriages'.

Consider next a simultanous increase of five persons of type 1- males and -females. The percentage change in the distributions are

$$\mu^{HM} = \begin{bmatrix} 0.25 & 0 & 0 & 0.25 & 0.11 & 0.11 \\ 0 & 0 & 0 & 0.13 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 & 0 \\ 0.25 & 0.13 & 0.10 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 0.26 & 0 & 0 & 0.28 & 0.12 & 0.12 \\ 0 & -0.04 & 0 & 0.12 & -0.02 & -0.02 \\ 0 & 0 & -0.01 & 0.14 & -0.01 & -0.01 \\ 0.28 & 0.12 & 0.14 & 0 & 0 & 0 \\ 0.12 & -0.02 & -0.01 & 0 & 0 & 0 \\ 0.12 & -0.02 & -0.01 & 0 & 0 & 0 \end{bmatrix},$$

$$\mu^{SBAM} = \begin{bmatrix} 0.47 & 0 & 0 & 0.30 & 0.05 & 0.12 \\ 0 & 0.15 & 0 & 0.15 & -0.07 & -0.01 \\ 0 & 0 & 0.15 & 0.15 & -0.07 & -0.01 \\ 0.30 & 0.15 & 0.15 & 0 & 0 & 0 \\ 0.05 & -0.07 & -0.07 & 0 & 0 & 0 \\ 0.12 & -0.01 & -0.01 & 0 & 0 & 0 \end{bmatrix}$$

and the number of couples increases by 3.7 in all models. Note that while in the CS-method

the same-sex matchings are substitutes (when the supply of type 1-population increase, their matching-price/value decrease which in the case of CS implies that the demand for same-sex matchings decrease, which is the definition of (gross) substitutes) the same-sex matchings are complements.

To consider the effect of a symmetric initial condition, in which there are female, as well as male, same-sex relationships, we consider an alternative initial observation given by

$$\begin{bmatrix} 1 & 0 & 0 & 9 & 2 & 1 \\ 0 & 1 & 0 & 4 & 9 & 5 \\ 0 & 0 & 1 & 1 & 2 & 9 \\ 9 & 4 & 1 & 1 & 0 & 0 \\ 2 & 9 & 2 & 0 & 1 & 0 \\ 1 & 5 & 9 & 0 & 0 & 1 \end{bmatrix}$$

in which case we obtain per centage change with only type 1-males increase, the marriage rate of the different changes as (HM)

$$(\rho, \eta) = ((0.65, 0.76, 0.76), (0.75, 0.7, 0.8)) \Rightarrow (\rho, \eta)^{HM} = ((0.58, 0.76, 0.76), (0.80, 0.71, 0.81))$$

and for CS

$$(\rho, \eta)^{CS} = ((0.6, 0.75, 0.76), (0.78, 0.71, 0.81))$$

thus we see that marriage rates respond more sensitive in this case compared to the case

in which only male relationships where present, and the distributions change by

$$\mu^{HM} = \begin{bmatrix} 0.25 & 0 & 0 & 0.11 & 0.11 & 0.11 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 0.43 & 0 & 0 & 0.12 & 0.17 & 0.18 \\ 0 & 0.03 & 0 & -0.05 & 0.00 & 0.00 \\ 0 & 0 & 0.02 & -0.06 & -0.01 & 0.00 \\ 0.12 & -0.05 & -0.06 & -0.13 & 0 & 0 \\ 0.17 & 0.00 & -0.01 & 0 & -0.04 & 0 \\ 0.18 & 0.00 & 0.00 & 0 & 0 & -0.03 \end{bmatrix},$$

$$\mu^{SBAM} = \begin{bmatrix} 1.42 & 0 & 0 & 0.12 & 0.24 & 0.29 \\ 0 & 0.52 & 0 & -0.11 & -0.02 & 0.02 \\ 0 & 0 & 0.41 & -0.14 & -0.05 & -0.01 \\ 0.12 & -0.11 & -0.14 & -0.48 & 0 & 0 \\ 0.24 & -0.02 & -0.05 & 0 & -0.37 & 0 \\ 0.29 & 0.02 & -0.01 & 0 & 0 & -0.31 \end{bmatrix}$$

and when we consider both male and female type-1 increase the marriage propensities are

$$(\rho, \eta)^{HM} = ((0.63, 0.78, 0.77), (0.72, 0.71, 0.81))$$

and

$$(\rho, \eta)^{CS} = ((0.65, 0.77, 0.77), (0.74, 0.7, 0.8))$$

and the distributions becomes

$$\mu^{HM} = \begin{bmatrix} 0.25 & 0 & 0 & 0.11 & 0.11 & 0.11 \\ 0 & 0 & 0 & 0.13 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 & 0 \\ 0.25 & 0.13 & 0.10 & 0.25 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mu^{CS} = \begin{bmatrix} 0.26 & 0 & 0 & 0.28 & 0.12 & 0.12 \\ 0 & -0.03 & 0 & 0.12 & -0.02 & -0.02 \\ 0 & 0 & -0.01 & 0.13 & -0.01 & -0.01 \\ 0.28 & 0.12 & 0.13 & 0.30 & 0 & 0 \\ 0.12 & -0.02 & -0.01 & 0 & -0.01 & 0 \\ 0.12 & -0.02 & -0.01 & 0 & 0 & 0.00 \end{bmatrix},$$

$$\mu^{SBAM} = \begin{bmatrix} 0.25 & 0 & 0 & 0.31 & 0.12 & 0.14 \\ 0 & -0.07 & 0 & 0.12 & -0.03 & -0.02 \\ 0 & 0 & -0.04 & 0.14 & -0.02 & -0.01 \\ 0.31 & 0.12 & 0.14 & 0.36 & 0 & 0 \\ 0.12 & -0.03 & -0.02 & 0 & 0.00 & 0 \\ 0.14 & -0.02 & -0.01 & 0 & 0 & 0.03 \end{bmatrix}$$

