Mathematical and statistical modelling of rounded data in historical populations

Eva Kačerová, Jiří Henzler

Abstract: One of the most cited sources of demographical data on the middle of the 17th century in Bohemia is the List of Inhabitants according to Faith of 1651. The List yields extensive data on serfs in several areas in Bohemia: on their age, profession, religion and marital status. Helas, data concerning the age of listed persons are in general rounded according to various rules. The aim of this article is to cope with this problem by means of mathematical modelling of age distribution and statistical analysis via moving averages[1].

Key words: Historical demography, Chrudim area, age structure, mathematics and statistics in demography

1. Introduction

A basic source for the study of the population structure according to the age and genders in the middle of the 17th century is the List of serfs according to faith of 1651 (Pazderová, 2002), whose origin has a connection with recatolisation efforts after the Thirty Years War. Immediate cause of the elaboration of the List was the edict of the vice-governors in Bohemia, made public in 4th of February, 1651. It stated shortcomings of the existing lists of non–catholic serfs and ordered to all suzerains to list, according to the enclosed form, all their serfs of both sexes, and to expedite it in the sex weeks term to the vice – governors office.

2. Chrudim area population

According to the List, 46 626 people lived in the Chrudim area, including 24 860 women. 2 104 persons missed the age record. Children of less than 12 years of age (age of the first confession) were listed rarely. The listing of these children was not complete in any dominion of the Chrudim area, so the children under 12 are not presented in the following analysis.

For the demographic structure study, the age record is the most important. The exact age of the listed persons played rather a secondary role in the aims of the originators of the List and therefore it is often rounded or even missed. Scribes tended to round the age of those, who did not know their age exactly, to multiples of 10 or 5. Even numbers were in it more popular then the odd ones. Rounding of age was dependent on the social categories of the serfs, too. Widows were often ascribed the age of 40, or 60 if older. „Alone women“ with children, i.e. probably unmarried mothers were frequently ascribed the age of 30, and so on. The measure of distortion can be measured by the index of cumulative age *ik*:

$$ik^p = (5 * \sum_{0}^{7} S^p_{25+5x}) / (\sum_{23}^{)62} S^p_x), \qquad (1)$$

where S_x represents the number of persons in a given age group and the upper index p describes sex.

Nevertheless, despite obvious inaccuracies and rounding the age data from the List represent valuable information of the population at that time (Fig. 1).

[1] This article came into being within the framework of the long-term research project 2D06026, "Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

