A method for calculating the strength and shape of age heaping

Onno Boonstra, afdeling Geschiedenis, Radboud University Nijmegen

Abstract

In their pioneering work on the Florence Catasto of 1427, Herlihy and Klapish-Zuber (1977), were astonished to see that the ages of the Florentine population were distributed very unevenly. There were much more people with an even numbered age than with an uneven numbered age; there was also distinct age heaping at multiples of 5, 10 and even 12.
Herlihy and Klapish-Zuber did not use statistics to prove the strength and shape of age heaping. More recent research does make use of a few statistical tools, but they are not very refined. A number of indices is being used, for instance Whipple's index, which is meant to detect age heaping at multiples of 5. However, it does not take into consideration age heaping at other multiples. For other multiples, there are other indices who do.
All these indices share a number of problems. First of all, although they do make clear that age heaping exists, they do not measure to what extent age heaping occurs. There are a few rules of the thumb, but these have not been validated statistically. And these rules cannot be compared to one another, so that it cannot become clear which form of age heaping is predominant. Next to that, the fact that there is always a kind of interaction between forms of age heaping (for instance, the heaping at age 60 can be the result of age heaping at multiples of 2, 3, 4, 5, 6, 10, 12, 20 or 30) is not taken into consideration.
A further problem with methods like Whipple's index is that it cannot be used in surveys or with sources where the ages of respondents are limited, like marriage registers.
In my paper, I will introduce a new, simple and straightforward method which (1) calculates the strength of age heaping statistically, (2) while measuring all possible kinds of shapes of age heaping, (3) making the results comparable. Finally, it is a method which (4) can be used in samples with a varying age distribution.

Age heaping and the need for a proper statistic

Age heaping is a phenomenon which occurs when some of the people who are interviewed during a survey, and asked to state their correct age, fail to do so. Instead, they choose a number of preference, based on a cognitive process called the "prototype recall process".[1] When the ages that have been mentioned in the survey are plotted in a graph, the graph does not show the smooth age distribution one normally would expect, but a spiked one, which sometimes is called a "hedgehog" or a "porcupine" distribution.[2] There are researchers who argue that the degree to which age heaping occurs within a society gives us a clue about the "innumeracy" of such a society. According to them, an awareness of age is a good proxy for numeracy. As a consequence, age heaping, which reflects the unawareness of age, can be used as a proxy for innumeracy. Some researchers even go further. They link age heaping even to illiteracy, because innnmercay is thought to be a proxy to illiteracy.[3] It is hard to find historians who support such claims, but the argument may gain weight if

[1] J.J. Vaske and J. Beaman, `Lessons learned in detecting and correcting response heaping: conceptual, methodological and empirical observations', *Human Dimensions of Wildlife,* 11 (2006), 285-296
[2] L.A. Clarkson, 'The Demography of Carrick-on-Suir, 1799', *Proceedings of the Royal Irish Academy. Section C: Archaeology, Celtic Studies, History, Linguistics, Literature*, 87C (1987), 13-36.
[3] Brian A'Hearn, Joerg Baten and Dorothee Crayen, 'Quantifying quantitative literacy, age heaping and the history of human capital', *The journal of economic history*, 69, 3 (2009), x-x, 6.

there was a statistic that can measure age heaping properly. At the moment, there isn't. A number of indices is being used, but they all have their faults and flaws. Take for instance the index that is used most often, Whipple's index. This index is meant to detect "digit" age heaping, i.e. age heaping at multiples of 5 and 10. Just like Bachi's index[4] and Beaman's index[5], it does not take into consideration age heaping at other multiples. Therefore, other indices have been created for other multiples, like the "even index" which measures heaping at multiples of 2, or the "disciple index", measuring heaping at multiples of 12.[6] All indices mentioned above share the problem that they do not take into account that an age distribution is shaped and structured by mortality. Meyer's index has been introduced to take this effect into account.[7] Next to that, a series of "overall" indices has been developed which try to combine the various indices into a single index, like the digit-specific modified Whipple's index[8], the modified total Whipple's index[9], Myers' Blended index[10], and so on.

