# A model of age heaping with applications to population graduation that retains informative demographic variation

**Dalkhat M. Ediev** [1]

**Abstract.** Age heaping remains an issue in both historical and contemporary demographic studies. Traditional smoothing techniques are problematic in dealing with age heaping in cases where the population age structure shows both the digit preference and informative variation by age that should not be lost during the graduation. Same applies to more recent modelling-based smooth reconstructions of the latent population distributions. We generalize and modify an earlier model where age rounding's propensity depends on the distance to the round age and the strength of age heaping at that age. Efficient and robust estimation method is proposed for parameterizing the model that allows reconstructing the latent population distribution by age. We test our model in comparison to the traditional alternatives on an ample set of empirical data. Our method is capable of removing age heaping without substantial distortions of the actual population variation by age. In comprehensive empirical testing, our method appears the best in both the quality of heaping removal and retaining the informative population variation. The method has good potential for a wide practical application.

**Keywords.** Age heaping, digit preference, graduation, smoothing population data, behavioural model.

**IPC keywords:** Age structure, Methodology, Applied demography, Mathematical demography.

## 1. Introduction

Despite dramatic improvements in population data coverage and accuracy over time, age heaping (inaccuracies emerging from rounding the ages of census/survey respondents) remains an issue in historical and contemporary demographic studies (Shryock and Siegel 1973; Myers 1940; Pardeshi 2010; Jowett and Li 1992; Szołtysek, Poniat, and Gruber 2018). Traditionally the extent of age heaping at ages divisible by five is measured by the Whipple index $K_5$:

$$K_5 = 100 \cdot \frac{mean(N_{25},N_{30},\dots,N_{60})}{mean(N_{23},N_{24},N_{25},\dots,N_{60},N_{61},N_{62})}, \quad (1)$$

where $N_x$ is the population at age $x$. The heaping index measures the excess of population size at ages ending in 5's and 0's (i.e., round ages divisible by 5) in the age range from 23 to 62 years. The heaping at 'more round' ages divisible by ten only is measured by a similar index $K_{10}$:

$$K_{10} = 100 \cdot \frac{mean(N_{30},N_{40},\dots,N_{60})}{mean(N_{23},N_{24},N_{25},\dots,N_{60},N_{61},N_{62})}. \quad (2)$$

Judging by these indices, even the recent census data collected by the UN (United Nations Statistics Division 2020) show substantial age heaping. About 13 per cent of the data referring to the current century show either of the heaping indices exceeding 120.

Outside the population studies, age heaping may even be a useful phenomenon indicative of underlying social and ethnographic conditions (e.g., A'Hearn, Baten, and Crayen 2009). (Also, see further down an unusual example of exceptional heaping at ages divisible by twenty in a population practicing the vigesimal counting system.) Yet, data affected to age heaping need to be cleansed from it before using in demographic analyses, such as computing the demographic rates, conducting

[1] (a) North-Caucasian State Academy; (b) Lomonosov Moscow State University (Department of Demography, HSMSS); ediev@nsa.ru; ediev@iiasa.ac.at. The research leading to these results has received funding from the Russian Foundation for Basic Research under Grant 18-01-00289 "Mathematical models and methods of correcting the distortions of the age structure and mortality rates of the elderly population". The author is thankful to the anonymous reviewer of the earlier draft of the paper for useful comments and suggestions.

population projections, etc. Age heaping is typically a sign of other problems with age data inaccuracy too, such as age exaggeration. While methods for dealing with the latter problem have been advanced (Horiuchi and Coale 1982; Mitra 1984; Ediev 2018, 2021), removing age heaping remains largely limited to general-purpose smoothing techniques, such as the moving average or the smoothing splines and polynomials (Shryock and Siegel 1973; United Nations 1983; Whittaker 1922; Yusuf et al. 2014). Recently, a range of new modelling-based approaches have been suggested in order to address digit preference in non-demographic data, such as blood units (Heller and Dunlop 2012), head circumference (Wang and Wertelecki 2013), income, weight and height (Groß and Rendtel 2016; Camarda, Eilers, and Gampe 2017; Zinn and Würbach 2016), etc. However, these works assume smoothness of the underlying latent distributions and, similar to the traditional graduations, tend to produce overly smoothed reconstructions when applied to demographic data. Such methods, in particular, may distort or smooth out informative differences in sizes of age groups caused to short- and medium-term fertility variation, migration, presence of specific age contingents, such as military personnel, etc.

Here, we offer another modelling-based approach fit to the needs of demographic graduation that is capable of both suppressing the age heaping and retaining the informative variation of sizes of age groups. Our approach develops ideas of Ediev's (2003) model that decomposes the age heaping into structural (which ages might be rounded to which ones) and heaping strength components. The following section provides a description of the model and the graduation method. The section is somewhat technical and may be partially skipped by readers interested in the idea and application results rather than the mathematical details of the model. In the third section, we present comparative results of testing our model and traditional graduation methods on a comprehensive set of demographic data. Discussion follows next, and the Appendix provides an R-code for the proposed method.

## 2. Age heaping model and graduation method

We generalize Ediev's (2003) behavioural age heaping model that explicitly describes which ages are rounded to which ones:

$$N_x = N_x^* + \sum_y \left( k_x p_y^x N_y^* - k_y p_x^y N_x^* \right), \quad (4)$$

here, $N_x$ is the observed population at age $x$, $N_x^*$ is the latent population distribution that would have been observed in the absence of age heaping, $k_x$ is the strength of age heaping at the given round age $x$, and coefficients $p_y^x$ determine the structure of age rounding, so as the product $k_x p_y^x$ gives the proportion of people aged $y$ who report the round age $x$ instead. The set of possible age rounding variants is limited by imposing a structure of non-zero $p_y^x$ coefficients that depend on the type of 'roundness' of age $x$ and the distance between the two ages involved:

$$p_y^x = f_{\{x\}}(x - y), \quad (5)$$

where $\{x\}$ denotes the type of round numbers that the age $x$ belongs to. Imposing the rounding structure (5) restricts possible re-distributions of the population age structure during the graduation process and, thereby, enables avoiding overly smoothed results. In the original model, $\{x\}$ referred to the last digit of the age ('0' or '5'). A more complex typology may apply to 'vigesimal'-type and other unusual heaping cases (see an example further down). Because the $k$- and $p$- parameters enter (4) in a multiplicative form, an additional scaling assumption is needed. In the original model, the scaling assumption was applied to the $k$'s ($k_{45} = k_{50} = 1$). Here, we use another scaling assumption:

$$\sum_y p_y^x = 1 \quad (6)$$

for any round age $x$.