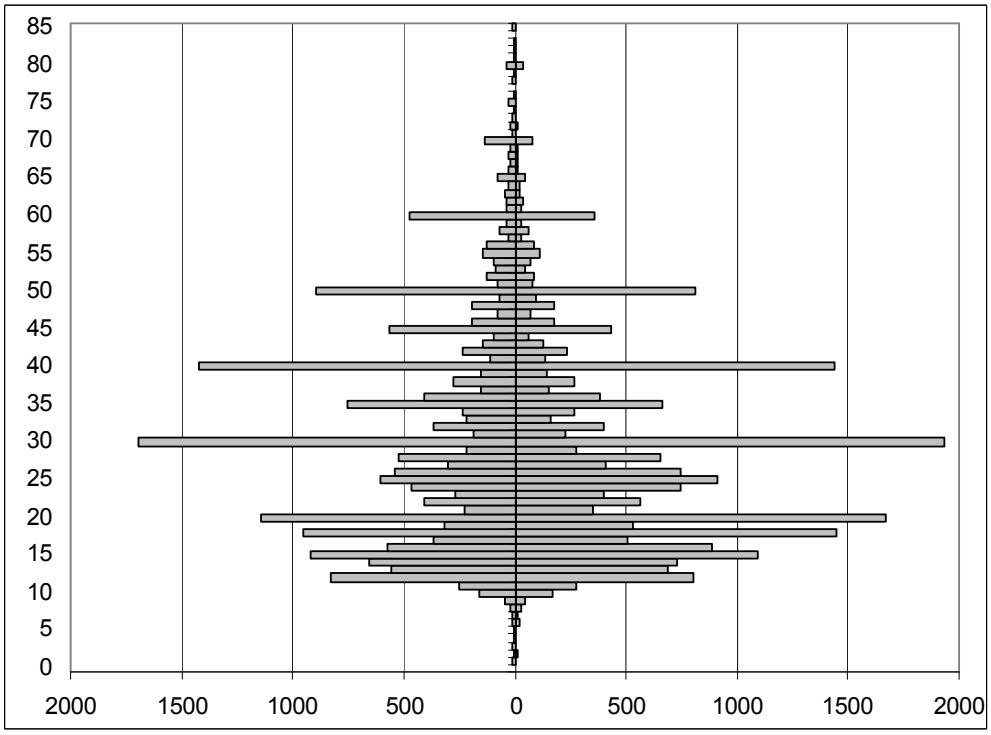

Figure 1: Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651

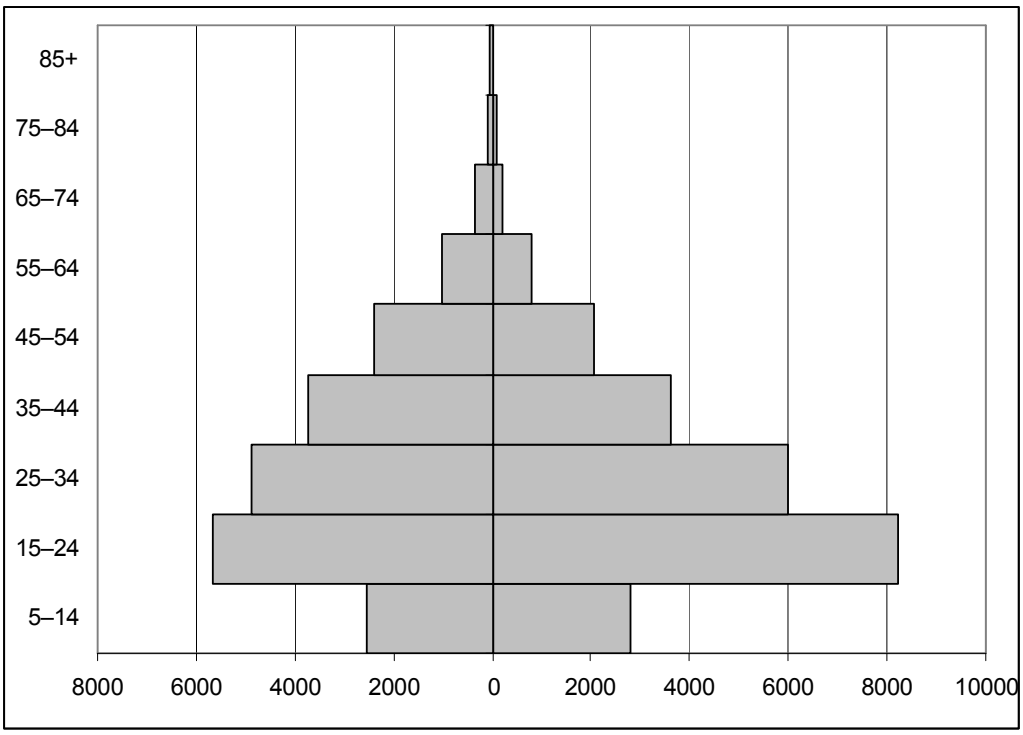

Figure 2: Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651 (decennial intervals)

Rounding of age leads to distorted results concerning the population age structure. This distortion could be diminished by applying decennial age intervals, ion which the most frequented value will be always in the middle, i.e. intervals 5–14, 15–24, etc. (Fig. 2). For the sake of comparability with the works of other authors dealing with the mentioned List. The commonly used age intervals 0–4, 5–9, 10–14, etc. (Fig. 3) are also given. In some studies,

decennial intervals 0–9, 10–19, etc. could be found. The choosing of type of age intervals could influence the weight of rounding in resulting age structure. In any case, it leads to the loss of information.