Again, it it extremely noteworthy that the changes in the distributions of ChooSiow and the SBAM are very similar, and in this case in which both same-sex relations are considered actually the magnitudes of the changes are almost the same. These results suggests that we can use the SBAM-approach in the microsimulation but use the interpretations from the CS-approach when interpreting the actual simulation results. A further notice is that the change in distribution with a proportional increase in the male and female of a given type is highest in the diagonal and same-sex of the increased male-female types. Compared to the asymmetric increase in which the sensitivities of the distributions of SBAM and CS are more dislike, however, they agree on the trend that the same-sex match of the increase type have uniformly the greatest sensitivity.

Like the rate of matchings is independent of the population size and distribution in our approach, the rate of separation is independent. While in the two other cases the

harmonic mean and ChooSiow the dissolvement rate changes as the size and distribution of the population changes.

4 Alternative microsimulation matching procedures

We shall here consider the matching process used in the DYNASIM and CORSIM/DYNACAN. Both methods takes the propensity to marriage as exogenous: the probability to be marriaged given age, gender, education and labour market status within a given period is independent of the population distribution.

4.1 DYNASIM-like procedures

In the DYNASIM model, the matching pool is then randomly queued of males and females in separate lines. Take the first male in the queue, call him i, and start in the female line by picking the first female, call her j, and assign the probability[7]

$$\Pr\{x_{ij} = 1\} = \exp\left\{-\frac{1}{2}\sqrt{(a_i - a_j)^2 + (e_i - e_j)^2}\right\}$$

of these two individuals to form a pair. If they are not matched the male i is trialed a match with the next female in the queue using the same procedure as above. If the male is not matched within 10 trials he is matched with the femail with the highest probability estimated in the trials. Any asymmtry in the number of male and females are handled by leaving the unmatched single and letting them indure the risk of marriage next year.

To calculate the average number of computations necessary to complete the matching proces, with only age as the determining parameter, consider the probability distribution of a 30-year old male, given that we match this male with a randomly chosen female the probability that they will be matched is the expectation of the distribution, e.g. ca. 3 per

[7]This is based upon the presentation of Perese (2002), but it seems an flawed presentation because there seems to be no parameters to guarantee that the distributions actually generate the actual data. One would perhaps estimate a difference-age and -education parameter.

cent, in the next step the conditional expectation is the same, given that the population is large, such that the probability of not having found a mate within $n \leq 10$ trials is $(1-3\%)^n$ and thus a probability that having found a mate after exactly n trials is $1-(1-3\%)^n$ and the probability that exactly at the n round is

$$\Pr\{\tau \leq n-1\} - \Pr\{\tau \leq n\}$$

which implies that there is about 70% chance that we need more than 10 trials to match the inidividual. Thus the expected number of trials are 8.5 trials per individual. Assume that there are $N = \min\{F, M\}$ individuals are to be matched, then the expected number of calculations are $8.5 * N$ calculations. The resulting distribution of matchings is a very complex and intransparent distribution which would generally depend upon the number of trials.

4.2 CORSIM-like procedures

In the CORSIM/DYNACAN model, the matching pool is as follows: let there be N persons in the matching pool, then the procedure estimates $N^2 = N * N$ probabilities using a estimate as follows: let there be N_0 individuals observed and variables $\left\{(y_{ij}, x_{ij})_{i,j=1,\ldots,N_0}\right\}$ in which y_{ij} is a dummy equal to one if individual i and j are married, and x_{ij} are variables such as the difference in age, difference in age squared, difference in years of education, labourforce status, difference in earnings etc. including interaction terms and then estimate the model

$$y_{ij} = \beta x_{ij} + \epsilon_{ij}$$

as a cross-sectional model. In the forecast at a given point in time there are, say N, individuals in the matching pool. The probabilities are then calculated

$$p_{ij} = \frac{\beta x_{ij}}{N_0}$$

$i, j = 1, \ldots, N$ and the matching matrix is the matrix of the form

$$A^{(N)} = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1N} \\ 0 & 0 & a_{23} & \cdots & a_{2N} \\ 0 & 0 & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{N-1,N} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

given recursively by $a_{12} = p_{12}$, $a_{13} = a_{12} + p_{12}$, $a_{23} = a_{13} + p_{23}$ and hence for every $1 < i < j \leq N$

$$a_{ij}^{(N)} = a_{i-1,j}^{(N)} + p_{ij}$$

and

$$a_{1j}^{(N)} = a_{j-2,j-1}^{(N)} + p_{1j}$$

The distribution is obtained as follows: start by drawing a number ψ between 0 and 1 then the pair (i_0, j_0) is formed whenever $\psi a_{NN}^{(N)} \in \left[a_{i_0,j_0-1}^{(N)}, a_{i_0,j_0}^{(N)}\right[$. Then the individuals (i_0, j_0) are removed from the mathcing pool, and $a_{i_0,j} = a_{i,j_0} = 0$, and we calculate $A^{(N-2)}$ in which

$$a_{ij}^{(N-2)} = a_{ij}^{(N)}$$

for every $i \neq i_0, j \neq j_0$, $a_{i_0 j_0}^{(N-2)} = a_{i_0-1,j_0}^{(N-2)}$ and

$$a_{ij}^{(N-2)} = a_{i-1,j}^{(N-2)} + p_{ij}$$

The procedure is repeated as long as there are any individuals left. It is easy to see that the ordering of the indices in the construction of A is irrelevant for the resulting matching since the number ψ is uniformly random.[8]Consider the case where we have 6 individuals in the