Given the set of $k$- and $p$- parameters, the linear system (4) may be resolved in vector-matrix form to graduate the population age structure:

$$\boldsymbol{N} = (\boldsymbol{E} + \boldsymbol{G})\boldsymbol{N}^* \Rightarrow \boldsymbol{N}^* = (\boldsymbol{E} + \boldsymbol{G})^{-1}\boldsymbol{N}, \quad (7)$$

where $\boldsymbol{G} = \begin{pmatrix} -\sum_y k_y p_0^y & k_0 p_1^0 & k_0 p_2^0 & \cdots & k_0 p_X^0 \\ k_1 p_0^1 & -\sum_y k_y p_1^y & k_1 p_2^1 & & k_1 p_X^1 \\ k_2 p_0^2 & k_2 p_1^2 & -\sum_y k_y p_2^y & & k_2 p_X^2 \\ \vdots & & & \ddots & \vdots \\ k_X p_0^X & k_X p_1^X & k_X p_2^X & \ldots & -\sum_y k_y p_X^y \end{pmatrix}$, $\boldsymbol{N} = \begin{pmatrix} N_0 \\ N_1 \\ N_2 \\ \vdots \\ N_X \end{pmatrix}$, $\boldsymbol{N}^* = \begin{pmatrix} N_0^* \\ N_1^* \\ N_2^* \\ \vdots \\ N_X^* \end{pmatrix}$, $\boldsymbol{E}$ is the identity matrix.

To find age heaping parameters of the model (4)-(7), an optimization procedure is used that maximizes the smoothness of the graduated age structure $\boldsymbol{N}^*$:

$$\boldsymbol{L} = \|\boldsymbol{N}^* - \boldsymbol{A}\| = (\boldsymbol{N}^* - \boldsymbol{A})^T(\boldsymbol{N}^* - \boldsymbol{A}) \rightarrow \boldsymbol{min}, \quad (8)$$

here, $\boldsymbol{A} = (A_0 \quad A_1 \quad A_2 \quad \ldots \quad A_X)^T$ is a benchmark population with a smooth, heaping-free age structure. We use the smoothing spline applied to the original population $\boldsymbol{N}$ as the benchmark $\boldsymbol{A}$ (after some experimentation, the spline's degrees of freedom are set at ten per cent of the number of age categories). Our method may also work with alternative smoothness goal functions not considered here. Note that the smooth benchmark itself is not the desired result of our graduation method, because it might be free of both the heaping signs and the informative variation of the sizes of age groups. The rounding structure (5) imposed over the graduation process will preclude the method from completely converging towards the benchmark distribution.

**Fitting the $k$-parameters given the $p$-parameters**

Typically, optimization (8) implies fitting about 20 $k$-parameters ($k_5, k_{10}, k_{15}, \ldots, k_{95}, k_{100}$) and 16 $p$-parameters ($p_0(\pm1), p_0(\pm2), \ldots, p_0(\pm5)$; $p_5(\pm1), p_5(\pm2), p_5(\pm3)$), 36 parameters altogether. The number of parameters is larger when dealing with more complex heaping structures. Such multidimensionality combined with multiple local minima may be an obstacle in practical applications. Here, we suggest a reduced optimization approach that is considerably faster and more robust than a full-scale optimization. The idea is to split the optimization (7) into two steps. Assume a given set of $p$-parameters (options for which will be discussed further down in the text). Given the $p$-parameters, model (4), (7) turns linear in terms of $k$-parameters. As a further simplification of the model (and improvement of its robustness in practice), we replace the unknown population $N_x^*$ that drives the heaping process in the right-hand side in (4), (7) by a smooth heaping-free proxy $M_x$[2]:

$$N_x = N_x^* + \sum_y \left(k_x p_y^x M_y - k_y p_x^y M_x\right). \quad (9)$$

The proxy population $M_x$ may be refined in iterations where the reconstructed structure $N_x^*$ in each iteration is taken (after being additionally smoothed) as the proxy structure $M_x$ in the next iteration. Here, we apply 20 such iterations, although the method gives good results when $M_x$ is simply set to the smooth benchmark $S_x$ too. Eq. (9) is linear in terms of the $k$-parameters and may be written in a vector-matrix form as:

$$\boldsymbol{N} = \boldsymbol{N}^* + \boldsymbol{\Pi}\boldsymbol{K}, \quad (10)$$

where $\boldsymbol{K} = \left(k_{(1)}, k_{(2)}, \ldots, k_{(l)}\right)^{\boldsymbol{T}}$ is the vector of $k$-parameters for all ages $x_{(1)}, x_{(2)}, \ldots, x_{(l)}$ presumed 'round', and $\boldsymbol{\Pi}$ is a matrix combined of products $p_y^x M_y$:

$$\boldsymbol{\Pi} = \begin{pmatrix} \ddots & & \cdots \\ & \sum_y p_y^{(i(x))} M_y & \\ \vdots & & \ddots \end{pmatrix} - \begin{pmatrix} p_0^{(1)} M_0 & p_0^{(2)} M_0 & \cdots & p_0^{(l)} M_0 \\ p_1^{(1)} M_1 & p_1^{(2)} M_1 & & p_1^{(l)} M_1 \\ \vdots & & \ddots & \end{pmatrix}, \quad (11)$$

here, $p_y^{(i)} \stackrel{\text{def}}{=} p_y^{x_{(i)}}$ is the $p$-parameter that describes rounding the age $y$ into the age $x_{(i)}$; in the first matrix, the non-zero rows are only those that correspond to round ages $x_{(i)}$, and in those rows, the

[2] With this assumption, Eq. (7) is simplified into: $\boldsymbol{N}^* = \boldsymbol{N} - \boldsymbol{GM}$.

non-zero element is only the one that is in the position $i(x)$ of the given round age in vector $\boldsymbol{K}$. Given (10), the optimization problem (8) turns into:

$$\boldsymbol{L} = (\boldsymbol{N} - \boldsymbol{A})^T(\boldsymbol{N} - \boldsymbol{A}) - \boldsymbol{2}(\boldsymbol{N} - \boldsymbol{A})^T\boldsymbol{\Pi}\boldsymbol{K} + \boldsymbol{2}\boldsymbol{K}^T\boldsymbol{\Pi}^T\boldsymbol{\Pi}\boldsymbol{K} \rightarrow \boldsymbol{min}. \quad (12)$$