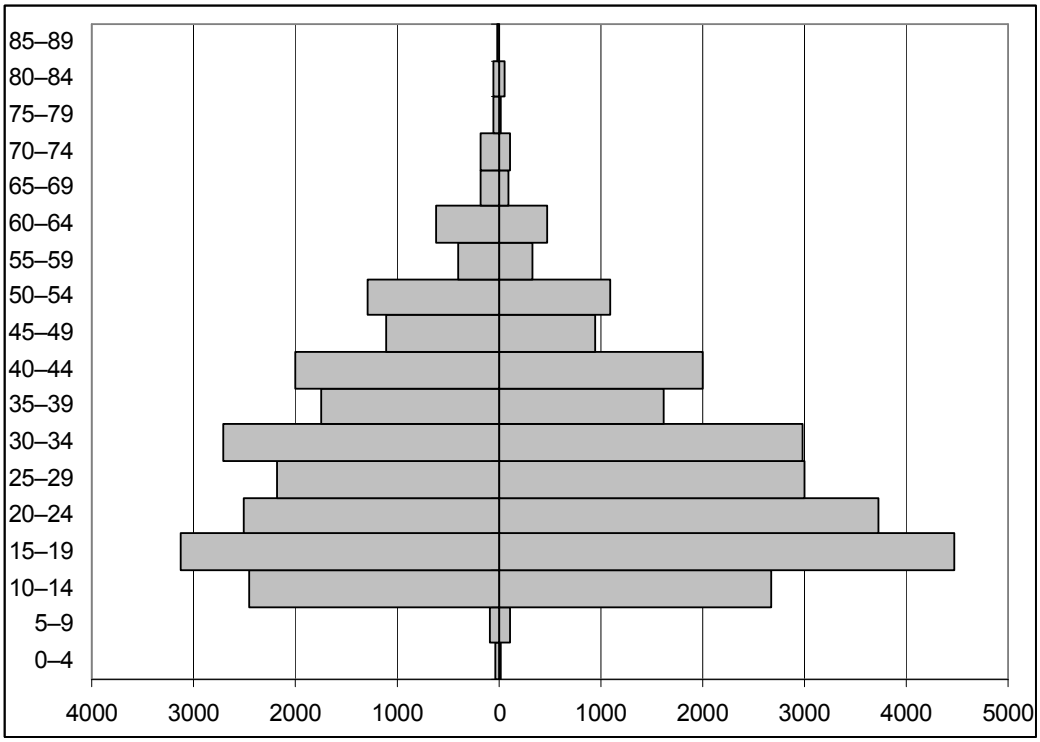

Figure 3: Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651 (five-year intervals)

3. Moving averages

Smoothing of the data using moving averages is a standard statistical technique. As the data were the most frequently rounded to the whole tens, the most suitable moving averages are those of ten subsequent values. In the mentioned List, age under 12 is significantly undervalued for two reasons: children under 12 were not included in the List or their data were included into the category of the confession age of 12 years. Therefore, the data concerning the age under 12 were not included into the processing. For avoiding distortion of the "smoothed" age pyramid in subsequent age categories by absence of the data under 12, are the moving averages corresponding to the years following the age of 12 computed on the base of shorter time intervals: for the age of 12 the original data are taken, the data for the age of 13 are smoothed by the average of three values for age 12, 13 and 14, the data for the age of 14 are smoothed by the average of five values for age 12, 13, 14, 15 and 16, the data for the age of 15 are smoothed by the average of seven values for age 12, 13, 14, 15, 16, 17 and 18, and finally the data for the age of 16 are smoothed by the average of nine values for age 12, 13, 14, 15, 16, 17, 18, 19 and 20. From the age of 17 onwards the data for a given age are smoothed by the moving average of ten subsequent values. As to the even number of the averaged values, the data of the age K, $K \geq 17$, will be smoothed by the arithmetic average of the two subsequent moving averages:

$$\frac{1}{2}\left(\frac{1}{10}\sum_{i=k-5}^{k+4} d_i + \frac{1}{10}\sum_{i=k-4}^{k+5} d_i\right),$$

where d_i is the original value from the List concerning the age i.