[8]A heurestic argument goes as follows: note that an ordering in the construction of A determines the ordering of the intervals $\left[a_{i_0,j_0-1}^{(N)}, a_{i_0,j_0}^{(N)}\right[$ but the length of each interval is unchanged when A is appropriately constructed. The random draw will thus have equal probability of matching the pair under

matching pool $\{1,2,3,4,5,6\}$, with an equal probability of matching then the matching matrix is

$$A^6 = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ 0 & 0 & a_{23} & a_{24} & a_{25} & a_{26} \\ 0 & 0 & 0 & a_{34} & a_{35} & a_{36} \\ 0 & 0 & 0 & 0 & a_{45} & a_{46} \\ 0 & 0 & 0 & 0 & 0 & a_{56} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{15} & \frac{2}{15} & \frac{4}{15} & \frac{7}{15} & \frac{11}{15} \\ 0 & 0 & \frac{3}{15} & \frac{5}{15} & \frac{8}{15} & \frac{12}{15} \\ 0 & 0 & 0 & \frac{6}{15} & \frac{9}{15} & \frac{13}{15} \\ 0 & 0 & 0 & 0 & \frac{10}{15} & \frac{14}{15} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Assume that a pair $(2,3)$ is matched in the first round by drawing a number $\gamma \in \left[\frac{2}{15}, \frac{3}{15}\right[$, then we alter the matching pool $\{1,4,5,6\}$ and the matching matrix becomes

$$A^4 = \begin{bmatrix} 0 & 0 & 0 & a_{14} & a_{15} & a_{16} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{45} & a_{46} \\ 0 & 0 & 0 & 0 & 0 & a_{56} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & a_{14} & a_{15} & a_{16} \\ 0 & 0 & a_{45} & a_{46} \\ 0 & 0 & 0 & a_{56} \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{6} & \frac{2}{6} & \frac{4}{6} \\ 0 & 0 & \frac{3}{6} & \frac{5}{6} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Next the pair $(1,4)$ is matched which leaves the pair $(5,6)$ as the final match.

What are the computational requirements of this procedure? There are two steps: 1) the probability estimates of each potential match between persons and 2) the matching procedure. The first step involves N^2 calculations while the second step is more difficult. For each of the $\frac{N}{2}$ iterations in the matching process we need to recalculate the matching matrix, but we do not need to recalculate the entire matrix - only the ordering greater than the pair matched in the subsequent match. There are potential $\frac{(M+1)^2}{2} - \frac{M+1}{2} = \frac{M+1}{2}M$ elements of a matching matrix $A^{(M)}$ with M persons to be matched, assuming that a

the two orderings.

person cannot be matched with himself. The expected or average number of recalculations is number of indices above the expected index to be matched: denote by π the distribution given by

$$\pi_s^{(M)} = \frac{p_{i(s),j(s)}}{a_{M-1,M}^{(M)}}$$

then the enumber of recalculations are the indices in which there are indices which are changed, denote this number by ERe. This is, in general, a complex expression so let us consider two extreme cases: *the diagonal distribution* and *the uniform distribution.*

4.2.1 The diagonal distribution

The diagonal distribution has $p_{ij} = 0$ whenever $i < j - 1$ and $a_{j,j+1} = \sum_{i \leq j} p_{i,i+1}$. This is the distribution which contains the most information and thus requires the least number of recalculations. Assume also that $p_{i,i+1} = \hat{p}$ for some positive number. Whenver $N = 4$ no need for recalculations are present. For $N = 6$ the expected number of recalculations are

$$\frac{1}{5}3 + \frac{1}{5}2 + \frac{1}{5}1 = \frac{6}{5}$$

since matching the two first individuals (with probability $\frac{1}{5}$) require 3 recalculations (the numbers a_{34}, a_{45} and a_{56}), matching the second and third (with probability $\frac{1}{5}$) require 2 recalculations (the numbers a_{45} and a_{56}) etc. When the first pair is matched and numbers are recalculated we are left with 4 individuals in which the expected number of recalculations are zero. For $N > 4$ even to be matched then

$$\pi_s^{(N)} = \pi_i^{(N)} = \frac{\hat{p}}{\hat{p}\,(N-1)} = \frac{1}{N-1}$$

and the expected index to be recalculated can be derived from the recursive formula

$$ERe_N = \sum_{s=1}^{N-1} \frac{1}{N-1}\,(N-2-s) + ERe_{N-2} = \frac{(N-2)^2}{N-1} - \frac{N}{2} + ERe_{N-2}$$

Thus, the average number of recalculations depends upon the number of individuals, N. The total number of calculations is thus the sum of random draws, two for each pair, hence $= 2 * \frac{N}{2} = N$ and the number of recalculations.

