PLOS ONE

RESEARCH ARTICLE

Entropy Based Modelling for Estimating Demographic Trends

Guoqi Li[1,2☯]*, Daxuan Zhao[2,3☯], Yi Xu[4☯], Shyh-Hao Kuo[2], Hai-Yan Xu[2], Nan Hu[2], Guangshe Zhao[5]*, Christopher Monterola[2]*

1 Department of Precision Instrument, Center for Brain-Inspired Computing Research, Tsinghua University, Beijing, P.R.China, **2** Institute of High Performance Computing, A*STAR, Singapore, Singapore, **3** School of Businesses, Renmin University of China, Beijing, P.R.China, **4** School of computing engineering, Nanyang Technological University, Singapore, Singapore, **5** School of Aerospace, Xi'an Jiaotong University, Xi'an, P. R. China

☯ These authors contributed equally to this work.
* liguoqi@mail.tsinghua.edu.cn (GL); zhaogs@mail.xjtu.edu.cn (GZ); monterolac@ihpc.a-star.edu.sg (CM)

Citation: Li G, Zhao D, Xu Y, Kuo S-H, Xu H-Y, Hu N, et al. (2015) Entropy Based Modelling for Estimating Demographic Trends. PLoS ONE 10(9): e0137324. doi:10.1371/journal.pone.0137324

Editor: Zhong-Ke Gao, Tianjin University, CHINA

Received: May 14, 2015

Accepted: August 16, 2015

Published: September 18, 2015

Data Availability Statement: Data are from the following website https://international.ipums.org/international/.

Funding: This work is supported by the Integrated City Planning Programme, Science and Engineering Research Council (SERC), Agency for Science, Technology and Research (A*STAR: http://www.a-star.edu.sg/) (Grant no. 1325000001), and Complex Systems Programme, SERC, A*STAR (Grant no. 1224504056). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare no competing interests.

Abstract

In this paper, an entropy-based method is proposed to forecast the demographical changes of countries. We formulate the estimation of future demographical profiles as a constrained optimization problem, anchored on the empirically validated assumption that the entropy of age distribution is increasing in time. The procedure of the proposed method involves three stages, namely: 1) Prediction of the age distribution of a country's population based on an "age-structured population model"; 2) Estimation the age distribution of each individual household size with an entropy-based formulation based on an "individual household size model"; and 3) Estimation the number of each household size based on a "total household size model". The last stage is achieved by projecting the age distribution of the country's population (obtained in stage 1) onto the age distributions of individual household sizes (obtained in stage 2). The effectiveness of the proposed method is demonstrated by feeding real world data, and it is general and versatile enough to be extended to other time dependent demographic variables.

Introduction

Predicting demographic trends (DT) [1] in the light of emerging complex processes [2] of the 21st Century continues to be an important and open research topic. Understanding developments and the changes in population is critical in assisting governments in targeting policies for the future and saving money for education, public health, retirement, transportation, energy consumption among others [3][4]. Specifically, DT refers to the changes in the joint distribution between population with time, age or other demographic factors, such as household's size, health measures, economic status, religious affiliation, education, marriage, etc [5][6][7].

Forecasting DT is a challenging task, and remains to be a fundamental concern in both basic and applied ecology [8]. The complexity lies in the DT'S intricate connectivity to the heterogeneous activities of a large group of individuals, and it is impacted by observed and

unobserved time dependent factors [9][10]. Existing methods such as the least square methods [11][12][13] and Bayesian inference [14], in spite of being the most extensively used procedures in estimating and predicting various engineering problems, fail to capture the driving mechanisms of complex processes that shapes DT [15]. There are very few literatures on building optimization models for understanding DT. Typical approaches involve incorporating factors such as environmental [16][17][18], demographic [19] and/or observer-related covariates [20]. However, data to support and verify such techniques is often not readily available as [21]–[23] suggesting that building an optimization model constrained by limited data to characterise DT is fundamentally important with a lot of potential applications.

Entropy-based methods, the measure of the uncertainty in random variables, have been successfully applied to many modelling and estimation problems, as seen in [24][25][26][27]. In this paper, we introduce the entropy-based method to estimate DT. We build the model motivated by our empirical observation that the age distribution of population follows an increasing entropy trend. The paradigm is based on minimizing the entropy-based objective function and incorporating some parameters describing the historical trends into the constraints where the dynamic and intrinsic properties can be reflected. We illustrate this procedure by estimating the evolution of demographic distributions over ages and household sizes. Our work involves a three-fold modeling stages. Firstly, an "age-structured population model" based on Leslie matrix [28][29][30][31] is used to predict the age distribution of a country's population. This makes the modelling of the demographic temporal distributions become possible, as one usually needs to project the age distribution of population into other factors. Secondly, the age distribution of each household size is estimated based on a proposed entropy-based model, where we propose an entropy formulated cost function and incorporate the DT into the constraint conditions. The model applied in this stage is called "individual household size model". Finally, the age distribution of the country's population (obtained in stage 1) is projected onto the age distributions of individual household size types (obtained in stage 2), which we refer to as "total household size model". Note that our estimation does not rely on any observed determinant on the formation of households. The evolution of the household size is estimated based on the historical information and the entropy principle.

To compare with existing works [3][32], our method predicts DT with limited information [33]. The output is a joint distribution of age and other demographic variables over time. Among its applications will be on policy analysis, economic forecasting and urban planning and so on. For the purpose of illustration, we use the population data from US Census and predict the age DT for each household size in 2010, based on the historical data in 2000 and 2006. The remaining parts of the paper are organized as follows. Section 2 lists the definitions and notations which are used throughout the article. Section 3 presents the three stages for the estimation of DT. The simulation results based US data are illustrated in Section 4 and we conclude the article in Section 5.

Methodology

Notations

In the following, we list the definitions and notations that will be used throughout the article:

t: The year index.

$.^T$: Matrix transpose.

$\hat{.}$: The estimation of a variable.

A_{upper}: The upper bound age.

i: The age index ($i = 0, 1, \ldots, A_{upper}$).

$P_i(t)$: The population for the people at age i (older than i but younger than $i + 1$) in the year t.

$P(t) = [P_0(t) \ldots P_i(t) \ldots P_{A_{upper}}(t)]^T$: The population vector for the people at all ages in the year t.

$N_i(m, t)$: The population for the male at age i in the year t.

$N_i(f, t)$: The population for the female at age i in the year t.

$N(m, t) = [N_0(m, t)\ N_1(m, t) \ldots N_{A_{upper}}(m, t)]^T$: The population vector for the male at all ages.

$N(f, t) = [N_0(f, t)\ N_1(m, t) \ldots N_{A_{upper}}(f, t)]^T$: The population vector for the female at all ages.

$D_i(m, t)$: The death rate for a male at age i in the year t.

$D_f(i, t)$: The death rate for a female at age i in the year t.

$B_i(t)$: The fertility rate for a female at age i in the year t.

$Ratio_{mf}(t)$: The ratio of the number of newly born boys to girls in the year t.

$Immig(m, t)$: The male immigrants vector in the year t.

$Immig(f, t)$: The female immigrants vector in the year t.

$Emig(m, t)$: The male emigrants vector in the year t.

$Emig(f, t)$: The female emigrants vector in the year t.

m_0: Total number of household sizes.

j: The household size index ($j = 1, 2, \ldots, m_0$).

k_0: Number of historical years' data used in the individual household size Model (12).

κ: An index applied on the historical data for the year $t - \kappa$ ($\kappa = 0, 1, \ldots, k_0$).

G_n: The people in the age interval $[0\ A_n]$ where A_n is an upper bound age of this group.

n: The group number index of G_n ($n = 1, 2, \ldots, n_0$) (as seen in Formula (10)).

n_0: Number of groups (G_n) in the individual household size Model (12).

$p_i^j(t)$: The probability (percentage) that people at age i in household size j in the year t.

$p_i(t) = [p_0(t) \ldots p_{A_{upper}}(t)]^T$: Age distribution of the population in the year t.

$p^j(t) = [p_0^j(t)\ \ldots\ p_{A_{upper}}^j(t)]^T$: Age distribution of household size j in the year t.

$q_i^j(t-\kappa) = [p_0^j(t-\kappa)\ \ldots\ p_{A_{upper}}^j(t-\kappa)]^T$: Age distribution of household size j in the year $t - \kappa$.

$Entropy(t)$: The entropy of the population distribution in the year t.

$\alpha_n^j(t+1)$: A ratio of people in group G_n to the population in household size j in Formula (10).

$\tilde{\alpha}_n^j(t+1)$: A parameter defined in Model (12).

$\{.\}_i$: The vector that contains the values of the variable {.} by changing subscript i.

ω_0^*: A weight of the objective function in Model (12).

$\omega_1, \ldots, \omega_{k_0}$: The weights defined in Model (12) and Formula (13).

$\xi_k^j(t+1)$: An error term in Formula (14).