Differentiating (12) with respect to vector $\boldsymbol{K}$ yields the optimality condition:

$$-\boldsymbol{2}(\boldsymbol{N} - \boldsymbol{A})^T\boldsymbol{\Pi} + \boldsymbol{2}\boldsymbol{K}^T\boldsymbol{\Pi}^T\boldsymbol{\Pi} = 0 \quad (13)$$

that may be resolved using the generalized matrix inverse:

$$\boldsymbol{K} = (\boldsymbol{\Pi}^T\boldsymbol{\Pi})^+\boldsymbol{\Pi}^T(\boldsymbol{N} - \boldsymbol{A}), \quad (14)$$

here, for any matrix $\boldsymbol{C}$, $\boldsymbol{C}^+$ is the (truncated as described next) generalized inverse that may be found from the singular-value-decomposition $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ as $\boldsymbol{C}^+ = \boldsymbol{U}\boldsymbol{\Sigma}^{-\boldsymbol{1}}\boldsymbol{V}^T$. If some of the singular values $\lambda_1, \lambda_2, \dots, \lambda_l$ turn zero (in practical applications, fall below a small threshold that is set 1e-6 here), the corresponding round ages show no heaping and their $k$-parameters should be set zero. To this end, we use truncated inverse with $\boldsymbol{\Sigma^{-1}} = \mathbf{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_t^{-1}, 0, \dots, 0)$, where $t$ is the number of singular values presumed non-zero.

Once the optimal $k$-parameters (14) are determined, the latent (graduated) population is obtained from (10) as:

$$\boldsymbol{N}^* = \boldsymbol{N} - \boldsymbol{\Pi}(\boldsymbol{\Pi}^T\boldsymbol{\Pi})^+\boldsymbol{\Pi}^T(\boldsymbol{N} - \boldsymbol{A}), \quad (15)$$

which concludes age heaping modelling and population graduation given the $p$-parameters.

### Finding the $p$-parameters

For the $p$-parameters, optimization[3] approach may turn less efficient, because the graduation results are less sensitive to the $p$-parameters and multiple local minima may exist. For a specific country-case, the optimization might be valid an option capable of revealing details of age rounding behaviours, although the particular structure of the $p$-parameters may also be set by an expert based on the empirical pattern of population distortions. As a general-purpose algorithm, we suggest a simplified approach with linear $p$-parameters in (5):

$$p_y^x = f_{\{x\}}(x - y) = \begin{cases} \frac{l_{\{x\}} + 1 - abs(x-y)}{l_{\{x\}}(l_{\{x\}}+1)}, abs(x - y) \le l_{\{x\}}, \\ 0, abs(x - y) > l_{\{x\}}, \end{cases} \quad (16)$$

here, $l_{\{x\}}$ is the round digit-specific lag that determines which ages may contribute to the heaping at the given round age (note, that (16) assures the scaling assumption (6))[4].

## 3. Application results

We tested our model by applying it to a vast collection of empirical population profiles, both affected to and free of age heaping. The collection combined: the Human Mortality Database (University of California (Berkeley) and The Max Planck Institute for Demographic Research 2020) and Canadian Human Mortality Database (2020) mostly, but not completely, free of age heaping; the UN database of official population data (United Nations 2020); census data for Russia, former USSR and Russian Empire and their regions (Demoscop Weekly 2021; Center for Demographic Research (Moscow/Russia) 2020) as well as data published by statistical agencies and archive data. Altogether, our collection comprises 33 332 entries, where 8 756 show various levels of age heaping ($K_5$ more than 105) of which 3 551 entries show extreme age heaping ($K_5$ exceeding 150).

---

[3] The $p$-parameters may be optimized either in a full-range optimization (8), or in iterations, alternating optimization of the $p$- and the $k$-parameters (given the latter parameters, model (4) turns linear in terms of $p$-parameters, and a procedure based on generalized matrix inverse similar to the one presented above may be devised).

[4] In empirical tests, we find that exponentially decaying $p$-parameters (at rate 27 percent per year of age-distance) also perform well. Same applies to a more data-driven approach, where the $p$-parameters are set proportional to the mean deviation from the smoothed population structure $1 - \frac{N_y}{M_y}$. These two approaches, however, were marginally less efficient in reducing the Whipple's index as compared to the linear approximation (16).

Three characteristic examples are presented in Figs. 2-4. Figure 2 depicts the original and the graduated population of Alexandropol of the Russian Empire (currently, Armenian city of Gumry) as of Census 1897. The town hosted a sizable military garrison that was responsible for the population peak at ages in the early 20s. At the same time, the largely illiterate local population reported ages with clear signs of heaping. Applying a traditional graduation method (e.g., a smoothing spline as illustrated in the Figure) would either fail to remove the heaping signs or distort the population structure at young ages. On the contrary, our method can both remove the heaping at older ages and preserve the details of the age structure where the presence of the military contignents, not the age misreporting, modifies it.

A more contemporary example (Figure 3) features a population (Census 2010 results for the city of Moscow) with age structure bearing signs of multiple demographic shocks in the past combined with a moderate age heaping (emigration and temporary absence seem to have resulted in massive proxy responding in the Census). Unlike conventional smoothing, our method is efficient in reconstructing the population age structure without apparent distortions.

The last example (Figure 4) presents an unusual case of age heaping that happened in different modes at ages divisible by 5, 10 and 20. The example features Karachay people (Northern Caucasus, Russia) that practised a vigesimal counting system that affected the age rounding in early Censuses and surveys. In this case, we adjust the heaping structure in the model to the observed pattern of population distortions by assuming round ages of four types: ages {10,30,50,70,90} with rounding lag five years; ages {20,40,60,80} with rounding lag ten years; ages {15,25,35,…,95} with rounding lag three years; and additional rounding at all the listed ages with lag one year to account for stronger rounding from adjacent ages apparent in the data. Small population size and extreme age distortions result in a rather rough graduation result that, nonetheless, is effectively cleaned from age heaping (for a smoother result, one may apply additional smoothing, e.g., by a moving average as illustrated in the Figure). Notably, the graduation indicates a substantial population bust at cohorts born in the late 1860s-1970s with a fast recovery and a baby boost in the following cohorts. That was the period when a substantial part of the people was moved to new settlements. That move, naturally, should have resulted in a fertility crisis and births' postponement.