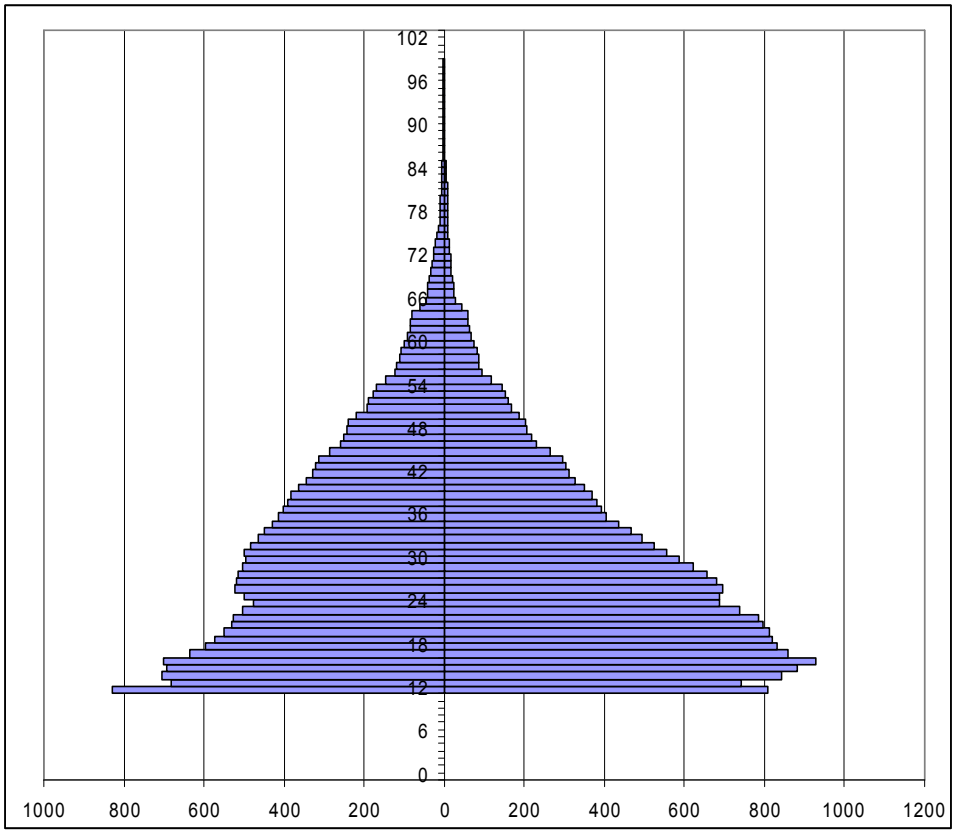

Figure 4: Moving averages

4. Mathematical models

Another option to smooth the data is to create simple mathematical model supposing that the rounding exerts some regularity. If the theoretical values computed on the base of such a model will not be systematically biased, a hypothesis on the possible way of rounding can be formulated.

In this paper, two mathematical models are presented. In both models we suppose that rounding to whole tens concerns ages not more than 4 years up and not more than 4 years down the rounded value and rounding to values ending by 5 concerns ages not more than 2 years up and not more than 2 years down the rounded value. Rounding to even numbers concerns neighbouring odd numbers.

In both models an interesting phenomenon was taken into account: From the original age structure (Fig. 1) one can easily see that from the age of 24 are the frequencies concerning the ages ending by 6 significantly bigger (somewhere more than two times bigger) than the frequencies concerning the ages ending by 4. The explanation of this feature may be as follows:

1. rounding the age of 23, of 33 etc. took place directly up to 25, 35 etc., not to neighbouring even numbers 24, 34 etc.,
2. rounding the age of 25, 35 etc. took place up to 26, 36 etc. and not down to 24, 34 etc.

To estimate the values h_{10}, h_5, h_3 that in original frequencies of ages ending by 0 or 5 and of even ages could be ascribed to rounding, and by using those values to estimate theoretical values Y_i from the original frequencies y_i, in both models we supposed that

neighbouring theoretical values Y_{28}, Y_{29}, Y_{30}, Y_{31}, Y_{32} differ by the same value Δ; the same supposition holds for values Y_{33}, Y_{34}, Y_{35}, Y_{36}, Y_{37}, etc.