$\bar{\xi}$: The upper bound of $\xi_k(t+1)$ and $\xi_k(t+1) \in [-\bar{\xi}, \ \bar{\xi}]$.

H: The hessian matrix.

$x^j(t)$: The number of household size j ($j = 1, \ldots, m_0$) in the year t.

$X(t) = [x^1(t) \ \ldots \ x^j(t) \ \ldots x^{m_0}(t)]'$: The vector contains the number of each household size.

W: A weighting matrix in the total household size Model (16).

τ: A parameter in the matrix W in the total household size model (Formula (18)).

u: A small positive weight parameter in the total household size Model (16).

$\hat{F}$: Predicted weighting matrix by collecting the predicted age distributions of all household sizes.

Three stages for forecasting the demographic trends

Fig 1 summarizes the three stages for forecasting the DT. Stage 1: using an "age-structured population model" to predict the population in the year $t + 1$. Stage 2: using an "individual household size model" to estimate the age distribution for each household size j based on data in the historical years where the DT reflected in the previous years can be incorporated into the constraint conditions. Stage 3: Combining the results from Stages 1 and 2, and employing a "total household size model" to predict the number of each household size. We detail in the next subsections each of the three stages shown.

Age-structured population model: for estimating age distribution of the population

We consider the population as a summation of all the organisms of the same group or species, who live in the same geographical area, and have the capability of inter-breeding. Quite

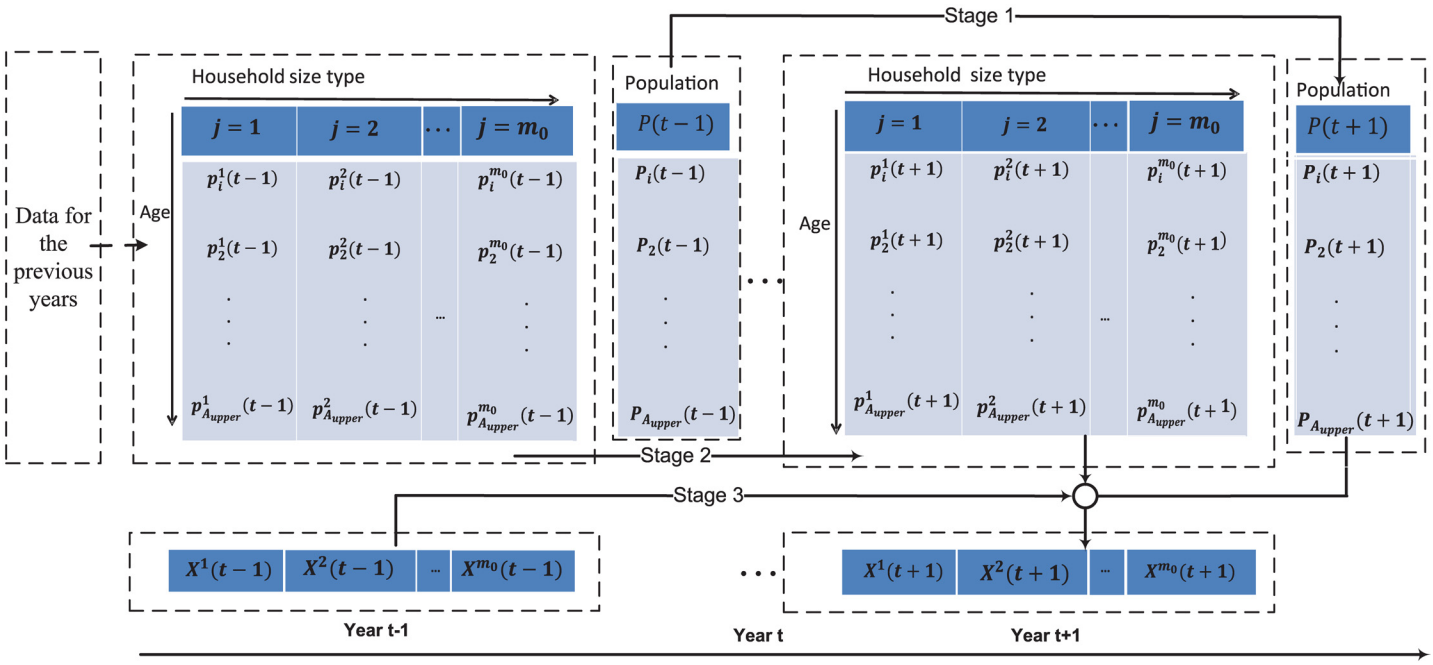

Fig 1. Illustration of the three stages for forecasting the demographic trends.

doi:10.1371/journal.pone.0137324.g001

frequently, the prediction of demographic temporal distributions is highly linked to the population's age-structure. Demographic temporal distribution modeling is achievable using the "age-structured population model" since it allows projection of the age distribution into other factors.

Assumptions. We apply the Leslie matrix method [28]–[31] that assumes:

a. There is no plague, disaster or war that will lead to abrupt changes in age specific death rate.

b. Statistical variables such as birth rates and birth ratio are slowly changed and predictable.

c. The fertility rate for both local residents and immigrants is the same.

d. All people who are older than A_{upper} are in the same age group. Here, we set $A_{upper} = 90$.

Problem formulation. We first consider the case without immigration and emigration. In the year $t + 1$, the number of people at age $i + 1$ is

$$\begin{aligned} P_{i+1}(t+1) &= N_{i+1}(m,t+1) + N_{i+1}(f,t+1) \\ &= [1 - D_i(m,t)]N_i(m,t) + (1 - D_i(f,t))N_i(f,t) \end{aligned} \tag{1}$$

where t and $t + 1$ denote the current year and the next year, respectively, and $i = 0, 1, \ldots, A_{upper} - 1$ is the age index. When $i = A_{upper}$, we have

$$\begin{aligned} P_{i\geq A_{upper}}(t+1) = \quad & [1 - D_{A_{upper}-1}(m,t)]N_{A_{upper}-1}(m,t) + [1 - D_{A_{upper}-1}(f,t)]N_{A_{upper}-1}(f,t) \\ & + [1 - D_{i\geq A_{upper}}(m,t)]N_{i\geq A_{upper}}(m,t) + [1 - D_{i\geq A_{upper}}(f,t)]N_{i\geq A_{upper}}(f,t) \end{aligned} \tag{2}$$

Let $[i_1, i_2]$ be the age interval that a female has the ability to give birth. Then, $P_0(t + 1) = N_0(m, t + 1) + N_0(f, t + 1)$ and

$$\begin{aligned} N_0(m,t+1) &= \frac{Ratio_{mf}(t)}{1 + Ratio_{mf}(t)} \sum_{i=i_1}^{i=i_2} B_i(t)N_i(f,t) \\ N_0(f,t+1) &= \frac{1}{1 + Ratio_{mf}(t)} \sum_{i=i_1}^{i=i_2} B_i(t)N_i(f,t) \end{aligned} \tag{3}$$

where $Ratio_{mf}(t)$ is a ratio of the newly born boys ($N_0(m, t)$) to the newly born girls ($N_0(f, t)$) at year t. Let

$$\begin{aligned} P(t) &= [P_0(t)\ P_1(t)\ \ldots\ P_{A_{upper}}(t)]^T \\ N(m,t) &= [N_0(m,t)\ N_1(m,t)\ \ldots\ N_{A_{upper}}(m,t)]^T \\ N(f,t) &= [N_0(f,t)\ N_1(f,t)\ \ldots\ N_{A_{upper}}(f,t)]^T \end{aligned} \tag{4}$$

be the vectors of the population, male population and female population, respectively, for ages between 0 and A_{upper} at year t.

Next, we extend the model to take into account of immigration effects. Let $Immig(m, t)$/$Emig(m, t)$ and $Immig(f, t)$/$Emig(f, t)$ be the respective immigrants and emigrants vector for

males/females at year t. We obtain the "age-structured population model" as follows:

$$\begin{aligned} P(t+1) &= N(m,t+1) + N(f,t+1) \\ N(m,t+1) &= A(t)N(m,t) + \frac{Ratio_{mf}(t)}{1+Ratio_{mf}(t)}B(t)N(f,t) \\ &\quad + Immig(m,t) - Emig(m,t) \\ N(f,t+1) &= [C(t) + \frac{1}{1+Ratio_{mf}(t)}B(t)]N(f,t) \\ &\quad + Immig(f,t) - Emig(f,t) \end{aligned} \tag{5}$$

where $A(t)$, $B(t)$ and $C(t)$ are the matrices constructed based on Eqs (2)–(4), and given by

$$A(t) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1-D_0(m,t) & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1-D_1(m,t) & 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1-D_{A_{upper}-1}(m,t) & 1-D_{i\geq A_{upper}}(m,t) \end{bmatrix} \tag{6}$$

$$B(t) = \begin{bmatrix} 0 & \dots & 0 & B_{i_1}(t) & \dots & B_{i_2}(t) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \tag{7}$$

$$C(t) = \begin{bmatrix} 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 1-D_0(f,t) & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 1-D_1(f,t) & \dots & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 0 & 1-D_{A_{upper}-1}(f,t) & 1-D_{i\geq A_{upper}}(f,t) \end{bmatrix} \tag{8}$$

Note that the population data we collected allows us to estimate the values of all the above parameters (such as the fertility rates and death rates). These parameters change slowly and are predictable which confirm the validity of our assumption. Thus, the population distribution for the coming years can be predicted based on the age-structured population Model (5), and its estimation is denoted as $\hat{P}(t+1)$ for the year $t+1$ as shown in Model (16) later.