4.2.2 The uniform distribution

In the case of uniform distribution $p_{ij} = \hat{p}$ for every pair $s = (i(s), j(s))$ such that

$$\pi_s^{(M)} = \frac{1}{\frac{M-1}{2}M} = 2\frac{1}{M(M-1)}$$

Again the number of recalculations for 4 individuals to be matched is zero, since the matching of one pair automatically match the other pair. For 6 individuals the number of possible matchings are 15 and having matched a pair typically requires 6 recalculations except whenever the matching involve individual 3 and a person above index 3 (5 required recalculations), individual 4 and a person above (3 required recalculations) and only 5 and 6 with no recalculations. Continue with $N = 8$ we find that there are 28 possible pairs, denote by s_i the number of recalculations required if individual i is paired with a person with a higher number than her self; one can convince oneself that matching a pair (i, j) the required recalculations are independent of j. Thus the number of expected recalculations are

$$\frac{2(N-i)}{N(N-1)}s_i$$

and the size of s_i is given by

$$s_i - s_{i-1} = -i$$

for every $i > 1$ and $s_1 = \frac{1}{2}(N-3)(N-2)$. Thus, we have that

$$s_i = s_1 - (i-1)i + (i-2) = \frac{1}{2}(N-3)(N-2) - (i-1)i + (i-2)$$

Adding the terms for each individual i we obtain that the expected recalculations for the 8-persons matching procedure is

$$\begin{aligned}\sum_{i=1}^{N-1}\frac{1}{N-1}\frac{2(N-i)}{N(N-1)}s_i &= \frac{1}{N-1}s_1+\frac{1}{N-1}\sum_{i=2}^{N-1}\frac{2(N-i)}{N(N-1)}\left(s_1-(i-1)i+(i-2)\right)\\ &= s_1\frac{1}{N-1}\left(1+\sum_{i=2}^{N-1}\frac{2(N-i)}{N(N-1)}\right)-\frac{1}{N-1}\sum_{i=2}^{N-1}\frac{2(N-i)}{N(N-1)}\left((i-1)i-(i+2)\right)\end{aligned}$$

It is quit a remarkable expression, but an interesting feature is that it always exceeds the number

$$\frac{1}{N-1}s_1=\frac{1}{2}\frac{N-2}{N-1}(N-3)$$

which increases in N. The total expected recalculations in the matching procedure is then this number plus the expected recalculations from the $N-2$ person matching procedure, thus

$$\begin{aligned}ERe_N &= \frac{(N-3)(N-2)}{2(N-1)}+\frac{1}{N-1}\sum_{i=2}^{N-1}\frac{2(N-i)}{N(N-1)}\left(\frac{1}{2}(N-3)(N-2)-(i-1)i+(i-2)\right)\\ &\quad +ERe_{N-2}\end{aligned}$$

and thus the *added* expected required recalculations in the procedure is then increasing, and thus the recalculations are convex of nature. We elaborate more on this expression to understand the result: note that for a given person i adding a new pair of individuals to the matching pool increases the number of recalculations for whenever this person is matched by

$$\frac{ds_i^N}{dN}=\frac{1}{2}(N-2+N-3)=N-\frac{5}{2}$$

and hence there is an "externality" of adding a new person. So each term

$$\frac{1}{2}(N-3)(N-2)-(i-1)i+(i-2)$$

increases and a new term in the sum is added too. For a given i the term

$$h_i = \frac{2\,(N-i)}{N\,(N-1)}\left(\frac{1}{2}\,(N-3)\,(N-2) - (i-1)\,i + (i-2)\right)$$

is increased when N increases.

One can view the recalculations of the matching matrix as a perturbation of the matching process and thus equivalent to the RAS procedure. Besides the recalculations the procedure also needs to locate the appropriate interval (or matrix entry) from which to match into. This is also very computational demanding and will in general also increase proportional to N the number of individuals in the population. Thus the total number of computations is of complexity $O\left(n^2\right)$.

4.3 The SBAM procedure

The SBAM procedure does as follows: We have obtained the RAS-adjusted matrix X which is a symmetric $T \times T$ matrix in which x_{ij} is the number of matches between individuals of type i and j. Then given a matching pool of N individuals such that

$$\sum_i x_{ij} = \sum_j x_{ij} = N_i = N_j$$

assign the first $2 * x_{11}$ persons of personsgroup N_1 randomly into two queues, A^{11} and B^{11}, and pair those persons. Remove those persons from the set N_1, i.e. the new set of type 1 persons in the matching pool is $N_1 \setminus A^{11} \cup B^{11}$. Next, randomly choose x_{12} persons from each of N_1 and N_2 an randomly queue them $A^{12} \subset N_1$ and $B^{12} \subset N_2$ and pair those persons. Remove from N_1 (N_2) the persons of A^{12} (B^{12}). More generally, randomly choose x_{ij} for $i \leq j$ from the residual sets N_i and N_j respectively and queue them $A^{ij} \subset N_i$ and $B^{ij} \subset N_j$ and pair those persons. Remove from N_i (N_j) the persons of A^{ij} (B^{ij}). Since the RAS procedure is independent of the population size, the calculations needed to

adjust the distribution of matches in the SBAM does not increase as the population size increases. The only number of computations dependent on the number of individuals to match is the pairing mechanism which randomly queue the individuals in the matching pool.