However, all these indices share a number of problems. First of all, although they do make clear that age heaping exists, they do not measure properly to what extent age heaping occurs. A few rules of the thumb have been formulated, but these have not been validated statistically. Secondly, the rules set for the various indices cannot be compared to one another, so that it cannot become clear which form of age heaping is predominant. Next to that, the fact that often heaping formats are piled (for instance, the heaping at age 60 can be the result of age heaping at multiples of 2, 3, 4, 5, 6, 10, 12, 20 and/or 30) is not taken into account.
A further problem with methods like Whipple's index is that the index is set up exclusively for ages between 23 and 62 inclusive, so it cannot be used properly in surveys or with sources where the age boundaries of respondents are smaller, larger or different, like for instance in the various population registers that so often are at the basis for historical research.

Therefore, there is a need for a straightforward method which (1) calculates the strength of age heaping statistically, (2) while measuring all possible kinds of age heaping formats, (3) making the results comparable. Finally, it should be a method which (4) can be used in data sets with various age boundaries. Such a method already has been created by Camarda et al[11], but in this paper a much simpler method is presented. The method will be described with the help of one of the most famous historical surveys that do show marked forms of age heaping, the 1427 Catasto of Florence.

Age heaping and the 1427 Catasto of Florence

On 24 May 1427, the Priors of the Republic of Florence decreed a tax survey for all citizens of the city of Florence and the inhabitants of the surrounding Florentine Contado and Distretto region. All

[4] R. Bachi,`The tendency to round off age returns: measurement and corrections'. *Bulletin of the International Statistical Institute, Proceedings of the 27th Session, Calcutta*, 33, 4 (1951), 195-222.
[5] J. Beaman et al., `Individual versus aggregate measures of digit preference', Human dimensions of wildlife, 2, 1 (1997), 71-80.
[6] T. De Moor en J.L. van Zanden, 'Van fouten kan je leren. Een kritische benadering van de mogelijkheden van 'leeftijdstapelen' voor sociaal-economisch onderzoek naar gecijferdheid in het pre-industriële Vlaanderen en Nederland', *Tijdschrift voor Sociale en Economische Geschiedenis*, 5, 4 (2008), 55-86.
[7] R.J. Myers, `Errors and biases in the reporting of ages in census data. *Transactions of the Acturial Society of, America,* 41 (1940), 395-415.
[8] Noumbissi A. 1992. L'indice de Whipple modifié : une application aux données du Cameroun, de la Suède et de la Belgique', *Population* 47, 4 (1992), 1038-1041.
[9] Spoorenberg Thomas , `Quality of age reporting : extension and application of the modified Whipple's index', *Populatino*, 62, 4 (2007), 729-741.
[10] Henry S. Shryock and Jacob S. Siegel, *Methods and Materials of Demography*. New York 1976: Academic Press.
[11] C.G. Camarda, P.H.C. Eilers and J. Gampe, `Modelling general patterns of digit preference', *Statistical Modelling,* 8, 4 (2008), 385-401.

family heads were interviewed about their property, their business, their income and the members of the household they headed. Within a few months, the survey was completed. For the city of Florence alone, 16,330 heads of the family had been interviewed, giving information about a total of 259,295 persons. In their pioneering statistical analysis on the Catasto, Herlihy and Klapish-Zuber reported that, according to the survey, the ages of the Florentine population were distributed very unevenly.[12] Figure 1 shows the age distribution for all men and women on the basis of the dataset Herlihy and Klapish-Zuber created of the Catasto.[13] It is obvious that this is not a smooth age distribution; on the contrary, even the word "hedgehog distribution" falls short. A very large proportion of the almost 300 thousand Florentines who were surveyed in the Catasto must have given a wrong age. For instance, a mere 258 Florentines were listed as being 39 years of age, whereas the number of people that were listed with an age of 40 was almost 50 times higher: 11,200!

Figure 1. Age distribution of the population of Florence, according to the Catasto, 1427. N=259,295.