To put the model to a more systematic and comprehensive test, we ran it on our entire dataset and compared it to the two most common graduation methods: the smoothing spline and the moving average. For each of the two alternatives, we tested a range of smoothness parameters (the number of degrees of freedom for the spline and the width of averaged age intervals for the moving average). For our model, we tested a single setup of parameters in all cases: round ages ending with 0's and 5's with rounding lags in (16) 5 and 3 years, respectively. Hence, the method's accuracy may be improved by choosing a better assumption for the parameters when dealing with a specific country case. To compare the methods, we analyzed two criteria characterizing the two opposing aspects of a graduation technique. The first criterion measures the degree of removal of age heaping measured as a mean post-graduation $K_5$ across cases with substantial initial age heaping (cases where the initial $K_5$ exceeded 120). Another criterion measures what distortions to the actual population structure might be inflicted by the graduation method. To this end, we calculate the mean deviation of the graduated population from the original data on entries with no substantial age heaping ($K_5$ in between 90 and 105). The deviation of one age structure from another is calculated as the mean quadratic deviation of age-specific populations as a per cent of the mean size of an age group in the initial population:

$$R(N_x^*, N_x) = 100 \frac{\sqrt{mean((N_x^* - N_x)^2)}}{mean(N_x)}, \quad (17)$$

here, $R(N_x^*, N_x)$ is the deviation of the graduated population structure from the initial one.

Results of the bi-criterial comparison of the graduation methods are presented in Figure 5. Apart from expected results for the traditional methods (an overall trade-off between smoothness and accuracy for the spline and moving average; heaping inversion for the moving average with an

inadequate length of the averaging frame), the Figure shows the advantage of the modelling approach implemented here. Indeed, our model appears the best in both the completeness of removing the heaping and minimizing the distortions to the population's real variation by age.

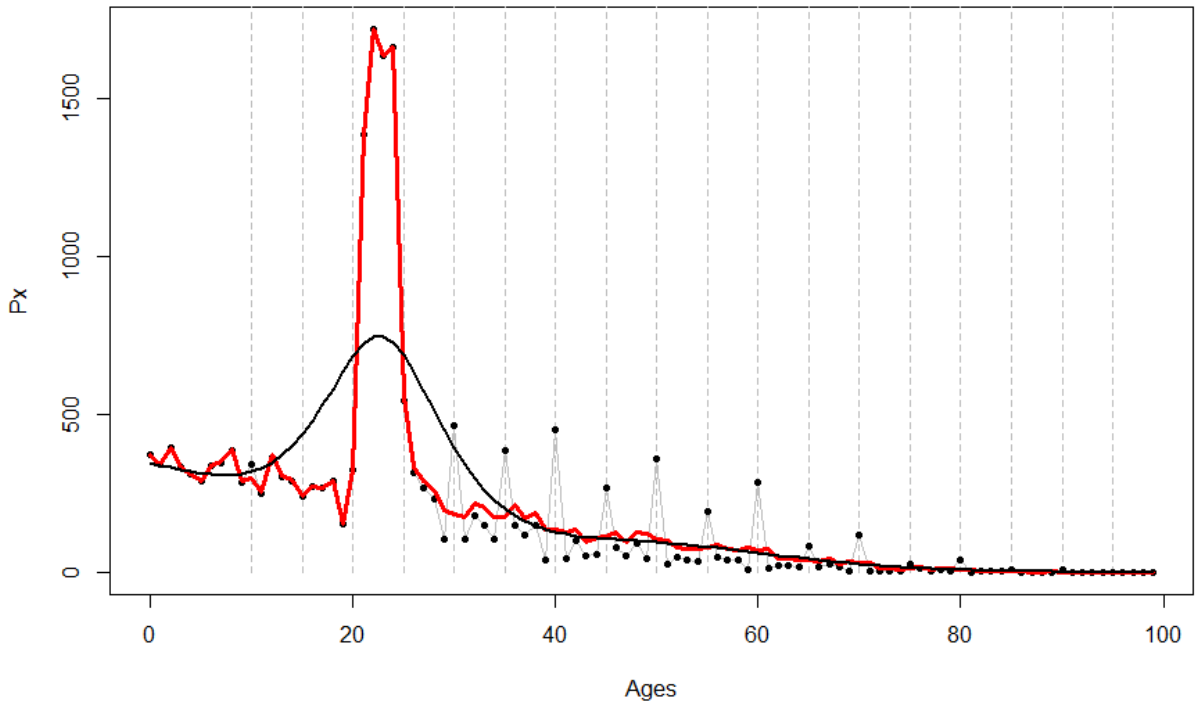


**Fig. 2.** The male population of Alexandropol city of the Russian Empire (currently, Gumry of Armenia) in Census 1897: crude data (grey lines, points), graduated by the proposed model (thick red line), and graduated by smoothing spline (solid black line).