Comparing the two alternatives: DYNACAN/CORSIM and SBAM, we see that the pairing mechanism, that is the actual part of the matching mechanism which make a pair on individuals a couple has the same computational requirements. However, the distributional adjustment differs in that the DYNACAN/CORSIM depends increasingly, convex on the total number of persons to match, while the SBAM adjustment is independent on the total number of persons to match.

5 An application

The SBAM method has been used in the development of a new Danish microsimulation model. The purpose of the model is to forecast the evolution and composition of Danish households and their demand for dwellings. The model works with a full sample of the Danish population of approximately 5,5 million individuals and 2,5 million households divided into 11 (geographical?) regions. The model describes demography, education, socio-economic status and housing choice.

Each individual will in every period (every year) with a given probability be included in the so-called *matching pool*. This probability depends on the characteristics of the individual. For example, a young person that is single, will have a high probability, while an older person living in a relationship, will have a lower probability. A person can be included in the matching pool either because it was previously single or it was previously matched but within a year the relationship had been separated and he/she had been matched with a new person.

5.1 Stock versus flow models

In this way, the present model differ from the models of Harmonic Mean and/or ChooSiow in that these alternative models use the stock-approach: the stock of matchings distributed on types at a given instant is determined by the matching function. The matching function applied here, the SBAM, is better suited to use the flow-approach: the stock of matchings distributed on types is determined partly by the flow of singles forming a new pair, a single and a previously matched person forming a new pair and previously matched persons separate from their previous matchings and becoming matched with each other, and partly by the separated couples within the period. The advantage of this new approach is that we can directly obtain the *gross* flows, while the stock-approach only determines the *net* flows as the change from the initial stock value into the ultimo stock value. When using microsimulation models it is often very important to know the *gross* flow as we model each individual separately. The stock approach is acceptable whenever the difference between the gross and net flows is neglectable as is the case in group-based population forecasts. The net flows can of course be estimated ex post, however, it is difficult to obtain the full effect of a changed matching distribution onto the conditional distributions. Also, the flow approach is important when the event, of e.g. a matching, is highly correlated with a different event, e.g. the migration event, which is also important in the housing demand model that we consider.

The model consists of at its most basic level of individuals and dwellings. At each point in time, an individual has some basic characteristics: gender, age, highest level of graduated education, citizenship, residence time and socio-economic status. Each individual is attached to a household, and each household is determined by the household members and the dwelling in which the household resides. The dwelling is characterized by its size, type and location. A household can be categorized as either a single or a couple, depending on the number of adult-members of the household, or as either an all-adult- or child-family, depending on the presens of a child in the household or not. The model distinguish between

individual-, household- and individual/household-specific events: an individual-specific event is the aging-event, education-event and socioeconomic-event. Household-specific events are events such as separation, construction, migration and fertility/expanded. A separation is caused by either one or two persons becomes either single or matched with another person, or that a person imigrates. A household is constructed either due to a separation from a current household, a matching of two new adults or the movement by an adult child away from his/her parents. A household can change location, size etc. of its dwelling. Finally, a household may be expanded either due to birth of a baby, adoption of a new child etc.

An important characteristic for a single person to be matched within a year is the event that the female gives birth to a child within the considered year. Furthermore, the likelihood of a couple's household to separation itself depend upon the birth of a child within that year, number of children etc. This implies that the initial state of a single person (or couple) does not perfectly determine that persons probability of matching (or separation) within the year. This requires the result of an in-year event of fertility to be determined ex ante the matching process begins. This casual ordering is the result from unobserved characteristics, i.e. the pregnancy decision.

Using the stock-approach the number of matchings and number of separations separately would not be determined, since only the *net* effect is known. However, the number of matches between two types may be unchanged, or a very small change in numbers, while this may cover up the fact that the actual effect is that an (almost) equal number of households are dissolved as the number of households constructed, and whoms gross number might be considerable. This, however, has consequences for the migration probabilities in that separation/construction of households involve a migration decision. This affects the demand for housing. To see how matching probabilities affects the migration propensities consider tabel X.

	Separation	Matching	Nb. pers.	Avg. per year	Migration prob
Adults	No	No	10.095.909	917.810	8,3
	Yes	No	924.932	84.085	62,2
	Yes	Yes	109.880	9.989	77,7
	No	Yes	1.046.790	95.163	61,5
Adult children	-	No	182.848	60.949	100,0
	-	Yes	206.034	18.730	96,2
Other	-	-	391.110	35.556	23,8

As can be seen, a person which experience a separation has a probability of migration of about 60 per cent as compared to a person which does not separate whom has a probability of migration of about 8 per cent. About one in fourth migration-event conincide with a separation- or matching-event (Kristensen, 2011). Since between 250 and 300 thousand adults migrate each year, this could potentially affect about 50-75 thousand migrations each year.

5.2 Empirical comparisons

In this way, a matching pool containing approximately 120,000 individuals arises each period. From this, the corresponding 60,000 pairs are formed. The SBAM algorithm is used for this. In the experiments reported in this paper, the individuals are divided into types on the basis of gender, age (15-65), 5 education levels and 11 regions. This results in 5,500 different types (=2*50*5*11). On a Windows-server (Intel Xeon CPU X5550, 2.67GHz), the matching takes approximately 20 seconds.

As we concluded above, using a few simple examples, the SBAM will use the same-sex matchings to compensate for any widening of the age-gender gap between males and females. The distribution of matchings is more concentrated towards the age-groups 22 to 30 in 2020 compared to the distribution in 2010. In the same period, the number of persons in these age-groups increase between 2010 and 2020. This follows the results from

our simple examples.