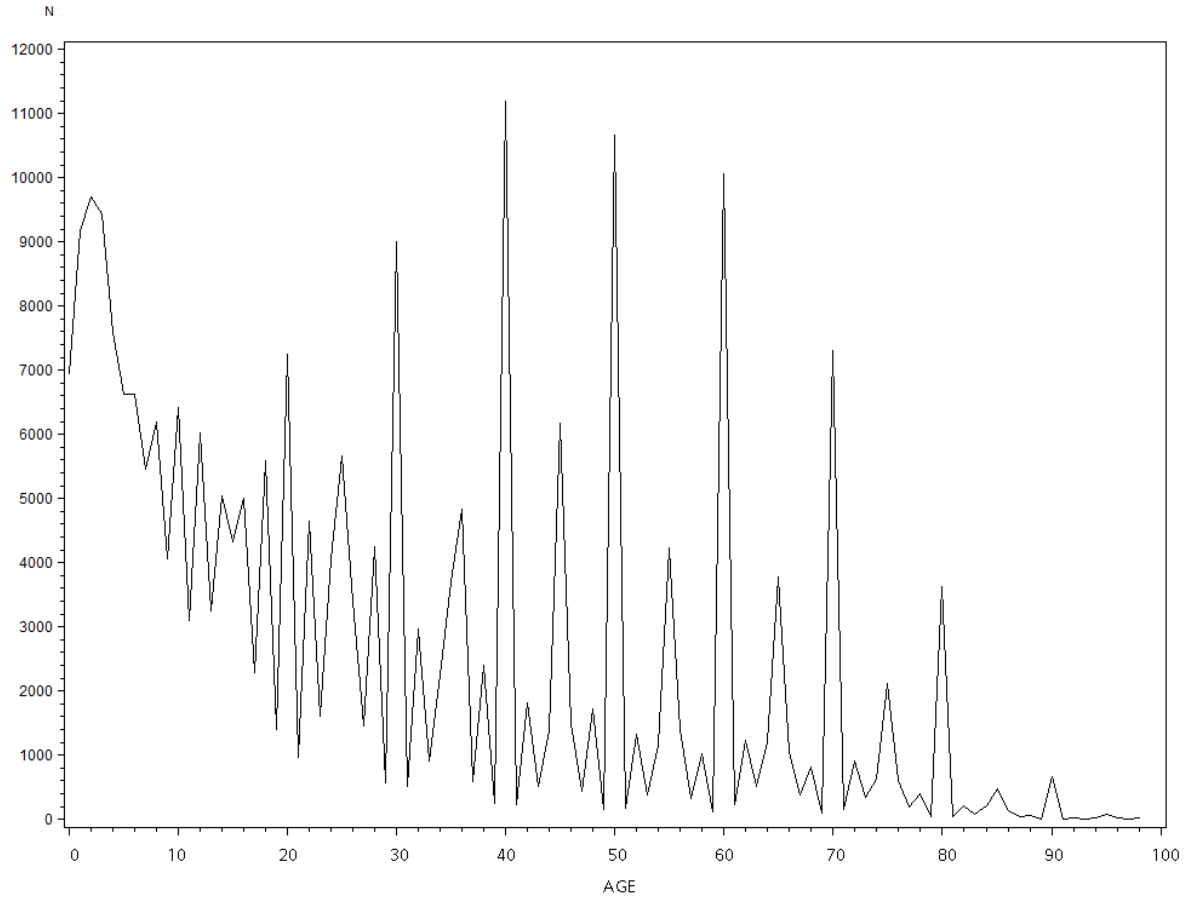


Figure 1 clearly gives us a clue about what age heaping formats were prevalent within the Florintine Catasto: there is immense age heaping at ages which are multiples of 10 (20, 30, 40 etc), at multiples of 5 (20, 25, 30, 35, etc) and also at multiples of 2 (20, 22, 24, 26, 28, 30, etc.). Klapish-Zuber and

[12] David Herlihy et Christiane Klapisch-Zuber, *Les Toscans et leurs familles: une étude du Catasto Florentin de 1427*. Paris 1978: Presses de la Fondation Nationale des Sciences Politiques, x-x.
[13] A digital version is available at http://www.stg.brown.edu/projects/catasto/

Herlihy also mention age heaping at multiples of 12, but apart from a spike at age 36, Figure 1 does not hint to such a heaping format .[14]

A statistical method for discovering age heaping formats within an age distribution