Data concerning the age under 12 were from the same reasons as in the preceding chapter excluded from our models.

5. Uniform model

Uniform model consists in supposition that numbers of cases rounded to age ended by 0 and 5 and to even age are for all neighbouring ages the same. E.g. the same number of $h_{10}(30)$ persons having a real age of 26, 27, 28, 29, 31, 32, 33, 34 was rounded to the age of 30, the same number of $h_5(35)$ persons having a real age of 33,34,37 was rounded to the age of 35 (as to the age of 36 see the 3rd paragraph of the chapter 4), etc.

From the equations

$$y_{28} = Y_{30} + 2\Delta - h_{10} + 2h_3(28) \quad (2)$$

$$y_{29} = Y_{30} + \Delta - h_{10} - h_3(28) \quad (3)$$

$$y_{30} = Y_{30} + 8h_{10} \quad (4)$$

$$y_{31} = Y_{30} - \Delta - h_{10} - h_3(32) \quad (5)$$

$$y_{32} = Y_{30} - 2\Delta - h_{10} + 2h_3(32) \quad (6)$$

we can compute unknowns $Y_{30}, h_{10}, h_3(28), h_3(32), \Delta$ and the remaining values Y_{28}, Y_{29}, Y_{31}, Y_{32} and same is valid in the similar equations for original data $y_{38}, y_{39}, y_{40}, y_{41}, y_{42}$, etc.