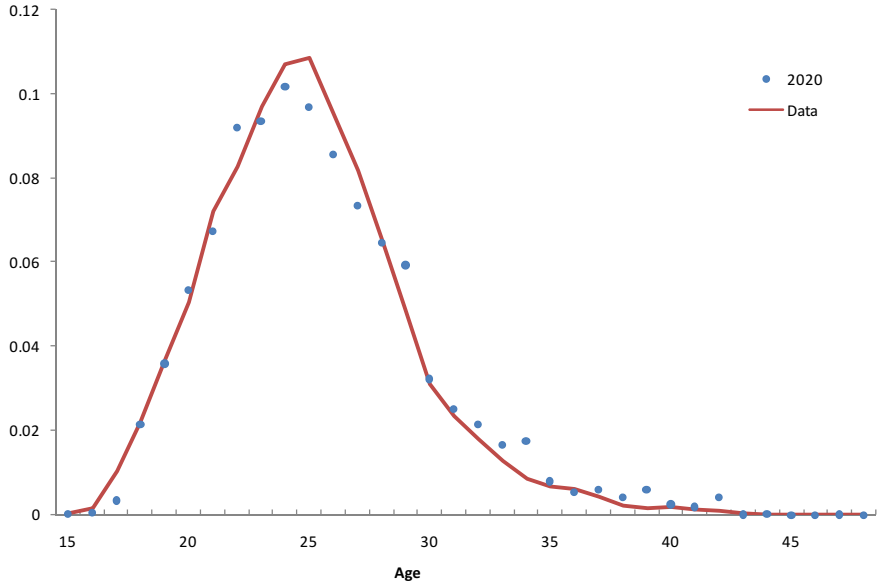
Figures 1-3 give examples of the results of the model in 2020. The figures show the distribution of newly formed pairs in the original data (from 2008) and in 2020. Figure 1 displays the age distribution of partners of 25-29 year old males. It is evident that SBAM is capable of generating an age distribution fairly consistent with data. The mean age of a partner is 25.0 in the data. In the forecast, the average is 25.4.

Figure 2 shows the educational distribution of partners for individuals with a vocational education. The SBAM algorithm finds it necessary to move the distribution slightly to ensure that the over-all matching is solved. In comparison to the original data, the proportions of partners with educational levels of "High school" and "Vocational" have thus fallen, while the proportions of "No education", "Medium" and "Long" have risen.

Figure 3 displays the regional distribution of partners for individuals living in "Copenhagen, environs". The Copenhagen area is divided into two regions: "Copenhagen, environs" (7) and "Copenhagen, city" (6). It is seen that approximately 50 per cent of new partners also live in the environs of Copenhagen. In addition, Copenhagen city and North Sealand (8) account for a significant proportion of new partners. It is evident that the SBAM method produces a distribution that is fairly consistent with the original data.

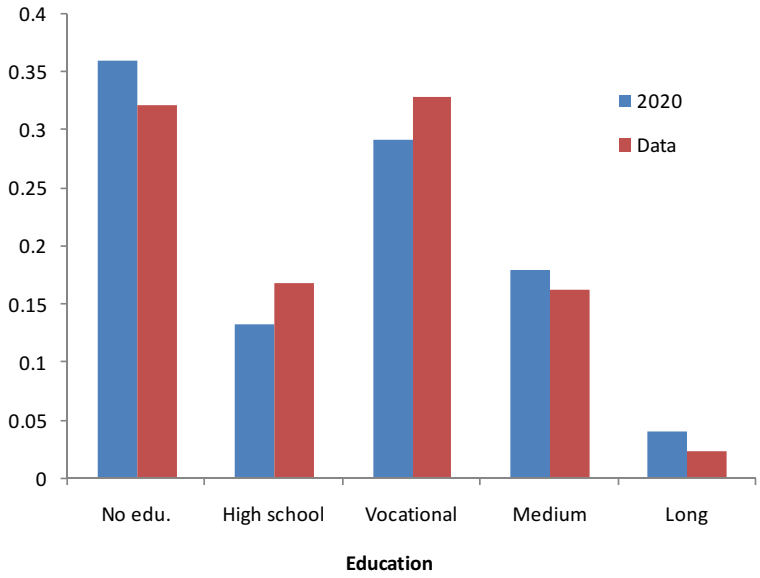
Figure 4 shows the distribution of matchings on the age difference. There is a significant change in the age-difference distribution as the matching ages, such that young couples tend to have a more lower age difference compared to later on in their life. When they turn age 35 the distribution is rather stationary.

Figure 1: Age distribution of partners. 25-29 year old males.