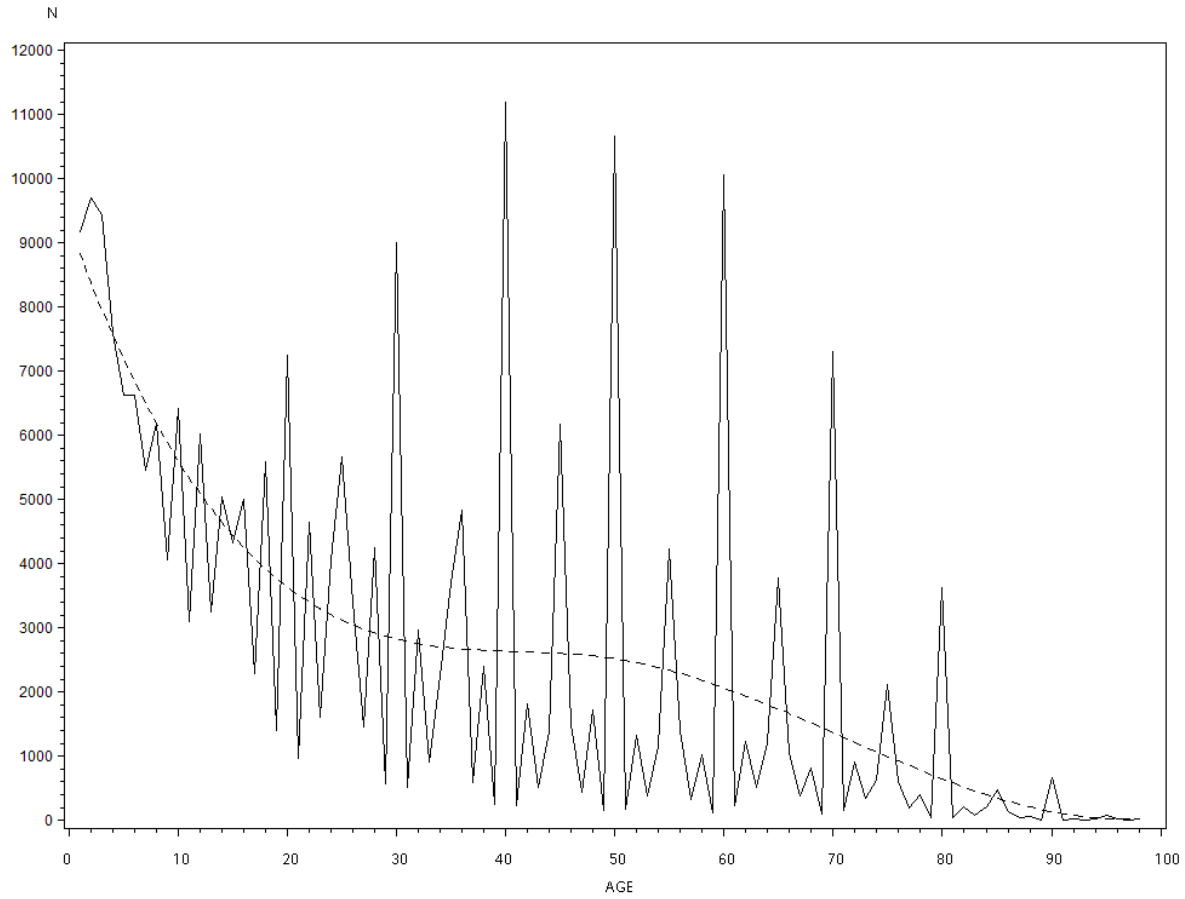
I propose to use a rather simple method that will help to discover statistically significant age heaping formats within an age distribution. For this purpose, the following five steps need to be taken:

1. *Elimination of people aged 0*
   In case of an even age distribution, at the moment a survey is taken, the number of people aged 0 is by definition only half of the number of people aged 1. This phenomenon may distort the age distribution smoothing process described in step 2. One may multiply the number of people aged 0, but as this number is nothing but an estimation, it is better to eliminate the people aged 0 altogether.

2. *Estimation of a smoothed age distribution*
   The actual, non heaped, age distribution can be estimated statistically with help of a smoothing operator. There is a large number of smoothing operators to choose from; it depends on the specific age distribution which one fits best. In the Florentine Catasto case, the best fitting operator is a spline curve with 1 knot, but for other distributions a different operator may be used. The smoothing operation results in a estimation of the real number of people for each age, the *predicted* age. The smoothed age distribution is shown in Figure 2.

[14] Herlihy et Klapisch-Zuber, *Les Tuscans et leur familles*,

Figure 2. Age and smoothed age distribution of the population of Florence, according to the Catasto, 1427. N=252,354.