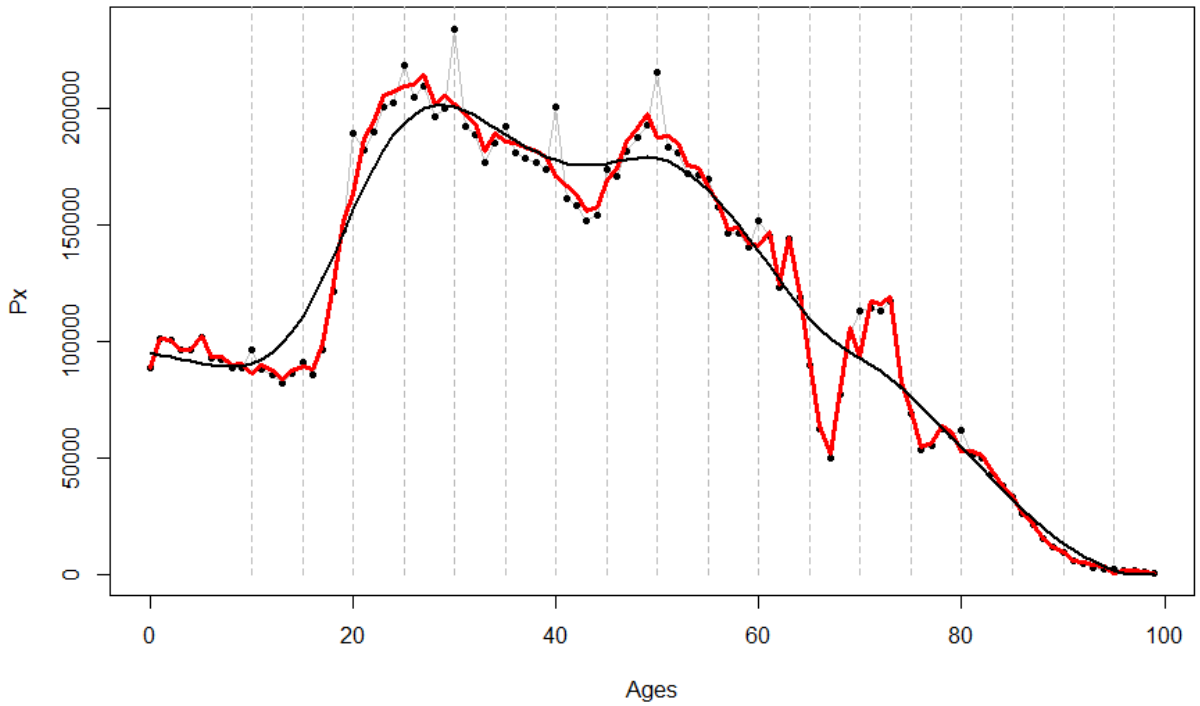


**Fig. 3.** The population of Moscow (Russia) in Census 2010: crude data (grey lines, points), graduated by the proposed model (thick red line), and graduated by the smoothing spline (solid black line).

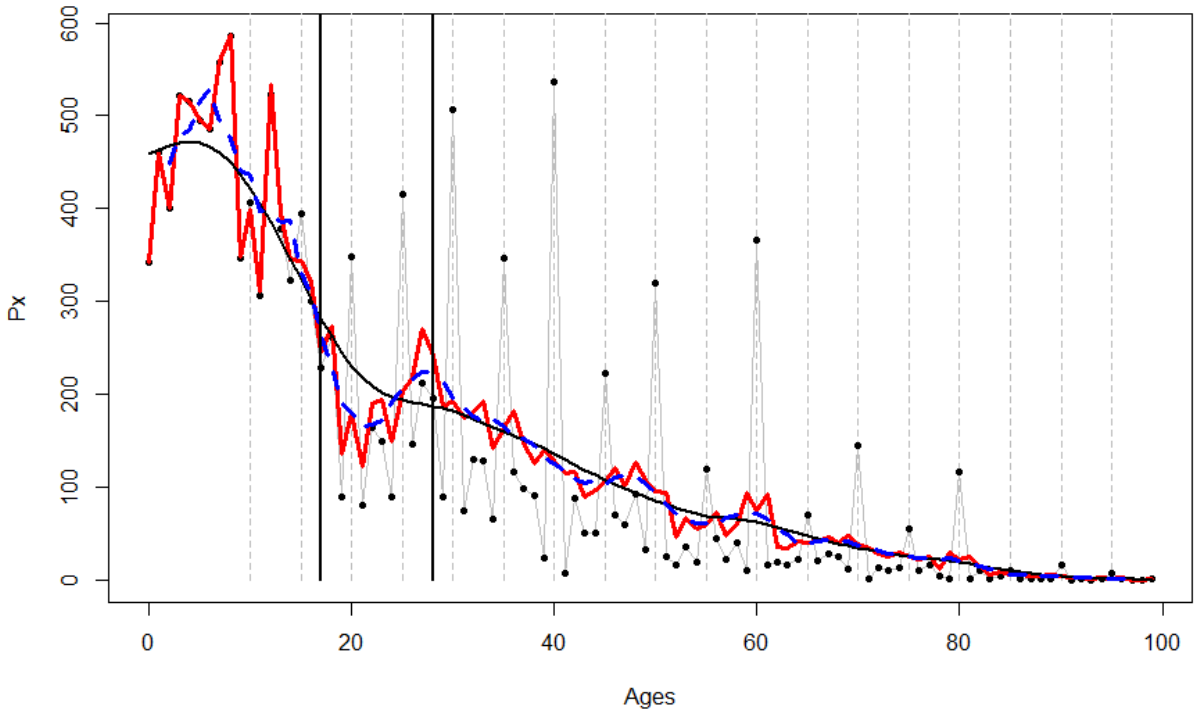


**Fig. 4.** Karachay female population (Northern Caucasus, Russia) in Census 1897: crude data (Ediev 2003) (grey lines, points), graduated by the proposed model (thick red line), additionally smoothed by a moving average over intervals of five years of age (the broken blue line), and graduated by smoothing spline (solid black line). Vertical lines: cohorts born in 1868-1879 (period of massive resettlement following the abolition of serfdom).