Individual household size model: for estimating age distribution for each household size

In this section, we will describe in detail our *individual household size model* that estimates the age distribution of each household size. The model is operated by minimizing an entropy based objective function and using the historical trends as constraints, where both the dynamic and intrinsic properties are reflected.

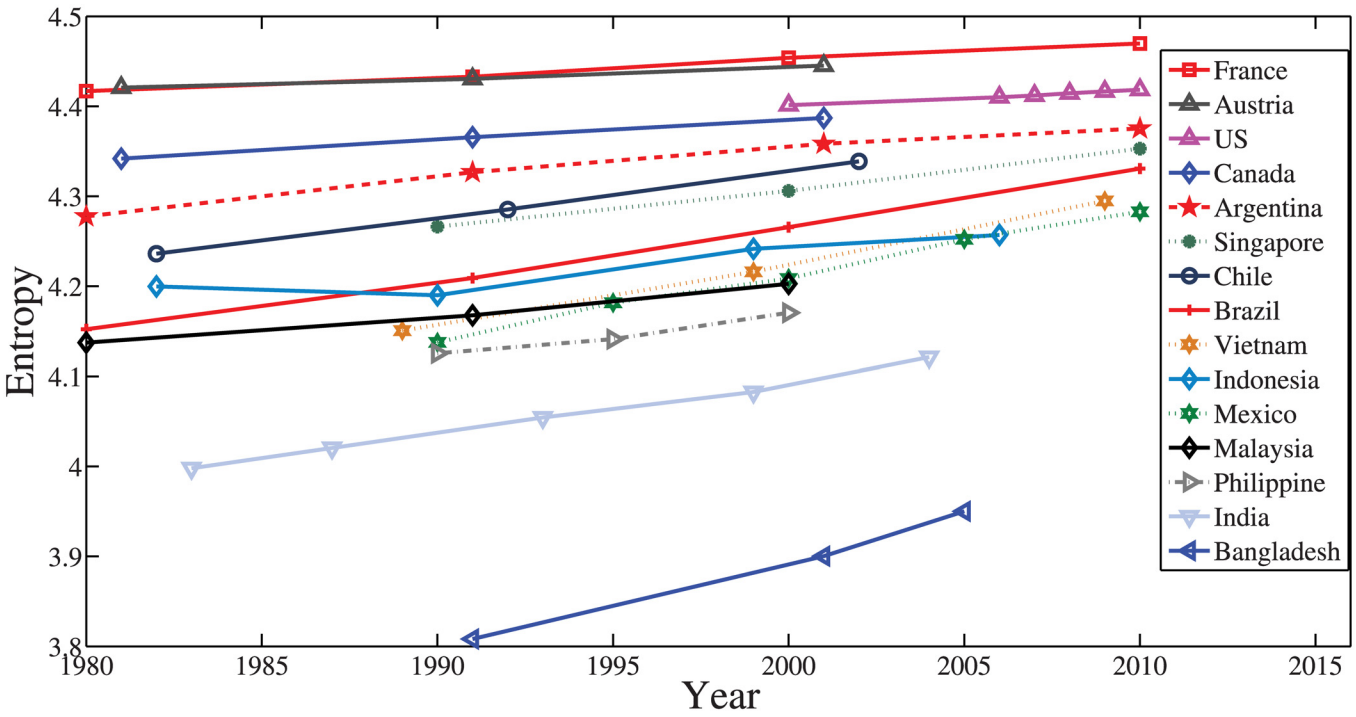

Fig 2. The age distribution entropy of selected countries as a function of time. Note that there is a slight decrease in the entropy of the Indonesia's population from 1990 to 2000 perhaps due to the difference in the statistics method used in the years (1990 vs 2000) considered as indicated in https://international.ipums.org/international/.

doi:10.1371/journal.pone.0137324.g002

Let $p_i(t)$ be the probability that a person is at age i in year t. We define an entropy function for year t as follows:

$$Entropy(t) = -\sum_{i=0}^{A_{upper}} p_i(t) ln(p_i(t)) \tag{9}$$

where $\sum_{i=0}^{A_{upper}} p_i(t) = 1$ and $p_i(t) \geq 0$.

Fig 2 plots the entropy of the age distribution based on the population data collected from six countries. In general, the entropy of the age distribution increases monotonically with respect to time in most countries. This observation suggests that we can estimate the age distribution of a particular household size based on entropy concepts. To this end, we divide the household size into n_0 types: i.e., 1 person per household, 2 persons per household, . . ., until n_0 persons per household.

Let j be the household size index and assume that we already have the age distributions for each household size j ($j \in \{1, \ldots, m_0\}$) in the years t, $t-1, \ldots, t-k_0$, which are denoted as $q_i^j(t-\kappa)$ for $\kappa = 0,1, \ldots, k_0$. Let $p_i^j(t+1)$ represent the percentage of persons at age i in household size j in the year $t+1$. This means we group the people whose ages are above 90 years together. Our objective is to estimate the age distribution $p_i^j(t+1)$ in the year $t+1$ based on the historical data.

We group the people from 0 to A_{upper} years old into n_0 groups, i.e., the groups G_n for $n = 1, \ldots, n_0$, where $n_0 \ll A_{upper}$. The age interval for the group G_n is $[0, A_n]$ and $0 < A_1 < A_2 < \ldots < A_{n_0} = A_{upper}$. It is easy to see that $G_{n-1} \subset G_n$. Define $\alpha_n^j(t)$ as a parameter such that

$$\alpha_n^j(t) = \sum_{i=0}^{A_n} p_i^j(t) \tag{10}$$

which means that $\alpha_n^j(t)$ is a ratio of people in group G_n, i.e., in the age interval $[0, A_n]$, to the population in household size j. Note that $\forall j \in \{1, \ldots, m_0\}$, $\alpha_{n_0}(t+1) = 1$ since $A_{n_0} = A_{upper}$.

Let

$$\tilde{\alpha}_n^j(t+1) = \alpha_n^j(t+1) - \alpha_n^j(t) \tag{11}$$

be the parameter which reflects the percentage change of the ratio $\alpha_n^j(t)$ from the year t to the next year $t+1$.

From here, we build an individual household size model to predict the age distribution $\{p_i^j(t+1)\}_i$ for each household size type j where $j = 1, 2, \ldots, m_0$, by optimizing the following:

$$\begin{aligned}
min \quad & \sum_{\kappa=0}^{k_0} \omega_{\kappa+1} \sum_{i=1}^{A_{upper}} p_i^j(t+1) ln\left(\frac{p_i^j(t+1)}{q_i^j(t-\kappa)}\right) \\
& + \omega_0^* \sum_{i=1}^{A_{upper}} p_i^j(t+1) ln\left(\frac{p_i^j(t+1)}{1/A_{upper}}\right) \\
s.t \quad & \sum_{i=0}^{A_1} p_i^j(t+1) \le \sum_{i=1}^{A_1} q_i^j(t) + \tilde{\alpha}_1^j(t+1) \\
& \vdots \\
& \sum_{i=0}^{A_n} p_i^j(t+1) \le \sum_{i=1}^{A_n} q_i^j(t) + \tilde{\alpha}_n^j(t+1) \\
& \vdots \\
& \sum_{i=0}^{A_{n_0}} p_i^j(t+1) \le \sum_{i=1}^{A_{n_0}} q_i^j(t) + \tilde{\alpha}_{n_0}^j(t+1) \\
& \sum_{\kappa=1}^{k_0+1} \omega_\kappa = 1 \\
& -p_i^j(t+1) \le 0 \quad for \quad i = 0, 1, ..., A_{upper}
\end{aligned} \tag{12}$$

Again, given that the entropy of the population is monotonically increasing with time, we can minimize an entropy based cost function under some constraints by employing the historical data. Compared with Eq (9), we omit the minus sign "–" such that the model becomes a minimization problem. The upper limit of such entropy as $t = +\infty$ is a uniform distribution with a histogram function having a constant $1/A_{upper}$ magnitude. Essentially, there are two parts in this cost function where ω_0^* is a small positive weight parameter. The first part is the cross entropy distance (KL distance [34]) between $\{p_i^j(t+1)\}_i$ and the historical data, and the second part is the relative entropy distance between $\{p_i^j(t+1)\}_i$ and population distribution when $t = +\infty$.