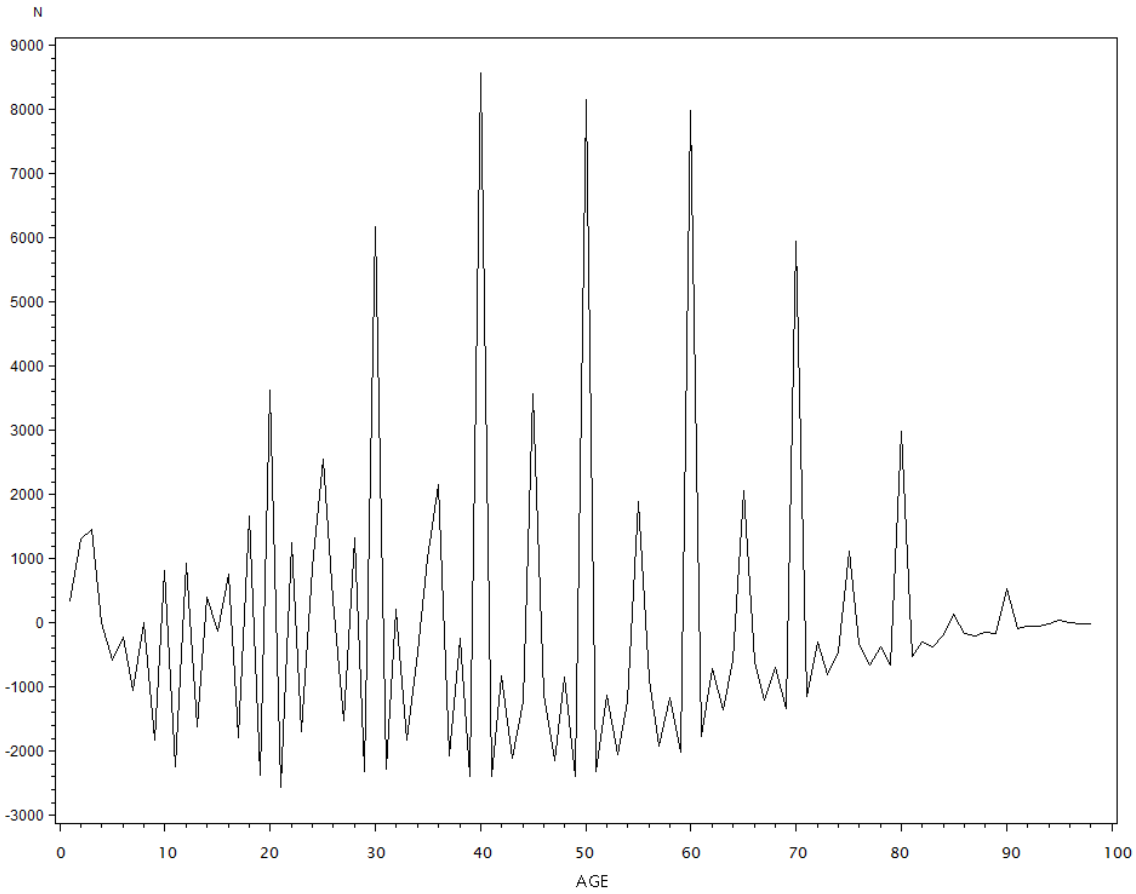


3. *Calculation of residuals*
   The next step is to calculate the residuals from the smoothed curve:

   residual(i) =N(i) - predicted(i)
   i=1-k, k being the last age mentioned in the survey

   The result is the graph as shown in Figure 3. In this figure, the spiking of ages at multiples of 2, 5 and 10 become even clearer than in Figure 1. There is distinct evidence of age heaping at the ages 5 and 10, whereas, from the age of 20 to the age of 70, age heaping at multiples of 10 is growing in importance. Next to that, there is also age heaping at multiples of 2. Before the age of 10, there is no sign of systematic age heaping.