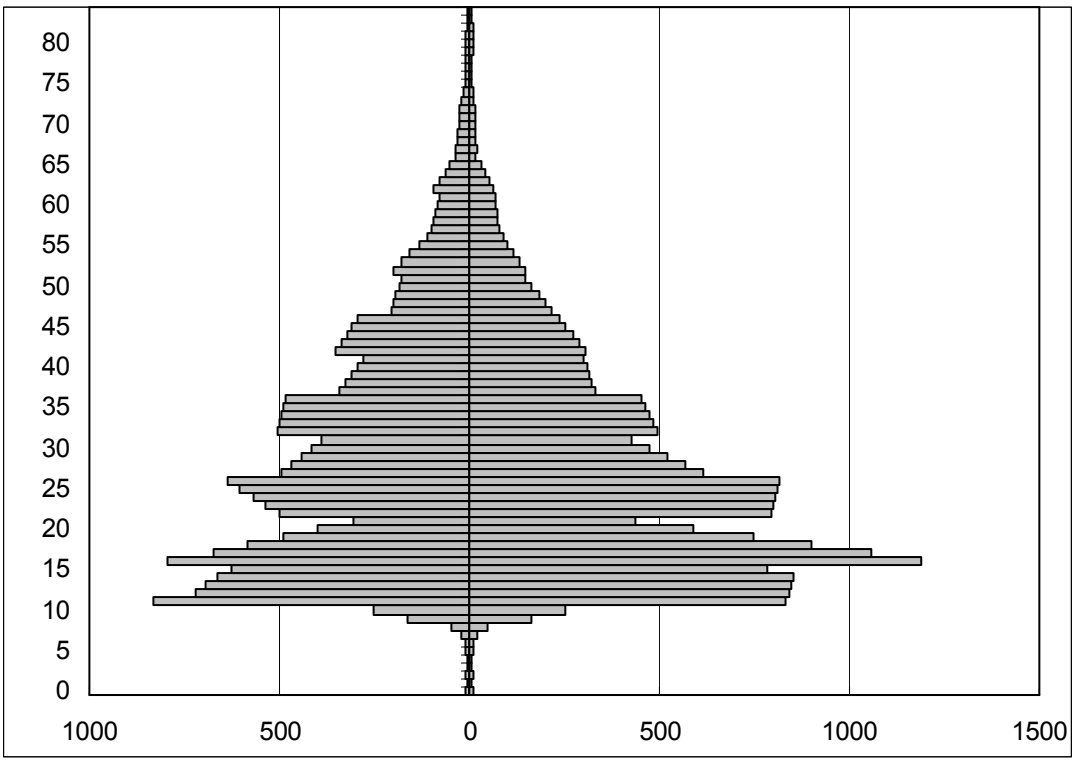

Figure 5: Uniform model

In a similar way we can compute the system of equations

$$y_{33} = Y_{35} + 2\Delta - h_{10}(30) - h_5 - h_3(34) - h_3(32) \quad (7)$$

$$y_{34} = Y_{35} + \Delta - h_{10}(30) - h_5 + h_3(34) \quad (8)$$

$$y_{35} = Y_{35} + 4h_5 - h_3(36) \quad (9)$$
$$y_{36} = Y_{35} - \Delta - h_{10}(40) - h_5 + 2h_3(36) \quad (10)$$
$$y_{37} = Y_{35} - 2\Delta - h_{10}(40) - h_5 - h_3(36) - h_3(38) \quad (11)$$

and similar systems of equations for $y_{43}, y_{44}, y_{45}, y_{46}, y_{47}$ etc.

6. Linear model

In linear model we suppose that that numbers of cases rounded to age ended by 0 and 5 and to even age are the bigger, the closer is rounded age to the age, to which the rounding takes place. E.g. by rounding to an age of 30 a number of $h_{10}(30)$ persons of real age of 26 and 34 would by rounded but dual number $2h_{10}(30)$ of persons of real age of 27 and 33, triple number $3h_{10}(30)$ of persons of real age of 28 and 32, and quadruple number $4h_{10}(30)$ of persons of real age of 29 and 31, and similarly by rounding to an age of 35. In the latter case, the effect of phenomenon 34, 36 will take place – see the 3rd paragraph of the chapter 4, etc.

Similarly as in the uniform model we solve the following system of equations

$$y_{28} = Y_{30} + 2\Delta - 3h_{10} + 2h_3(28) \quad (12)$$
$$y_{29} = Y_{30} + \Delta - 4h_{10} - h_3(28) \quad (13)$$
$$y_{30} = Y_{30} + 20h_{10} \quad (14)$$
$$y_{31} = Y_{30} - \Delta - 4h_{10} - h_3(32) \quad (15)$$
$$y_{32} = Y_{30} - 2\Delta - 3h_{10} + 2h_3(32) \quad (16)$$