Note that we can never know the value of $\tilde{\alpha}_n^j(t+1)$ at the year t as we do not know $\alpha_n^j(t+1)$. However, it can be estimated from the historical data as:

$$\begin{aligned}
\hat{\tilde{\alpha}}_n^j(t+1) &= \sum_{\kappa=1}^{k_0} \omega_\kappa(\alpha_n^j(t-\kappa) - \alpha_n^j(t))/\kappa \\
&= \sum_{\kappa=1}^{k_0} \omega_\kappa\left(\sum_{i=0}^{A_n} q_i^j(t-\kappa) - \sum_{i=0}^{A_n} q_i^j(t)\right)/\kappa
\end{aligned} \tag{13}$$

where ω_κ for $\kappa = 1, \ldots, k_0 + 1$ are decreasing weights, which implies that the more recent data is more valued. Let $\xi_n^j(t+1) = \frac{\tilde{\alpha}_n^j(t+1) - \hat{\tilde{\alpha}}_n^j(t+1)}{\tilde{\alpha}_n^j(t+1)}$ be an error term of the estimation, then we have

$$\tilde{\alpha}_n^j(t+1) = \hat{\tilde{\alpha}}_n^j(t+1)(1 + \xi_n(t+1)) \tag{14}$$

the distribution of $\xi_n^j(t+1)$ is known and bounded within $[-\bar{\xi}, \bar{\xi}]$. Usually $\xi_n^j(t+1)$ can be assumed as a random variable uniformly distributed in $[-\bar{\xi}, \bar{\xi}]$. We now have that:

Theorem 1. The optimization problem defined in Model (12) is a strict convex optimization.

Proof. Note that the Hessian matrix H of the objective function is given by:

$$H = \begin{bmatrix} \frac{\sum_{\kappa=0}^{k_0} \omega_{\kappa+1}}{p_1^j(t+1)} & 0 & 0 & \dots & 0 \\ 0 & \frac{\sum_{\kappa=0}^{k_0} \omega_{\kappa+1}}{p_2^j(t+1)} & 0 & \dots & 0 \\ & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{\sum_{\kappa=0}^{k_0} \omega_{\kappa+1}}{p_{A_{upper}}^j(t+1)} \end{bmatrix} \tag{15}$$

Since $p_i^j(t+1) \geq 0$ for all i and j, it is easy to see that H is a positive definite matrix. On the other side, it is known that the constraints of the optimization problem in the Model (12) are linear. Therefore, the feasible domain is a convex set. Both the objective function and the feasible domain are convex, hence the problem is a convex optimization. Note that one only needs to find a local minimum point of a convex optimization to obtain the global minimum point [35][36][37][38].

Total household size model: for estimating the number of each household size

In this section, we build a *total household size model* to further estimate the number of each household size j for $j = 1, 2, \dots, m_0$ based on the predicted age distribution of population and age distribution of each individual household size. Here, our objective is to estimate the number of household size j for $j = 1, 2, \dots, m_0$ in the year $t + 1$.

Let $x^j(t)$ be the number of household with size j in the year t and denote that $X(t) = [x^1(t) \ \dots \ x^{m_0}(t)]^T$. We hope to estimate the vector $X(t+1) = [x^1(t+1) \ \dots \ x^{m_0}(t+1)]^T$. As mentioned, the first stage is to obtain the estimated total population distribution $\hat{P}(t+1)$ based on the current fertility rate and death rate. The second stage is then to obtain the estimated age distribution of each household type j denoted as $\hat{p}^j(t+1)$. Now we estimate the household number distribution by solving the following total household size model:

$$\begin{aligned} min \quad & (1-u) \cdot ||\hat{F} \cdot X(t+1) - \hat{P}(t+1))||^2 + u \cdot ||W \cdot X(t+1) - X(t))||^2 \\ s.t \quad & X(t+1) > 0 \end{aligned} \tag{16}$$

where $||.||$ is the L_2 norm, and $X(t+1) \geq 0$ means each component of $X(t+1)$ is nonnegative, and $\hat{F}$ is a weighting matrix collected from the the predicted age distributions of all household sizes:

$$\hat{F} = \begin{bmatrix} 1 \cdot \hat{p}_1^1(t+1) & 1 \cdot \hat{p}_2^1(t+1) & \dots & 1 \cdot \hat{p}_{A_{upper}}^j(t+1) \\ 2 \cdot \hat{p}_1^2(t+1) & 2 \cdot \hat{p}_2^2(t+1) & \dots & 2 \cdot \hat{p}_{A_{upper}}^2(t+1) \\ \vdots & \vdots & \dots & \vdots \\ m_0 \cdot \hat{p}_1^{m_0}(t+1) & m_0 \cdot \hat{p}_2^{m_0}(t+1) & \dots & 1 \cdot \hat{p}_{A_{upper}}^{m_0}(t+1) \end{bmatrix} \tag{17}$$

and W is a diagonal weighting matrix for different household size type given by

$$W = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 2^{\tau} & \dots & \vdots \\ \vdots & \vdots & j^{\tau} & \vdots \\ 0 & 0 & \dots & m_0^{\tau} \end{bmatrix} \tag{18}$$

The above objective function contains two parts with u being a small positive weight parameter. The first part is the distance between the estimated age distribution for population and the accumulative of the age distribution for all household sizes. The other part is the weighted distance of the estimated $X(t+1)$ (denoted as $\hat{X}(t+1)$) to $X(t)$. As there are j persons in the household size j, we construct a diagonal weighting matrix W with a given power $\tau > 0$ in Eq 18. As shown in Theorem 2, the optimization of Eq (16) is also convex.

Theorem 2. The optimization problem defined in Model (16) is convex.

Prof. The proof is similar to Theorem 1. The Hessian matrix of the objective function in Eq (16) is

$$H = (1-u)\hat{F}^T\hat{F} + uW^TW \tag{19}$$

Obviously H is a positive definite matrix and we have this theorem holds.

Simulations

In this section, we illustrate the procedure we have discussed above using the US's Census population data. We predict the demographic distribution in the year 2010 based on the historical data in years 2000 and 2006. The prediction is then compared with the actual Census data in the year 2010. We show that the method we described here accurately captures the actual statistics. As mentioned, there are three stages in the estimation:

Stage 1: Estimating the age distribution of the population by employing the *age structure based population model* in Section 3A.

Stage 2: Estimating age distribution for each household size type by employing the *individual household size model* in Section 3B.

Stage 3: Estimating the number of different household size type by employing the *total household size model* in Section 3C.

In Stage 1, we collect the population data from the US Census and get the values of all parameters that are required in Model (5). By solving this model, we obtain the estimation of the population in the year 2010 based on the the data in the year 2000 and 2006 in Fig 3.

In Stage 2, by letting ω = [0.95 0.025 0.025] and assuming that the error term bound $\bar{\xi} = 0$, we divide the population into 9 groups (G_n, n = 1, 2, . . ., 9) and let $A_n = n{\cdot}10$. By solving the individual household size model (12), the age distributions for the household sizes j = 1 and j = 2, . . ., 7 are obtained in Figs 4–10, respectively. It is seen that the individual household size model predicts accurately the age distribution of all household sizes.

In stage 3, the numbers of each household size are estimated by solving the total household size Model (16). As seen in Fig 11, the difference between the estimation and the real values is quite close, which again shows the accuracy of our proposed method.

In addition, we also look at the cases when the error term bound $\bar{\xi} \neq 0$. Let's say, $\bar{\xi} = 2$ implying that $\bar{\xi}_k(t+1)$ is randomly distributed in the interval [−2 2]. By repeating the

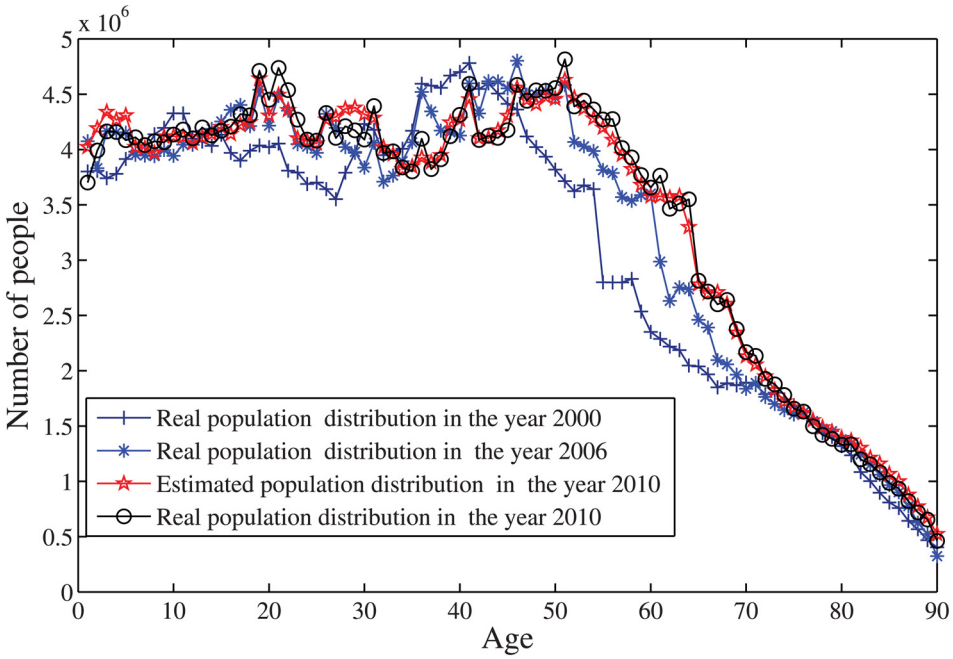

Fig 3. The predicted age distribution of US population for the year 2010 based on the population data in the years 2000 and 2006.