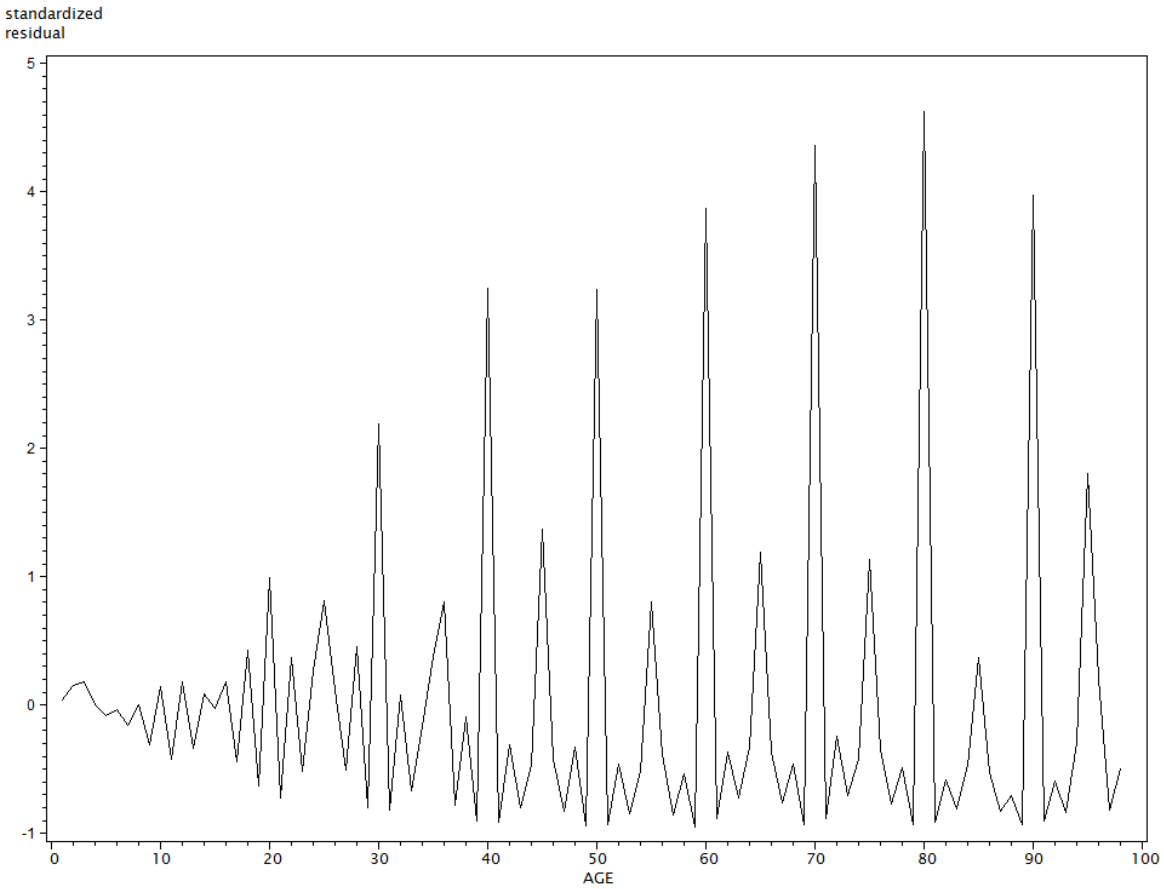
Figure 3. Residuals of the smoothed age distribution of the population of Florence, according to the Catasto, 1427. N=252,354.



Instead of calculation "raw" residuals, one may chose to calculate standardized residuals, taking into account the number of observations for each age. The formula then becomes:

standardized residual(i) = (N(i) - predicted(i))) / predicted(i)
i=1-k, k being the last age mentioned in the survey

Figure 4. Standarized residuals of the smoothed age distribution of the population of Florence, according to the Catasto, 1427. N=259,295.



For the Catasto dataset, this standardization process does not make that much of a difference. It only makes clear that age heaping at multiples of 10 is growing in importance with the ageing of the respondents up until the age of 80.

4. *Creation of dummy variables*

The next step will be to do an ordinary kind of OLS regression analysis, in which the degree to which variation of the standardized or non-standardized residuals that is caused by various heaping formats can be estimated. In order to do so, a set of dummy variables needs to be created. The dummy variables have a value of 0 for all ages, unless that age is a multiple of some kind: in that case, its value is 1. Dummy variable T0 and T1 are not created, so the first dummy variable is T2, which has the value 1 for all multiples of 2. The next one, dummy variable T3 has a value of 1 for all multiples of 3. In Table 1, the age, number of respondents, standardized residual and dummy variables T2-T16 are shown for the ages 27 to 32. Age 27 scores “1” with dummy variables T3 and T9, age 28 with dummy variables T2, T4, T7 and T14.

Table 1. Part of the dataset with dummy variables

age	N	standard-ized residual	T 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
27	1447	-0,52655	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
28	4241	0,40155	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0
29	564	-0,81077	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	9007	2,06318	1	1	0	1	1	0	0	0	1	0	0	0	0	1	0
31	514	-0,82031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	2963	0,0309	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1

There is a limit to the set of dummy variables that can be created; the last possible dummy variable is T(N(k)-4). However, normally the number of dummy variables will be restricted to a smaller number, for instance 20.