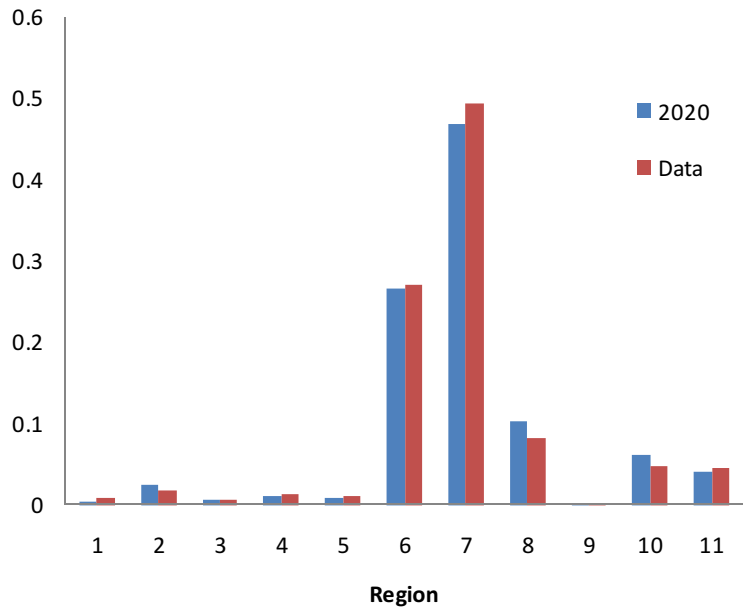
Source: Own calculations.

Figure 2: Educational distribution of partners. Vocational (?).



Source: Own calculations.

Figure 3: Regional distribution of partners. Copenhagen, environs.

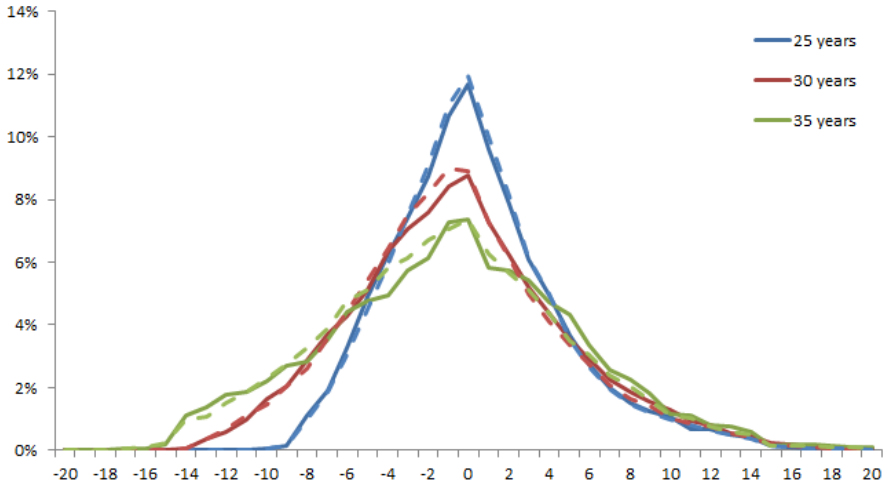


Note: Copenhagen, city=6, Copenhagen, environs=7, North Sealand=8.
Source: Own calculations.

6 Conclusions

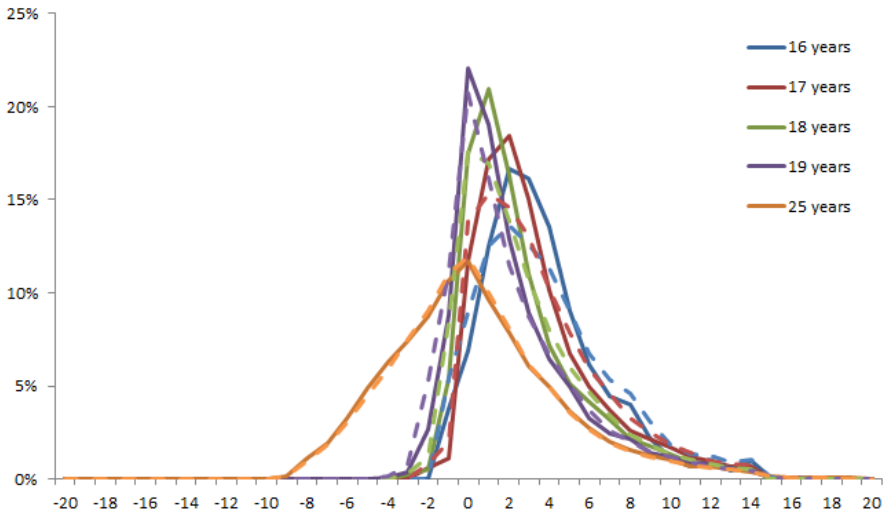
Most current microsimulation models are based upon a sample of individuals from the population, which limits the computational requirements in the matching process which is usually a quadratic complexity process and thus would increase quadratic as the sample size increases. Few models exceeds a sample size above 100.000 individuals which would imply a matching pool of approximately 2.500 persons. This low sample size is problematic when distributing the simulated population on a larger number of characteristics and limit the fineness of the distribution. This has invoked the modellers to apply rather coarse methods which does not allow matching distributions to vary with the age-level only the age-difference, improvements have been imployed by allowing for a distinction between first time marriages and higher order marriages but also satisfactory implications from

Figure 4: Distribution age-difference matching, historic vs. estimated



Note: *Solid lines are estimated shares, while dotted lines are observed. Data is from the period 2001-2008. The horisontal vertex is the age difference of a matching couple. Source: Own calculations.*