doi:10.1371/journal.pone.0137324.g003

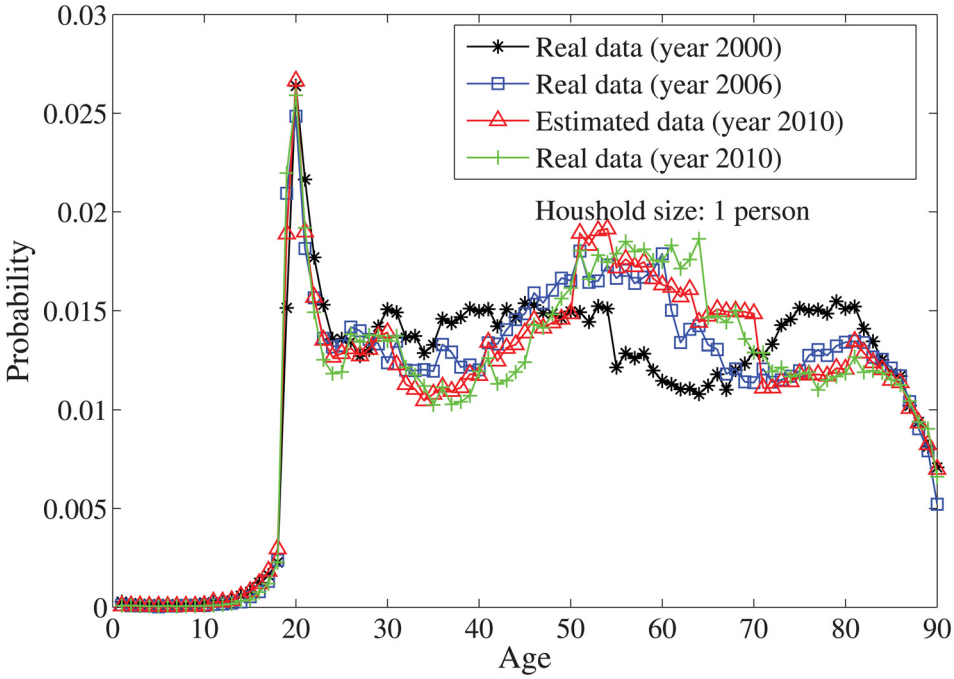

Fig 4. The estimated age distribution for household size type 1 in the year 2010 (x axis is the age index and y axis is the probability).

doi:10.1371/journal.pone.0137324.g004

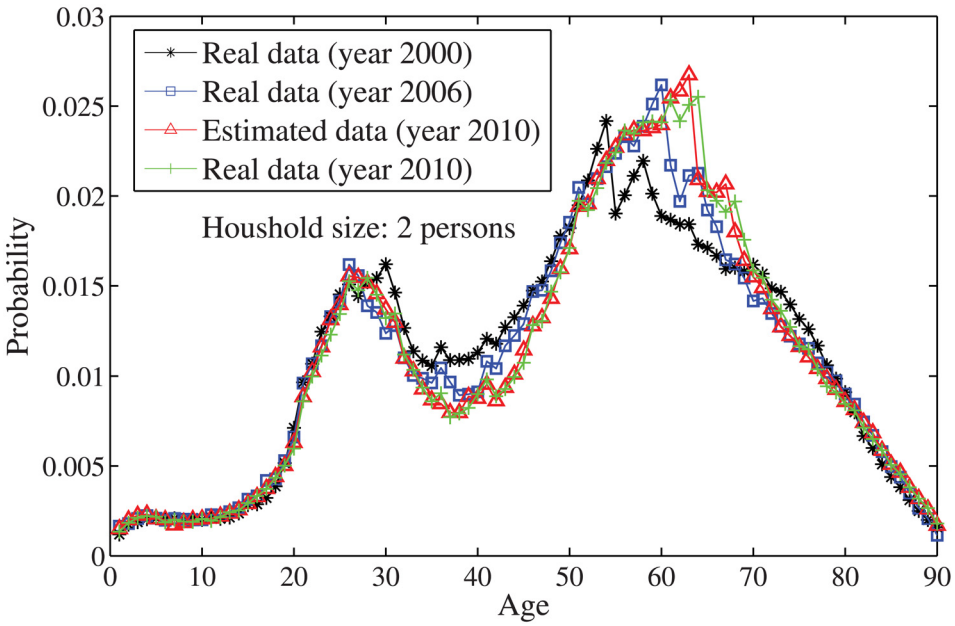

Fig 5. The estimated age distributions for household size 2 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g005

simulations many times over (200 times in our case), we can verify the robustness of our estimations. We take the household size 1 as an example. As seen in Fig 12, most of the real values of the age distribution are located inside the red area generated by the estimations.

To investigate how the parameters τ and u affect the performance of the estimation, we set different values for them and see how the error term $Error(\hat{X}(t+1)) = \frac{\|\hat{X}(t+1)-X(t+1)\|_1}{\|X(t+1)\|_1}$ changes,

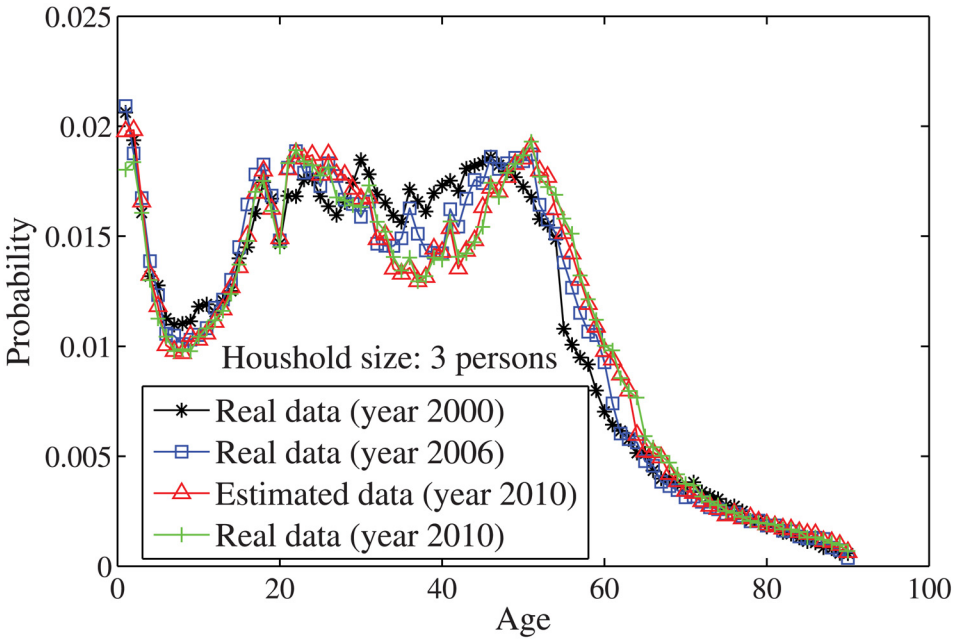

Fig 6. The estimated age distributions for household size 3 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g006

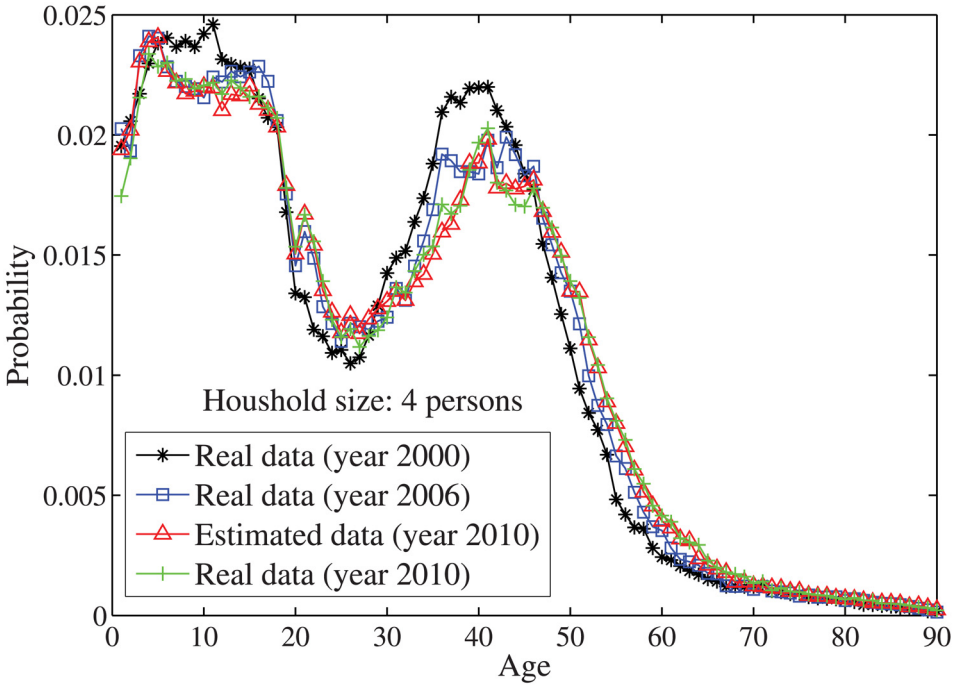

Fig 7. The estimated age distributions for household size 4 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g007