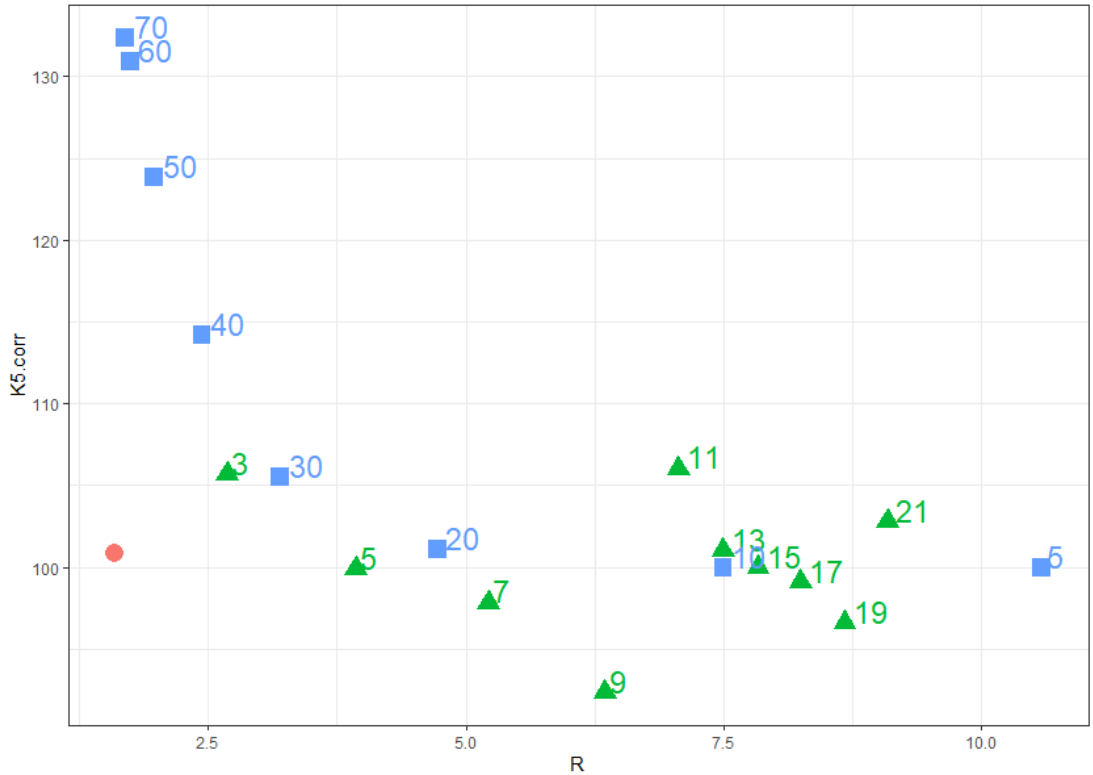


**Fig. 5.** Summary criteria for three graduation methods: the model proposed here (the red circle), the smoothing spline (blue squares with the number of degrees of freedom indicated next to them), and the moving average (the green triangles succeeded by the length of the averaging frame). The criteria: the mean Wipple's index $K_5$ after applying the method, averaged over entries with substantial age

heaping (the vertical axis) and the measure of population structure's distortion (17), per cent, averaged over data with no age heaping (the horizontal axis).

## Concluding remarks

Our strategy of building the graduation method on a foundation of explicitly modelling the mechanism of age rounding turned successful in both removing age heaping signs and minimizing the distortions inflicted to the actual variation of population size by age. The model presented here outperforms both conventional graduation methods, the smoothing spline and the moving average. Its robustness and ability to near completely remove age heaping while preserving the actual population variation suggest its potential for broad practical usage.

Replacing the full optimization (8) in fitting the model parameters by a partial optimization using the generalized matrix inverse (14) produces efficient estimates without compromising the graduation output. That said, a full optimization may also produce useful results in particular country cases. However, it may not be recommended as a general-purpose algorithm because of the multiple local minima problem and the need for deeper expert involvement.

An advantage of our method as compared to the traditional methods is that the strength of age heaping observed in the data (as described by the *k*-parameters) and, therefore, the extent of smoothing needed during graduation are automatically detected by the model. This makes it unnecessary to test the data for the presence of age heaping and selecting the strength of the graduation smoothness before applying the method.

The linear approximation (16) suggested for the $p$-parameters also appears to be efficient in our empirical applications. At the same time, some cases show population patterns inconsistent with (16). Figure 3, for example, might indicate non-symmetric age roundings, while approximation (16) may only produce symmetric patterns of age rounding. Figures 2 and 4 also show a non-linear pattern: ages adjacent to the round ones appear to be rounded at a higher probability than suggested by a linear model. In such cases, one might either turn to the full optimization (5)-(8) or opt for a data-driven set of $p$-parameters (see note 3). Yet, the simple linear approximation (16) may also be efficiently used in these cases with a slight modification. For the non-symmetric patterns of age rounding, one may introduce two identical sets of round ages – one with only roundings from the younger ages (setting $f_{\{x\}}(x-y)=0$ at $x<y$ in (16)), and another with rounding from elder ages only (setting $f_{\{x\}}(x-y)=0$ at $x>y$). For cases when ages adjacent to the round ones show higher probabilities of rounding, one may re-introduce the same round ages with an additional lag of one year (as we did for the example presented in Figure 4) – hence, having two modes of rounding in the model. Conveniently, our procedure allows for duplicating the same round ages several times in the $\boldsymbol{K}$ vector (10), (14). However, in most cases, such modifications would be unnecessary or contributing only marginally to the model performance.

## References


A'Hearn, Brian, Jörg Baten, and Dorothee Crayen. 2009. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital." *Journal of Economic History* 69 (3): 783–808. https://doi.org/10.1017/S0022050709001120.

Bermúdez, Lluís, Dimitris Karlis, and Miguel Santolino. 2017. "A Finite Mixture of Multiple Discrete Distributions for Modelling Heaped Count Data." *Computational Statistics and Data Analysis* 112 (August): 14–23. https://doi.org/10.1016/j.csda.2017.02.013.

Camarda, Carlo G., Paul H.C. Eilers, and Jutta Gampe. 2017. "Modelling Trends in Digit Preference Patterns." *Journal of the Royal Statistical Society. Series C: Applied Statistics* 66 (5): 893–918. https://doi.org/10.1111/rssc.12205.

Canadian Human Mortality Database. 2020. “CHMD Canadian Human Mortality Database.” 2020. http://www.bdlc.umontreal.ca/chmd/.

Center for Demographic Research (Moscow/Russia). 2020. “Russian Fertility and Mortality Database.” 2020. http://demogr.nes.ru/index.php/ru/demogr_indicat/data.

Demoscop Weekly. 2021. “Census Data at Demoscop Weekly (Online Magazine).” 2021. http://www.demoscope.ru/weekly/ssp/rus_age1_10.php.

Ediev, Dalkhat M. 2003. *Demographic Losses of Deported Soviet Peoples [Demograficheskiye Poteri Deportirovannykh Narodov SSSR]*. Stavropol: AGRUS.

———. 2018. “Expectation of Life at Old Age: Revisiting Horiuchi-Coale and Reconciling with Mitra.” *Genus* 74 (1). https://doi.org/10.1186/s41118-018-0029-7.

———. 2021. “On Existence and Uniqueness of Remaining Life Expectancy Estimates in the Model of Stable Population.” *Mathematical Modelling*, no. 5: 1–12.

Groß, Marcus, and Ulrich Rendtel. 2016. “Kernel Density Estimation for Heaped Data.” *Journal of Survey Statistics and Methodology* 4 (3): 339–61. https://doi.org/10.1093/jssam/smw011.

Heller, Gillian Z., and Lindsay C. Dunlop. 2012. “A Modelling Approach for Blood Units Transfused after Stem Cell Transplantation.” *Statistics in Medicine* 31 (28): 3649–55. https://doi.org/10.1002/sim.5415.