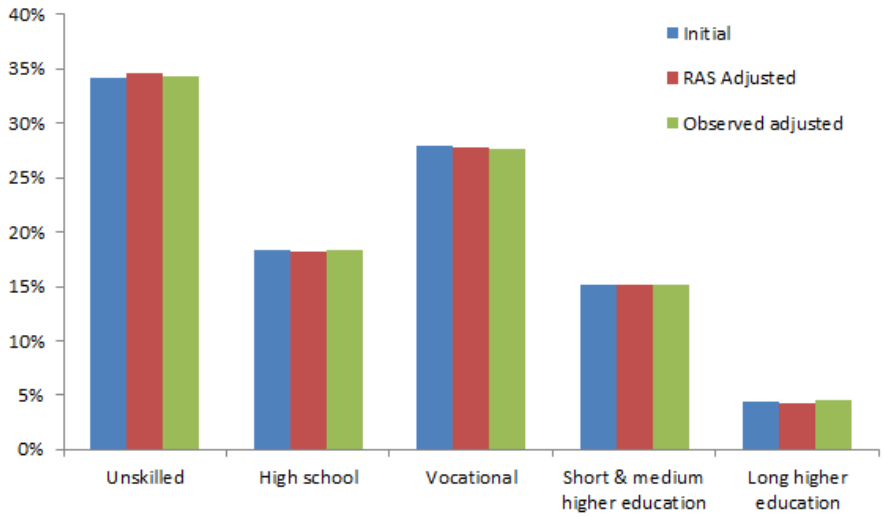
Figure 5: Age difference distribution, young ages



Note: *The probability of a person of a given age is matched with a person of a given age-difference. Solid lines are adjusted, dotted lines are observed.*

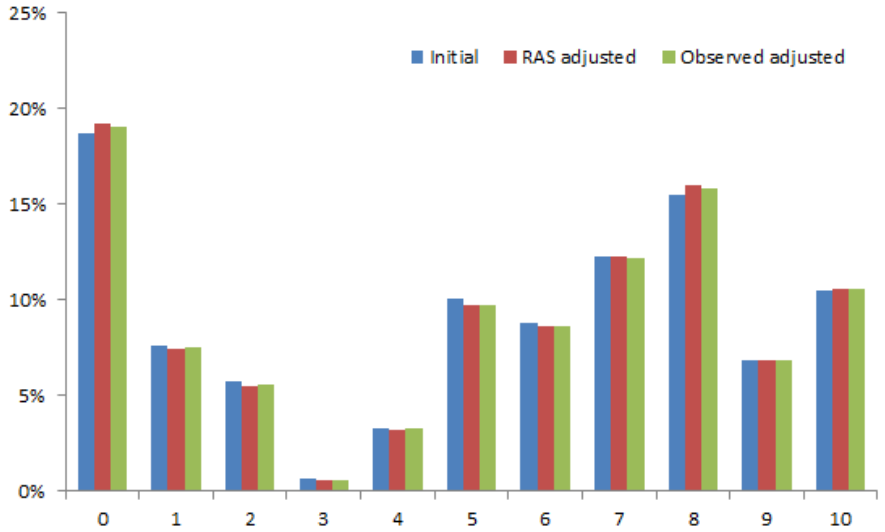
Source: *Own calculations.*

Figure 6: Distribution on education



Note: *The probability of a person matched in 2006 with a match of a given education.*
Source: *Own calculations*

Figure 7: Distribution on geographical region



Note: *The categories are: Copenhagen city, Copenhagen suburb, Northern sealand, Bornholm, Eastern sealand, Western/southern sealand, Fyn, Southern Jutland, Eastern jutland, Western jutland, Northern jutland. The probability that a person matching a person in 2006 is matched with a person in a given region.*
Source: *Own calculations*

this approach arises due to the male-centric assumption.

We introduce a new method for computational efficient matching based on the RAS method of rebalancing matrices and the linked-lists method of C#. Overall the method mimics the stable approach emencing from changes in population distributions establishing the theoretical fundation on solid consistent grounds while retaining the observed distribution, which is often a problem for stable approaches that tends to produce bimodal distributions that cluster too much around the center of the distribution. On the other hand the method is rather efficient in its computational requirements.

Applications of the method to the historical period 2001-2009 of observed matchings compared with the SBAM method reveals that the method in general perform very well in predicting the distribution on age differentials while the most serious errors are observed in a small groupe of 16-18 years.

7 References

(1) McDougall, Robert A. (1999) ”Entropy Theory and RAS are Friends”. GTAP Working Paper 5-14-1999

(2) Bregman, Lev M. (1967) “Proof of the convergence of Sheleikhovskii’s method for a problem with transportation constraints”, USSR Computational mathematics and mathematical Physics, 1(1), 191-204, 1967.

(3) Gale, D. and Shapley, L. (1962) “College admissions and the stability of marriage”, American Mathematical Monthly 69, pp. 9-14

(4) Schneider, Michael H. and Zenios, Stavros A. (1990) “A Comparative Study of Algorithms for Matrix Balancing”. Operations Research, Vol. 38, No. 3 (May - Jun., 1990), pp. 439-455.

(5) Choo, E. & Siow, A. (2006), “Who marries whom and why”, Journal of Political Economy vol .114, No. 1, pp. 175-201

(6) Shannon, C.E. (1948) “A mathematical theory of communication”. Bell System Technical Journal, 27:379–423, 623–659.

(7) Schoen, Robert (1988) “Modeling multigroup populations”, Plenum Press, New York

(8) Kristensen Joakim B. (2011) “Det danske boligmarked i 2000’erne - kortlægning af boligbestand og flyttebevægelser”, DREAM working paper 2011:3.

(9) Easther, Richard and Jan Vink. 2000. “A Stochastic Marriage Market for CORSIM.” Strategic Forecasting Technical Paper.