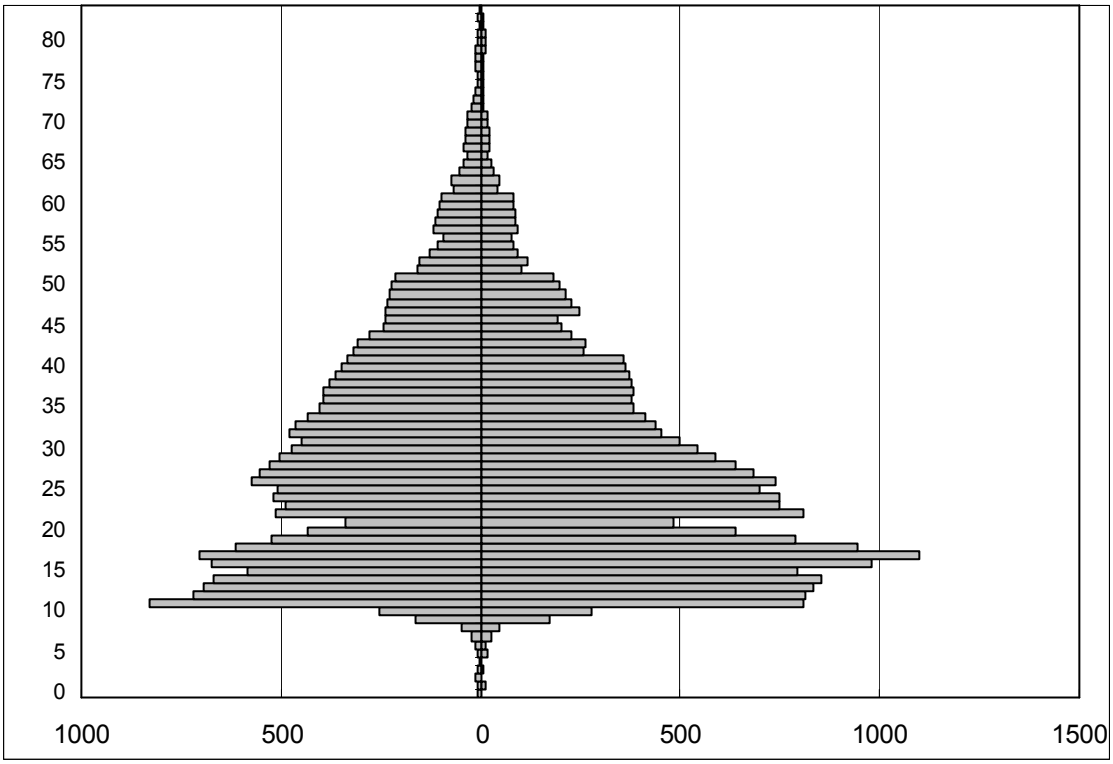

Figure 6: Linear model

and the system of equations

$$y_{33} = Y_{35} + 2\Delta - 2h_{10}(30) - h_5 - h_3(34) - h_3(32) \quad (17)$$

$$y_{34} = Y_{35} + \Delta - h_{10}(30) - 2h_5 + h_3(34) \quad (18)$$

$$y_{35} = Y_{35} + 4h_5 - h_3(36) \quad (19)$$

$$y_{36} = Y_{35} - \Delta - h_{10}(40) + 2h_3(36) \quad (20)$$

$$y_{37} = Y_{35} - 2\Delta - 2h_{10}(40) - h_5 - h_3(36) - h_3(38) \quad (21)$$

7. Conclusions

From the theoretical bihistograms (Fig. 5 and 6) there is obvious that for the studied real population the linear model exerts smaller systematic deviation (periodicity of theoretical values) than the uniform model. On the other hand, the uniform model seems to be better in assessing the female part of population. Bihistogram constructed on the basis of moving averages yields global view on the population only. The crucial difference between the statistical approach based on the moving averages and mathematical approach using uniform or linear model consists in the following fact. Whereas mathematical models bring into the data hypotheses based partly on the present knowledge on construction such 17th century lists, the statistical approach is based on existing data only.

8. References

[1] Ducháček, K. – Fialová, L. – Horská, P. – Répásová, M. – Sládek, M.: On using the 1661–1839 lists of subject of the Třeboň dominion to study the age structure of the population, in: HD 13, 1989, s. 59–75.

[2] Fialová, L. – Kučera, M. – Maur, E. – Horská, P. – Musil, J. – Stloukal, M.: Dějiny obyvatelstva českých zemí [History of the population of the Czech Lands], Praha, 1998.

[3] Henzler, J. a kol.: Matematika pro ekonomy [Mathematics for economists], Oeconomica, Praha, 2007.

[4] Kačerová, E.: Struktura obyvatelstva Choceňska v roce 1651[Population structure of the Chrudim area], Demografie [online], 2008, s. 1–4, http://www.demografie.

[5] Kačerová, E., Henzler, J.: Matematické modelování věkových struktur historických populací [Mathematical models of age structure of historical population], *Forum Statisticum Slovacum*, 2009, roč. 5, č. 1, s. 25–30.

[5] Maur, E.: Problémy demografické struktury Čech v polovině 17. století [Problems of the demographic structure of Bohemia in the middle 17th century], in: ČsČH XIX, 1971, s. 849–850.

[6] Pazderová, A.: List of serfs according to faith – The Chrudim area, Národní archív, 2002.

[7] Žváček, J. – Henzler, J.: Statistika pro ekonomy, interaktivní text na CD [Statistics for economists, interactive text on CD], MUP, Praha, 2009.

Address of the authors:

Eva Kačerová
Department of Demography FIS VŠE
3, W. Churchill
130 67 Praha 3
kacerova@vse.cz

Jiří Henzler
Department of Mathematics FIS VŠE
3, W. Churchill
130 67 Praha 3
henzler@vse.cz