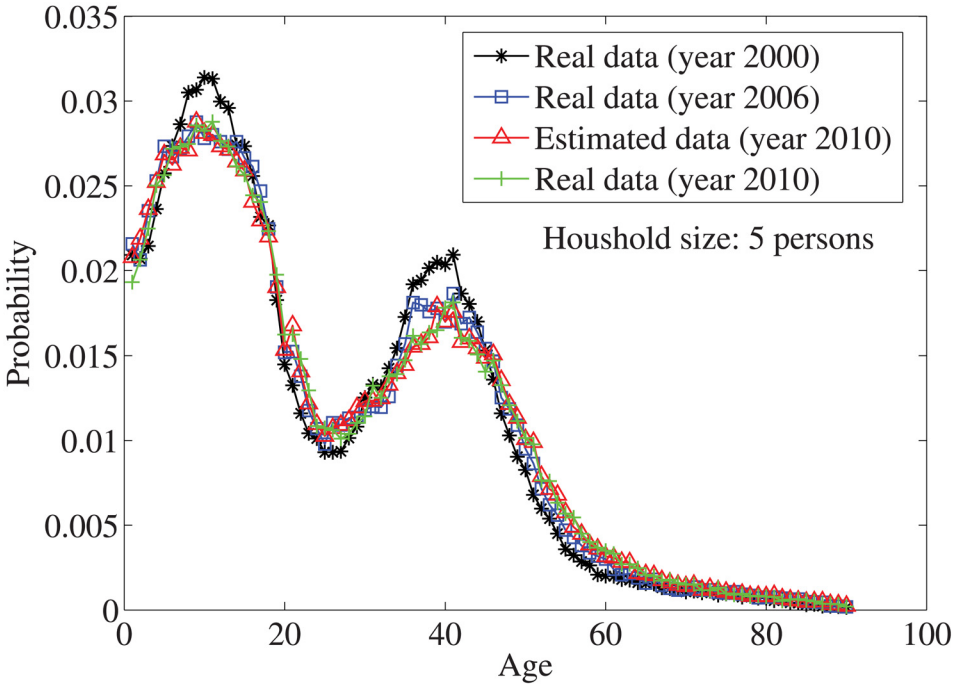

Fig 8. The estimated age distributions for household size 5 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g008

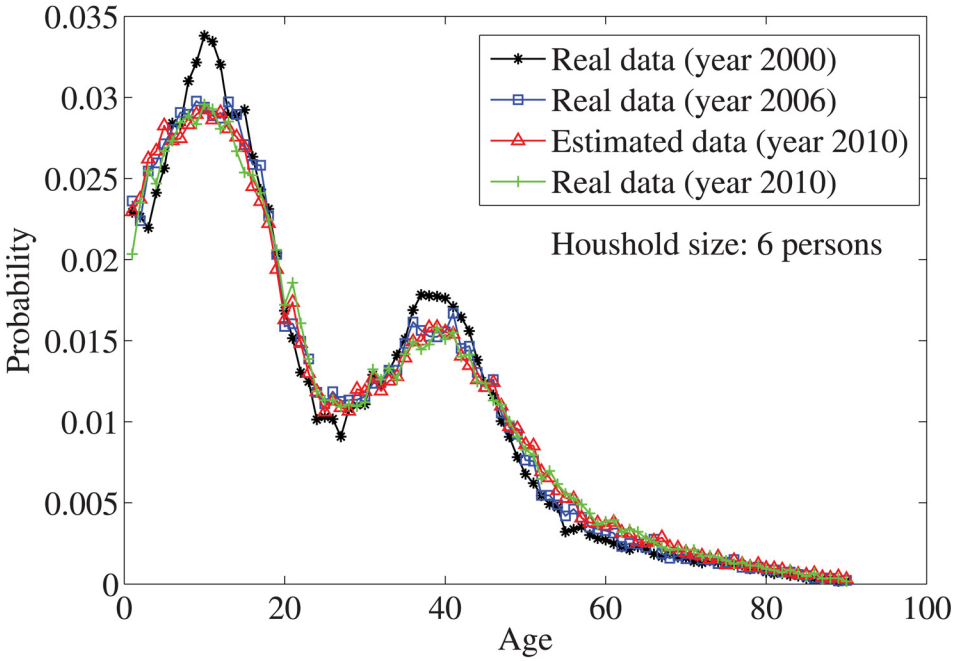

Fig 9. The estimated age distributions for household size 6 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g009

where $\|.\|_1$ denotes the L_1 norm. Fig 13 shows the error term against different values of τ. We observed that when τ is around 0.5, the proposed algorithm achieves its best performance. This is reasonable since the diagonal elements $W^T W$ is just the number of persons in each household size when $\tau = 0.5$. On the other hand, Fig 14 shows how the parameter u impacts the

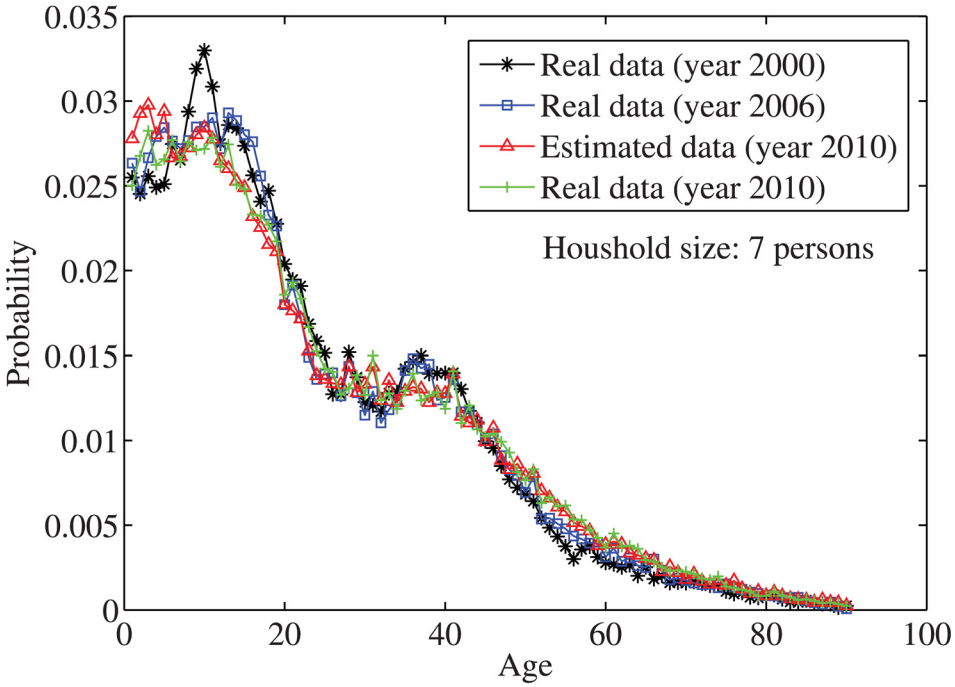

Fig 10. The estimated age distributions for household size 7 in the year 2010 using US data.