Horiuchi, S., and Ansley J. Coale. 1982. “A Simple Equation for Estimating the Expectation of Life at Old Ages.” *Population Studies* 36 (2): 317–26. https://doi.org/10.2307/2174203.

Jowett, A. John, and Yuan Qing Li. 1992. “Age - Heaping: Contrasting Patterns from China.” *GeoJournal* 28 (4): 427–42. https://doi.org/10.1007/BF00273112.

Mitra, S. 1984. “Estimating the Expectation of Life at Older Ages.” *Population Studies* 38 (2): 313–19. https://doi.org/10.2307/2174079.

Myers, R.J. 1940. “Errors and Bias in the Reporting of Ages in Census Data.” *Transactions of the Actuarial Society of America* 41 (II): 395–415.

Pardeshi, Geeta S. 2010. “Age Heaping and Accuracy of Age Data Collected during a Community Survey in the Yavatmal District, Maharashtra.” *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine* 35 (3): 391–95. https://doi.org/10.4103/0970-0218.69256.

Shryock, H. S., and Jacob S. Siegel. 1973. *The Methods and Materials of Demography*. Washington D.C.: United States Bureau of the Census.

Szołtysek, Mikołaj, R. Poniat, and S. Gruber. 2018. “Age Heaping Patterns in Mosaic Data.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51 (1): 13–38. https://doi.org/10.1080/01615440.2017.1393359.

United Nations. 1983. *Manual X: Indirect Techniques for Demographic Estimation. Department of International Economic and Social Affairs Population Studies, No 81*. Vol. 4. United Nations.

———. 2020. “UNdata.” 2020. http://data.un.org/Data.aspx?d=POP&f=tableCode%3A1.

United Nations Statistics Division. 2020. “Demographic Statistics Database. Population by Age, Sex and Urban/Rural Residence.” 2020. http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22.

University of California (Berkeley), and The Max Planck Institute for Demographic Research. 2020.

"Human Mortality Database. Online Database Sponsored by University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)." 2020. www.mortality.org.

Wang, B., and W. Wertelecki. 2013. "Density Estimation for Data with Rounding Errors." *Computational Statistics and Data Analysis* 65 (September): 4–12. https://doi.org/10.1016/j.csda.2012.02.016.

Whittaker, E. T. 1922. "On a New Method of Graduation." *Proceedings of the Edinburgh Mathematical Society* 41: 63–75. https://doi.org/10.1017/s0013091500077853.

Yusuf, Farhat, Jo. M. Martins, David A. Swanson, Farhat Yusuf, Jo. M. Martins, and David A. Swanson. 2014. "Testing the Quality and Smoothing of Demographic Data." In *Methods of Demographic Analysis*, 255–78. Springer Netherlands. https://doi.org/10.1007/978-94-007-6784-3_12.

Zinn, S., and A. Würbach. 2016. "A Statistical Approach to Address the Problem of Heaping in Self-Reported Income Data." *Journal of Applied Statistics* 43 (4): 682–703. https://doi.org/10.1080/02664763.2015.1077372.