5. *Do a regression analysis*
   With the above dataset, either a standard or a stepwise OLS regression analysis can be performed with the residual or the standardized residual as the dependent variable, and the dummy variables as independent variables. In the stepwise procedure, the most significant heaping format, i.e. the dummy variable which significantly contributes most in explaining variation in the residuals is selected first, after which the second most significant heaping format comes in and so on, until none of the remaining dummy variables make a significant contribution. With the Catasto dataset, this procedure leads to the following results:

Table 2. Regression results

Variable	*Estimate*	*Standardized estimate*	*P*
T2	0,407	0,173	.003
T5	1,481	0,497	.000
T9	0,415	0,107	.027
T10	1,672	0,410	.000
T14	0,352	0,077	.126
T16	0,554	0,113	.025

R^2 = 80,0; N=98

First of all, it is clear that the explained variance of the model is very high. This shows that the model is very accurate in explaining age-specific residual variations from the predicted values. Secondly, Table 2 shows that there are six dummy variables who make a significant contribution to the model. T5 and T10 are the most important ones by far. T2 comes in third, followed by T16, T9 and, though not quite significant statistically, T14. One therefore can conclude that in Florence, 1427, three distinct heaping formats were used, with multiples of 2, 5 and 10. There is a hint that a few other heaping formats existed as well: multiples of 16 for instance, or multiples of 9. The standardized estimates of these formats are hardly significant, however, and may be caused by exceptional heaping at very specific ages. For instance, a look at Figure 3 may lead to the conclusion that there is no general heaping format at multiples of 9, but there is rather large additional heaping at the ages 36 and 45.

Robustness of the proposed method

In order to test for the robustness of the method, a few alterations to the model have been made. First of all, only "Whipple's" ages have been included in the model, causing the percentage explained variance to rise tremendously, and the minimum percentage age heaping to rise from 24,3 to 38,0 percent. In line with theory, age heaping grows in importance when people get older. Secondly, the use of non-standardized instead of standardized residuals as our dependent variable only makes a difference within the age range 1-100, where standardized residuals do better, but not within the age range 23-62.

Table 3. Tests of robustness

Ages	*Standardized residuals*	*Stepwise regression*	R^2	*Significant heaping formats (in descending order of significancy)*
1-100	yes	yes	80,0	T5, T10, T2, T16, T9, T14
1-100	yes	no	82,0	T5, T10, T2, T14
1-100	no	yes	68,7	T5, T10, T2, T12, T20
1-100	no	no	71,1	T5, T2, T10
23-62	yes	yes	96,5	T5, T10, T12, T6*, T2, T9, T14
23-62	yes	no	97,3	T10, T5, T12, T14, T2, T9, T6*
23-62	no	yes	95,7	T5, T10, T2, T4, T9
23-62	no	no	97,6	T5, T10, T2, T12, T14