doi:10.1371/journal.pone.0137324.g010

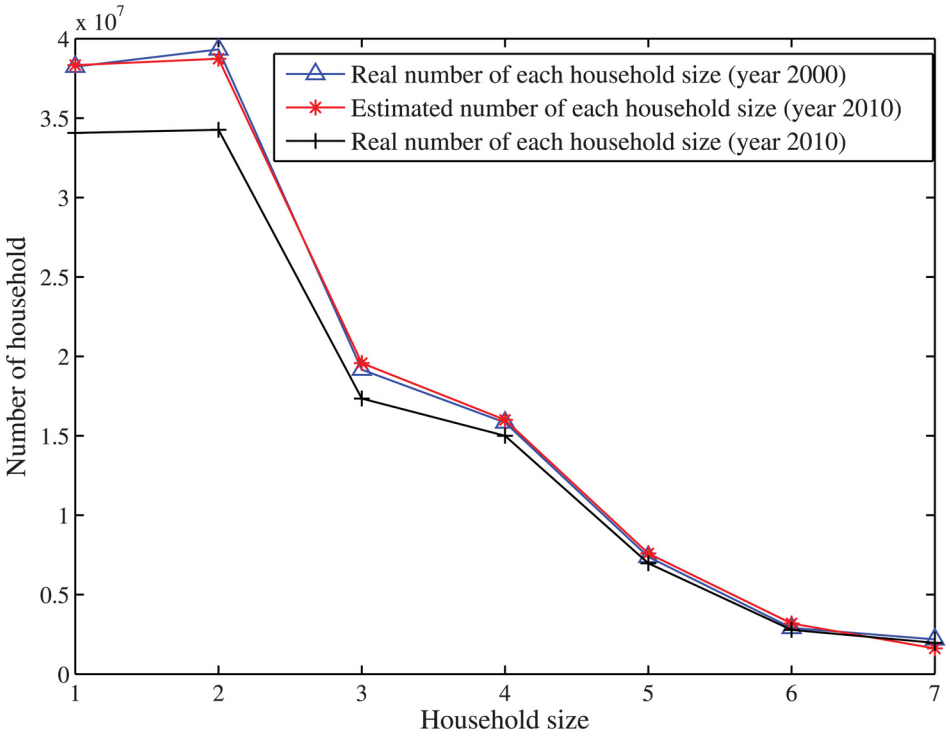

Fig 11. The prediction of number of household size distribution for US in the year 2010.

doi:10.1371/journal.pone.0137324.g011

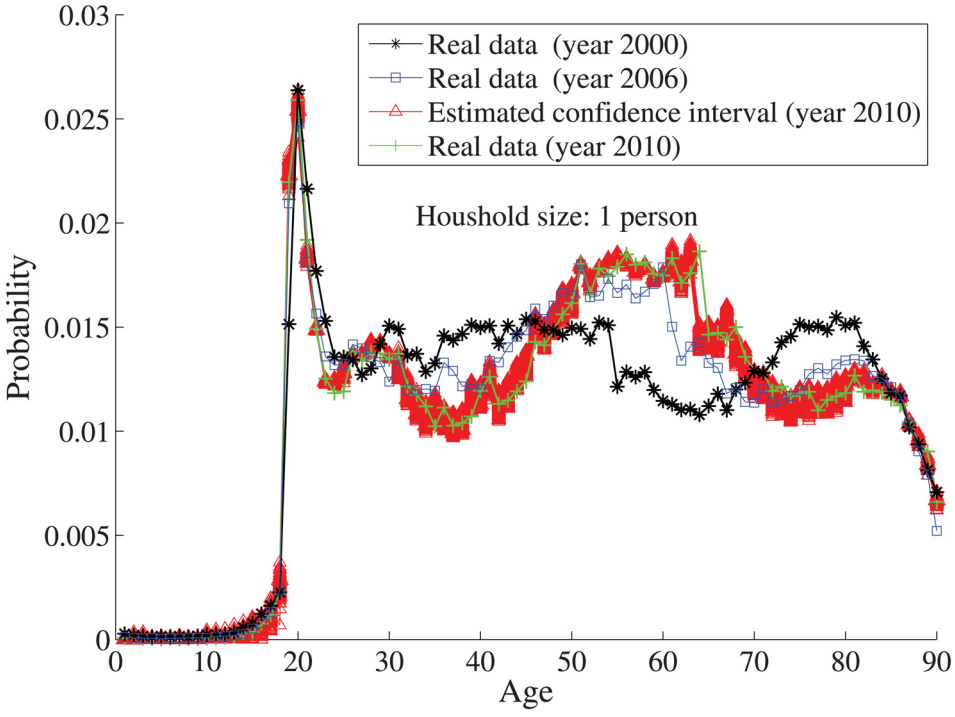

Fig 12. The illustration of the robustness of the estimation compared with their real values for household size 1.

doi:10.1371/journal.pone.0137324.g012

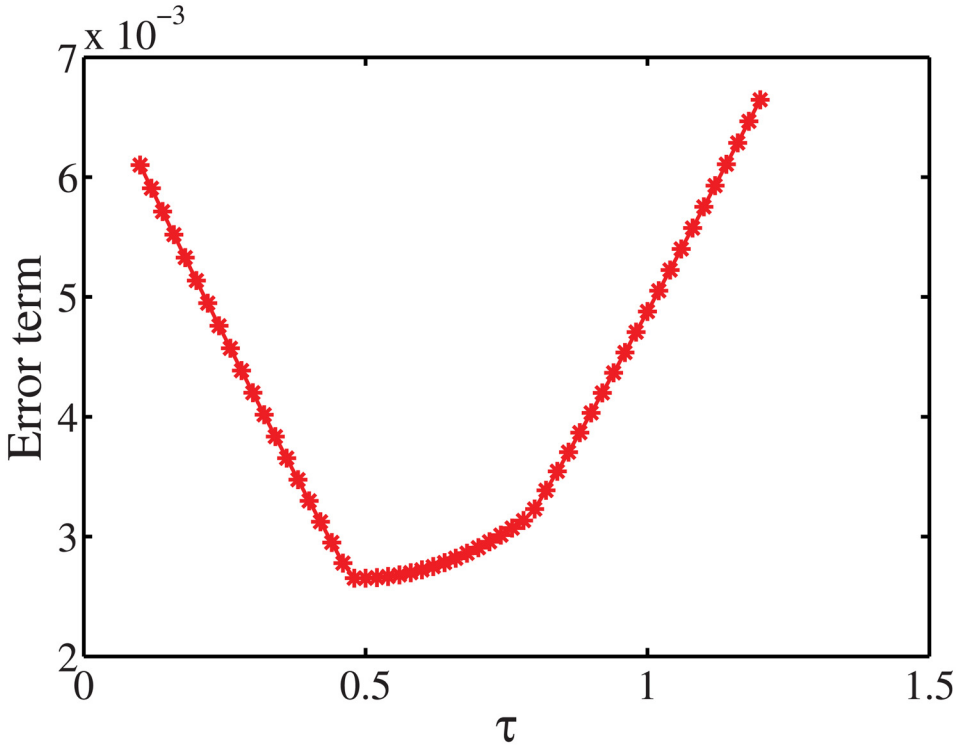

Fig 13. Error term with respect to the parameters τ.

doi:10.1371/journal.pone.0137324.g013

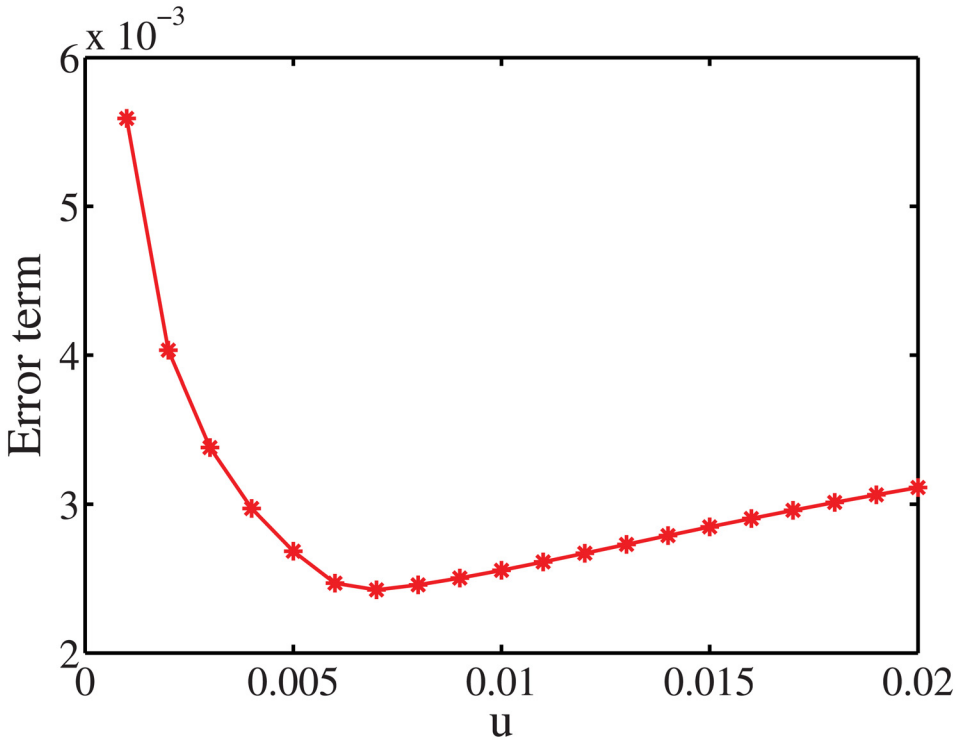

Fig 14. Error term with respect to the parameters *u*.

doi:10.1371/journal.pone.0137324.g014

estimation error. By setting different values for u in the interval [0 0.2], we observe that the performance of the algorithm provides the best fit when $u \in [0.005\ 0.01]$.