## Appendix A. R-code for the model

```
Goptim.unheaping.ediev<-function(Px,RDigits=data.frame(dmin=c(10,15),dmax=99,step=10,lag=c(4,2)),
                        Ps=Ps_RDigits(RDigits),Ages=0:(length(Px)-1),iter=20,smoothPs=FALSE,svdPs=FALSE,
                        df.MA=0.1,iterate.target=FALSE,d.tol=1e-6) {

 #pseudo-inverse matrix
 psInv<-function(R,d.tol=d.tol) {
  SV <- svd(R)
  return(SV$v %*% diag(ifelse(abs(SV$d)<d.tol,0,1/(SV$d))) %*% t(SV$u))
 }

 #Generate P_matrix given p-parameters
 Pmatr_Ps<-function(Ps,RDigits,Ages){
  Pmatr<-matrix(rep(0,length(Ages)^2),nrow=length(Ages))
  pLen<-0
  for(acLevel in 1:(dim(RDigits)[1])){
   rd<-RDigits[acLevel,]
   ds0<-seq(rd$dmin,rd$dmax,rd$step) #incl ages beyond Px
   ds<-ds0[ds0%in%Ages]
   lag<-rd$lag
   lags<-(-lag):lag
   lags<-lags[lags!=0]
   ps0<-Ps[pLen+1:(length(lags))]
   pLen<-pLen+length(lags)
   #Pmatr[Ages %in% 25:55,Ages %in% 30:50]
   for(x in ds){
    ys0<-x+lags
    ys<-ys0[ys0 %in% Ages]
    ps<-ps0[ys0 %in% Ages]
    Pmatr[Ages %in% ys,Ages==x]<-Pmatr[Ages==x,Ages %in% ys]+ps
   }
  }
  if(pLen!=length(Ps)) return(NA)
  return(Pmatr)
 }

 #optimal Ks and uPx given Pmatr (Pi-matrix is built within)
```

```
  deAccPx_Pmatr<-function(Px,Pmatr,M=pmax(0,smooth.spline(x=0:(length(Px)-1), y = Px,
df=round(df.MA*length(Px),0))$y),
                          A=M,alfa=0,df.MA=0.1,d.tol=d.tol){
    PiMatr<-Pmatr*M
    S<-diag(colSums(PiMatr))
    Amatr<-t(S-PiMatr) %*% (S-PiMatr) + alfa * diag(rep(1,length(Px)))
    K<-psInv(Amatr,d.tol=d.tol) %*% t(S-PiMatr) %*% (Px-A)
    K<-pmax(0,K)
    return(data.frame(uPx=Px-(S-PiMatr) %*% K,K=K))
  }
  #optimal Ks and uPx given Ps
  deAccPx_Ps<-function(Px,Ps,RDigits,iter=20,uPx0=Px,M=pmax(0,smooth.spline(x=0:(length(Px)-1), y = uPx0,
df=round(df.MA*length(Px),0))$y),
                       A=M,alfa=0,df.MA=0.1,iterate.target=FALSE,d.tol=d.tol){
    Pmatr<-Pmatr_Ps(Ps,RDigits,Ages)
    roundAges<-NULL
    for(acLevel in 1:(dim(RDigits)[1])){
      rd<-RDigits[acLevel,]
      ds<-seq(rd$dmin,rd$dmax,rd$step)
      roundAges<-c(roundAges,ds)
    }
    for(i in 1:iter){
      deacc<-deAccPx_Pmatr(Px,Pmatr,M,A,alfa,df.MA,d.tol)
      uPx<-deacc$uPx
      M=pmax(0,smooth.spline(x=0:(length(Px)-1), y = uPx, df=round(df.MA*length(uPx),0))$y)  #refine M (accum
driver), AND? A (smooth target)
      if(iterate.target) A=M
    }
    K<-deacc$K
    Ks<-K[roundAges+1] # ages must start at x=0 with step =1
    return(list(Ps=Ps,K=K,Ks=Ks,Px=Px,uPx=uPx))
  }

  #approximate Ps based on RDigits structre and deviations from M
  Ps_PxRDig<-function(Px,RDigits,Ps=NA,M=pmax(0,smooth.spline(x=0:(length(Px)-1), y = Px,
df=round(df.MA*length(Px),0))$y),A=M,
                      smoothPs=FALSE,svdPs=TRUE,df.MA=0.1){
    Lags<-RDigits$lag
    Pslen<-2*sum(Lags)
    if (length(Ps)!=Pslen) Ps<-rep(NA,Pslen)
    cumlags<-c(0,2*cumsum(RDigits$lag))
    cOrder<-order(Lags)
    cRDigits<-RDigits[cOrder,]
    uPx<-Px #initiate iteratively adjusted profile
    #plot(Ages,Px)
    #lines(Ages,M)
    for(i in 1:(dim(cRDigits)[1])){
      orig.row<-cOrder[i]
      pind0<-cumlags[orig.row]+1
      pind1<-cumlags[orig.row+1]
      ps.orig<-Ps[pind0:pind1]
      if (is.na(sum(ps.orig))){
        rd<-cRDigits[i,]
        ds0<-seq(rd$dmin,rd$dmax,rd$step) #may incl ages beyond Px
        ds<-ds0[ds0%in%Ages]
        lag<-rd$lag
        z<-(-lag):lag
        Z<-matrix(rep(ds,length(z)),nrow=length(z),byrow = TRUE)
        Z<-Z+z
        PZ<-matrix(uPx[Z+1],nrow=length(z))
        MZ<-matrix(M[Z+1],nrow=length(z))
        AZ<-matrix(A[Z+1],nrow=length(z))
        PsZ<--(PZ-AZ)/MZ #+1
        PsZ[PsZ<0]<-0 #retain only rounding ages (no 'negative' rounding at non-round ages)
```

```
      PsZ[PsZ>100]<-NA #clean from extreme low Ms

      if(svdPs){ #apply SVD to PsZ matrix - gives ks and Ps simulatneously
        svd.PsZ<-svd(PsZ[,!is.na(colSums(PsZ))])
        ps0<-svd.PsZ$u[,1]/sum(svd.PsZ$u[z!=0,1])
      } else ps0<-rowMeans(PsZ,na.rm = TRUE)
      pssm0<-pmax(0,ps0) # make ps positive, monotone and smooth after lag=1

      if(smoothPs){
        pssm0<-pmax(1e-6,pssm0) # make ps positive, monotone and smooth after lag=1
        if(lag>=3) {
          zzz<-z[z < -1]+2
          ppp<-pssm0[z < -1]
          logreg<-lm(log(ppp)~zzz)
          if (coef(logreg)[2]<0) {
            a<-min(log(pssm0[z == -1]),mean(log(ppp)))
            b<-0
          } else {
            a<-min(log(pssm0[z == -1]),coef(logreg)[1])
            b<-coef(logreg)[2]
          }
          pssm0[z < -1]<-exp(a+b*zzz)

          zzz<-z[z > 1]-2
          ppp<-pssm0[z > 1]
          logreg<-lm(log(ppp)~zzz)
          if (coef(logreg)[2]>0) {
            a<-min(log(pssm0[z == 1]),mean(log(ppp)))
            b<-0
          } else {
            a<-min(log(pssm0[z == 1]),coef(logreg)[1])
            b<-coef(logreg)[2]
          }
          pssm0[z > 1]<-exp(a+b*zzz)
        }else if (lag==2){
          pssm0[z== - 2] = min(pssm0[z== - 2],pssm0[z== - 1])
          pssm0[z== + 2] = min(pssm0[z== + 2],pssm0[z== + 1])
        }
      }
      ps<-pssm0[z!=0]
      Ps[pind0:pind1]<-ps/sum(ps) #scale Ps automatically
    }
    cPs<-ifelse(is.na(Ps),0,Ps)
    deacc<-deAccPx_Ps(uPx,cPs,RDigits)
    uPx<-deacc$uPx
    M=pmax(0,smooth.spline(x=0:(length(Px)-1), y = uPx, df=round(df.MA*length(uPx),0))$y)
    A=M
   }
   return(Ps)
 }

 Ps<-Ps_PxRDig(Px,RDigits,Ps=Ps,smoothPs = smoothPs, svdPs=svdPs,df.MA=df.MA) #only Ps==NA will be
estimated
 deacc<-deAccPx_Ps(Px,Ps,RDigits,iter,df.MA=df.MA,iterate.target=iterate.target,d.tol=d.tol)
 Ks<-deacc$Ks
 uPx<-deacc$uPx
 return(list(uPx=uPx,Ps=Ps,Ks=Ks))
}

#approximate Ps based on RDigits structure and simple models/empirics
Ps_RDigits<-
function(RDigits=data.frame(dmin=c(10,15),dmax=99,step=10,lag=c(4,2)),model=c("Linear","Exponential","Rectangula
r","Empirical")[1],q=0.73) {
 Ps<-NULL
```

```
  for(i in 1:(dim(RDigits)[1])){
    rd<-RDigits[i,]
    dmin<-rd$dmin
    smallerdigits<-dmin-(0:10)*rd$step
    digit<-min(smallerdigits[smallerdigits>0])

    lag<-rd$lag
    z<-(-lag):lag
    z<-z[z!=0]
    if (model=="Linear") ps<-(lag+1-abs(z))/(lag*(lag+1)) else {
      if (model=="Exponential") ps<-0.5*(1-q)/(1-q^lag)*q^(abs(z)-1) else {
        if (model %in% c("Rectangular","Uniform","Fixed","Constant")) ps<-rep(1,length(z))/(2*lag) else return(NA)
      }
    }
    Ps<-c(Ps,ps)
  }
  return(Ps)
}
```