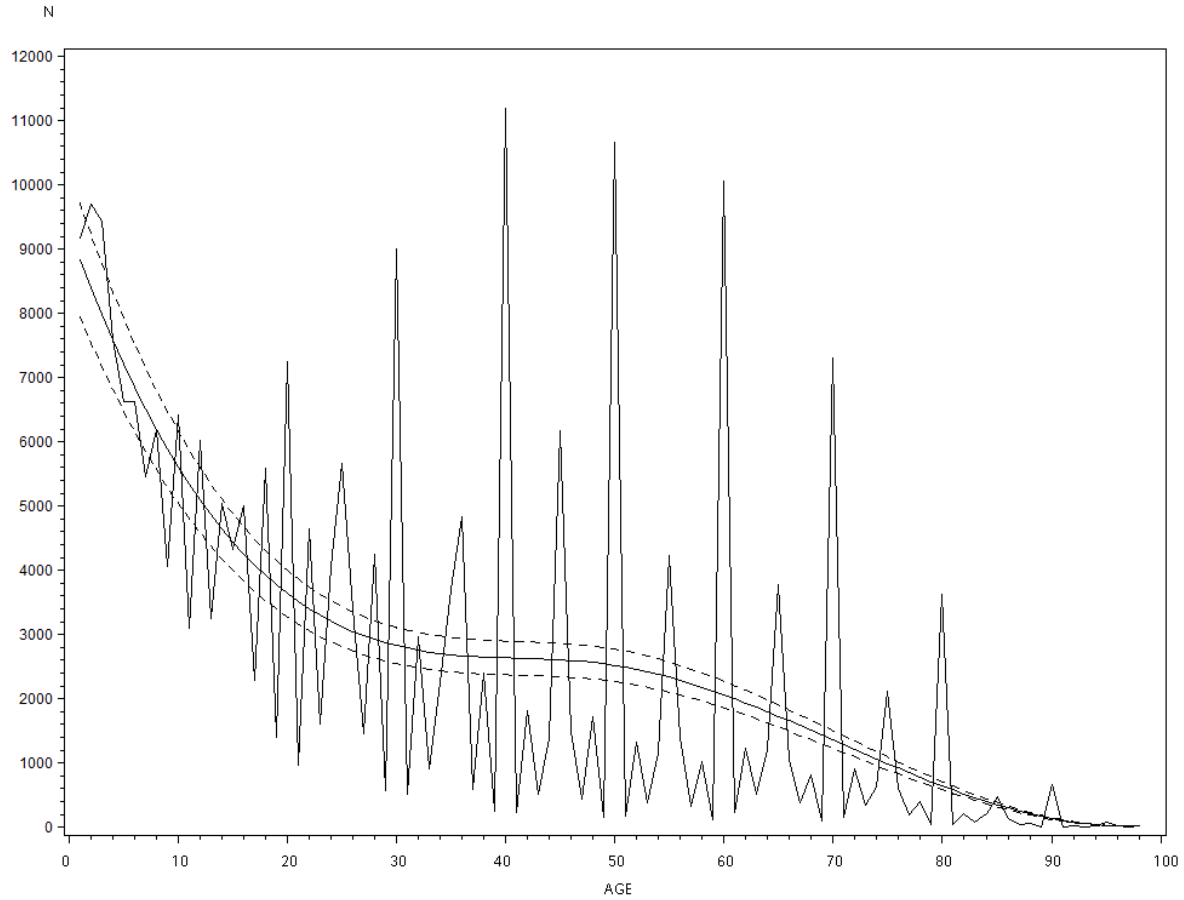
* Negative heaping effect

The tests show different sets of significant heaping formats, however. T2, T5 and T10 are always present, but T4, T6, T9, T12, T14, T16 and T20 appear less often. Within the age range 23-62, an analysis with standardized residuals shows that T12 performs better than T2, but with an analysis with unstandardized residuals, it is the other way around.

Calculation of minimum percentage of age heaping within a survey

The difference between an observed age distribution and a smoothed one already gives us an idea about the degree to which a survey is plagued by age heaping. But it is possible to make an adequate calculation of the percentage of age heaping within the survey. Consider Figure 2. Of course, even a non-heaped age distribution never is as smooth as the smoothed distribution created in Step 2. For each and every age, small deviations from the smoothed distribution are to be expected. But normally these deviations are distributed randomly. Therefore, there is no reason to believe that especially ages that are heaped are outliers because of variations in the age distribution. But one cannot be sure. Therefore, an interval around the predicted values is used which is based on the historical experience that the number of people with age^{t} is hardly ever 10% higher or lower than the number of people with age^{t-1} or age^{t+1}. In Figure 5, such a confidence interval of 10% has been used.

Figure 5. Confidence intervals around the smoothed age distribution



If we assume that those who state their age(i) when N(i) is below the confidence interval do so correctly, one is able to calculate the minimum percentage of age heaping within a survey. All observations that exceed the confidence interval are thought to be the result of age heaping. This allows us to measure the percentage of age heaping within the population sample:

*percent age heaping= ((∑(resid(i)-(pred(i)+(pred(i)*0,1))/∑(i))*100)*
I=1-k

For the Florence Catasto survey, the minimum percentage of age heaping amounts to 24,3% when all ages 1-100 are taken into account, and 38,0% when only the afes 23-62 are taken into account.

Concluding remarks

The method proposed here is simple and straightforward. It gives a definitive answer to question which heaping formats are being used in the survey and what their contribution is to the overall heaping effect. Next to that, the method also gives a good idea of the total amount of people who were inclined to age heaping. In the case of the Florentine Catasto, this is not totally true, because not everybody stated their own age; this was done by the family heads. But for other historical surveys, the method will do fine.