Discussion and Conclusions

In this paper, we have demonstrated a new method that estimates the development of age and household's size distributions. The procedure consists of three models in three coupled stages, we referred to as: *the age-structured population model* in stage 1 where the age distribution of countries' population was predicted; *the individual household size model* in stage 2 where the age distribution of each individual household size was estimated; and *the total household size model* in stage 3 where the number of different household sizes was derived by projecting the age distribution of total population onto the age distributions of individual household sizes. The procedure described here indicates that demographic trends can be accurately estimated using entropy as an optimisation variable, which we believe will be of potential interest to both academics and practitioners alike. We have illustrated and validated the correctness and accuracy of the proposed method using US data. While we have considered age and household size distributions in this article, we note that the method we have demonstrated is general and versatile enough to be extended to other time dependent demographic variables.

Acknowledgments

This work is supported by the Integrated City Planning Programme, SERC, A*STAR (Grant no. 1325000001), and Complex Systems Programme, SERC, A*STAR (Grant no. 1224504056).

Author Contributions

Conceived and designed the experiments: GL DZ YX CM. Performed the experiments: GL SHK HYX NH. Contributed reagents/materials/analysis tools: GZ CM. Wrote the paper: GL DZ YX SHK HYX NH GZ CM.

References

1. Lee R, The demographic transition: three centuries of fundamental change. The Journal of Economic Perspectives. 2003; 17:167–190. doi: 10.1257/089533003772034943
2. Gao ZKet al, Multiscale complex network for analyzing experimental multivariate time series. Europhysics Letters. 2015; 109: 30005p1–30005p6. 2015. doi: 10.1209/0295-5075/109/30005
3. Swanson D and Hough G, An evaluation of persons per household (PPH) estimates generated by the american community survey: A Demographic Perspective. Population. 2012; 31:235–266.
4. Xu HY, Kuo SH, Li G, Legara EFT, Zhao G and Monterola CP, Generalized Cross Entropy Modeling for Estimating Joint Distribution from Incomplete Information, A*STAR working paper, 2014.
5. O'Neilla BC, Daltonb M, Fuchsc R, Jianga L, Pachauric S and Zigovad K, Global demographic trends and future carbon emissions, Proceedings of the National Academy of Sciences of the United States of America (PNAS). 2010; 107:17521–17526. doi: 10.1073/pnas.1004581107
6. Stephenson J, Newman K and Mayhew S, Population dynamics and climate change: what are the links? Journal of Public Health. 2010; 32:150–156. doi: 10.1093/pubmed/fdq038 PMID: 20501867
7. Salvo JJ, and Brown WA, Population estimates and the needs of local governments, Paper Presented at U.S. Census Bureau Conference on population Estimates: Meeting User Needs (Alexandria VA). 2006.
8. Humbert JY, Mills LS, Horne JS and Dennis B, Journal compilation. A better way to estimate population trends. 2009; 118:1940–1946.
9. Gao ZK and Jin ND, A directed weighted complex network for characterizing chaotic dynamics from time series. Nonlinear Analysis-Real World Applications. 2012; 13:947–952. doi: 10.1016/j.nonrwa.2011.08.029

10. Gao ZK, Fang PC, Ding MS and Jin ND, Multivariate weighted complex network analysis for characterizing nonlinear dynamic behavior in two-phase flow, Experimental Thermal and Fluid Science. 2015; 60:157–164. doi: 10.1016/j.expthermflusci.2014.09.008
11. Csiszar I, Why least squares and maximum entropy? An axionatic approach to inference linear inverse problem, Annual of Statistics. 1991; 19:2032–2066.
12. Markovsky I and Huffel SV, Overview of total least-squares methods, Signal processing. 2007; 87:2283–2302. doi: 10.1016/j.sigpro.2007.04.004
13. Li G and Wen C, Identification of Wiener systems with clipped observations, IEEE Transactions on Signal Processing. 2012; 60:3845–3852. doi: 10.1109/TSP.2012.2190404
14. Aster RC, Borchers B and Clifford H, Parameter estimation and inverse problems, Second Edition. Elsevier. 2012.
15. Saadi S and Rahman A, Evidence of non-stationary bias in scaling by square root of time: Implications for Value-at-Risk, Journal of International Financial Markets, Institutions and Money. 2008; 18:272–289. doi: 10.1016/j.intfin.2006.12.001
16. Dennis B, and Otten MRM, Joint effects of density dependence and rainfall on abundance of San Joaquin kit fox, The Journal of Wildlife Management. 2000; 64:388–400. doi: 10.2307/3803237
17. Beyene J, Fallah S, Bull SB, Tritchler D, Chan V and Knight J, Modeling complex disease with demographic and environmental covariates and a candidate gene marker. Genetic Epidemiology. 2001, 21:423–428.
18. Trawinski PR and Mackay DS, Identification of environmental covariates of West Nile virus vector mosquito population abundance. Vector Borne and Zoonotic Diseases. 2010; 10:515–26. doi: 10.1089/vbz.2008.0063 PMID: 20482343
19. Owens J, Dickerson S and Macintosh DL, Demographic covariates of residential recycling efficiency. Environment and behaviour. 2000; 32:637–650. doi: 10.1177/00139160021972711
20. Link WA and Sauer JR, New approaches to the analysis of population in Land Birds: Comment. Ecology. 1997; 78:2632–2634. doi: 10.1890/0012-9658(1997)078%5B2632:NATTAO%5D2.0.CO;2
21. Fagan WF, Meir E, Prendergast J, Folarin A and Karieva P, Characterizing population vulnerability for 758 species. Ecology Letters. 2001; 4:132–138. doi: 10.1046/j.1461-0248.2001.00206.x
22. Inchausti P and Halley J, Investigating long-term ecological variability using the global population dynamics database. Science. 2001; 293:655–657. doi: 10.1126/science.293.5530.655 PMID: 11474102
23. Brook BW, and Bradshaw CJA, Strength of evidence for density dependence in abundance time series of 1198 species. Ecology. 2006; 87:1445–1451. doi: 10.1890/0012-9658(2006)87%5B1445:SOEFDD%5D2.0.CO;2 PMID: 16869419
24. Abbas AE, Entropy methods for joint distributions in decision analysis. IEEE Transactions on Engineering Management. 2006; 53:146–159. doi: 10.1109/TEM.2005.861803
25. Phillips SJ, Anderson RP and Schapire RE, Maximum entropy modeling of species geographic distributions. Ecological Modeling. 2006; 190:231–259. doi: 10.1016/j.ecolmodel.2005.03.026
26. Leyk S, Nagle NN and Buttenfield BP, Maximum entropy dasymetric modeling for demographic small area estimation. Geographical Analysis. 2013; 45:285–306. doi: 10.1111/gean.12011
27. Salois MJ, Regional changes in the distribution of foreign aid: An entropy approach, Physica A, 2013; 392:2893–2902. doi: 10.1016/j.physa.2013.02.007
28. Kot M, Elements of mathematical ecology, Cambridge. Cambridge University Press, 2001.
29. Leslie PH, On the use of matrices in certain population mathematics. Biometrika. 1945; 33:183–212. doi: 10.1093/biomet/33.3.183 PMID: 21006835
30. Leslie PH, Some further notes on the use of matrices in population mathematics. Biometrika. 1948; 35:213–245. doi: 10.1093/biomet/35.3-4.213
31. Guckenheimer J, Oster GF and Ipaktchi A, The dynamics of density dependent population models. Journal of Mathematical Biology. 1977; 4:101–147. doi: 10.1007/BF00275980
32. Smith SK, Nogle J and Cody S, A regression approach to estimating the average number of persons per household. Demograhy. 2002; 39:697–712. doi: 10.1353/dem.2002.0040
33. Miller DJ and Liu W, On the recovery of joint distributions from limited information. Journal of Econometrics. 2002; 107: 259–274. doi: 10.1016/S0304-4076(01)00123-3
34. Contreras-Reyes JE, Asymptotic form of the KullbackLeibler divergence for multivariate asymmetric heavy-tailed distributions. Physica A. 2014; 395:200–208. doi: 10.1016/j.physa.2013.10.035
35. Kuhn HW and Tucker AW, Nonlinear programming. Proceedings of 2nd Berkeley Symposium (Berkeley: University of California Press). 1951; 481–492.

36. Li G, Ning N, Ramanathan K, He W, Pan L and Shi L, Behind the magical numbers: hierarchical chunking and the human working memory capacity, International journal of neural systems. 2013; 23:1350019. doi: 10.1142/S0129065713500196 PMID: 23746292

37. Li G, Wen C and Zhang A, Fixed point iteration in identifying bilinear models, Systems & Control Letters. 2015; 83:28–37. doi: 10.1016/j.sysconle.2015.06.008

38. Li G, Wen C, Zheng WX and Zhao G, Iterative identification of block-oriented nonlinear systems based on biconvex optimization, Systems & Control Letters, 2015; 79:68–75. doi: 10.1016/j.sysconle.2015.01.003