A case study of missing historical heights: Estimation using Multiple Imputation

Quaranta, Luciana

2007

Luciana Quaranta

A case study of missing historical heights in Friuli (Italy)

Estimation using Multiple Imputation[1]

The aim of this paper is to study how aggregative analyses are influenced by missing data, in particular when using military records. An attempt will be made to explain why mean height calculations cannot accurately describe the population as a whole in cases where measurements of certain members are not available. The way in which missing data can be estimated using Multiple Imputation and the advantages of having complete data will afterwards be described.

Living standards can be analysed by using many different types of indices, for example GDP (gross-domestic product) or HDI (human development index). These indices cannot, however, be employed in historical contexts, since the information needed to construct them is not always available. Anthropometric indicators are used instead. For example, the distribution of heights in historical populations is of considerable interest in the biomedical and social sciences, since it provides information about nutrition and health and therefore indirectly also about living standards. Stature is, in fact, a long term indicator of the nutritional status of an individual.

Military records are the main source of anthropometric information. Important problems can arise, however, in cases where there is missing data, especially if it is not missing with a random pattern. For example, the trend of mean heights cannot precisely reflect the changes in the living standards of the population as a whole if people from a particular group were not measured. Many techniques have been developed as an attempt to solve this problem. One of these is Multiple Imputation, an estimation procedure which can be very effective in cases where several variables present missing data.

This paper is organized as follows. I first describe, generally, the problem of missing data, the method of Multiple Imputation, and its implementation through the ICE package of Stata. Successively I evidence why there can be missing information in military records. A brief description of the data used will be given, showing the effects of missing measurements when

[1] This work was developed within the ASTRI project, carried out by the *Archivio di Stato di Udine* in conjunction with the Department of Statistics of the University of Udine.

studying mean heights. Afterwards, I will describe how Multiple Imputation can be employed to estimate data. Lastly, results will be presented graphically, to evidence how studies can improve when complete datasets are available.

1. Missing data and the use of Multiple Imputation

Missing data is a difficulty widely faced in research, and one which leads to three major problems (Rubin 1987, 1-23). On the one hand, reduced sample sizes lead to less efficient estimations. Secondly, it is not possible to carry out analyses that involve complete data methods. Lastly, data can be biased, mainly because subjects with missing information are not necessarily the same as those with complete data. In fact, missing data generally does not occur completely at random and, therefore, the differences between measured and unmeasured subjects can be due both to known and unknown factors.

A solution adopted by some statistical packages is to use “complete-case analysis”, which consists in excluding units that have missing values for any of the variables involved in the study (Little and Rubin 2002, 3-59). Two other possible solutions are to apply different weights to variables that have effectively been measured or to assign mean values to missing data. All of these procedures, however, may cause biases.

When indicators of missingness hide true values that are important for the analysis, it is appropriate to impute missing values, in other words, to obtain means or draws from a predictive distribution of missing data, created based on the observed information. With Multiple Imputation, a technique which was introduced by Rubin (1987), two or more values are drawn from this distribution and are used to replace missing data, thus producing *m* datasets, each containing different imputed values for the missing datum. Inferences from the combination of these datasets are very effective and reflect three aspects: uncertainty derived from non-response, sampling variability that results from missing values, uncertainty about the correct model used to estimate missing data.

Additional advantages of this technique are that the newly created datasets can be analysed by employing the same standard analytical methods adopted for complete datasets, and also that the estimated data can be used for many different purposes, while it is only necessary to perform imputation once.

Most statistical software contain Multiple Imputation procedures. In this work I particularly concentrate on the ICE package of Stata, which was originally introduced by

Patrick Royston. An in-depth description is given in Royston (2004; 2005a; 2005b) and van Buuren, Boshuizen, Knook (1999).

ICE performs multivariate imputation by using regression switching in an iterative procedure. This allows to estimate, in the same process, missing values for different variables. Initially variables which contain missing data are ordered randomly, and their observed values are replicated across missing cases. Therefore, at a first stage, data are filled in at random. Afterwards, the distribution of each variable is sampled conditional on the distribution of the remaining covariates. For each variable in turn, missing values are thus imputed by applying an univariate procedure. This step is repeated as many times as the number indicated in the `cycles` option, and at each cycle previous imputations are replaced with newly calculated values. With the completion of these cycles, a single imputation sample is created. The procedure is then repeated *m* times independently, to obtain *m* imputations. Each dataset can be analysed separately, but parameters of interest are averaged across the *m* copies of the data, thus giving one single estimate.

2. Military records and problems of missing data

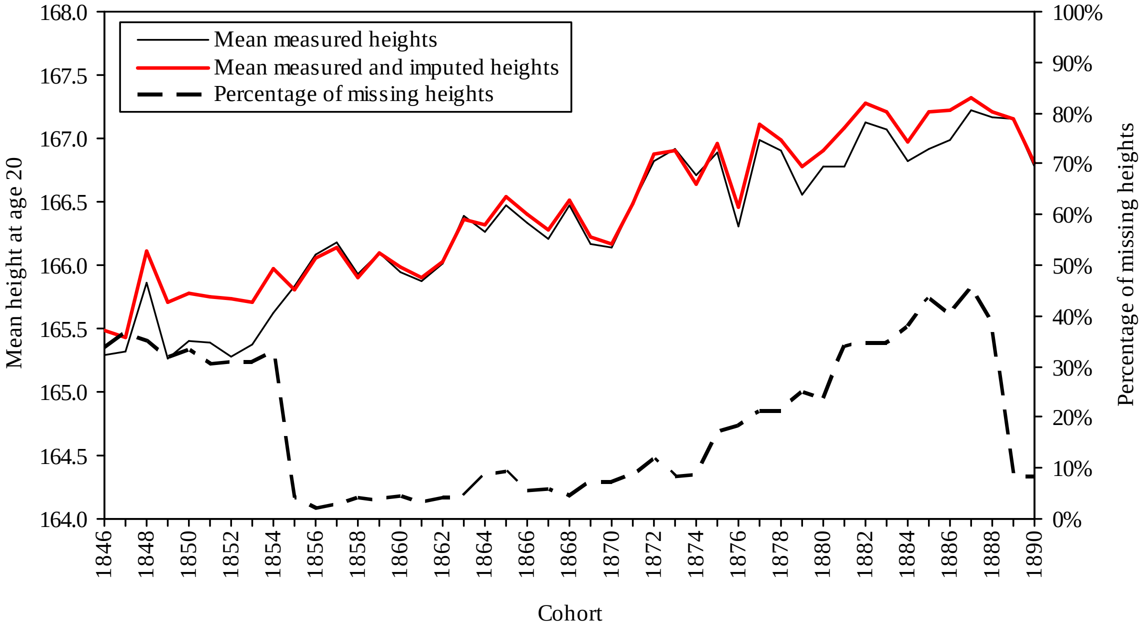
In the particular case of military records, the largest difficulty regarding missing data is faced in areas where conscription was not universal, but voluntary. In fact, in places like the UK or the USA, height distributions are truncated on the left tail, since only males who exceeded the minimum height requirements presented themselves to the medical examination. Many studies have dealt with this problem; for example, Floud, Wachter and Gregory (1990), A’Hearn (2004) and Komlos (2004). In Italy, on the contrary, conscription was universal until the year 2004, meaning that all males who were around twenty years old had to attend a medical examination[2]. During this examination conscripts could be declared able (in some cases with the possibility of being exempted), or unfit, either temporarily (these males were required to return to the medical examination the following year) or definitively, in other words they were rejected. Using information from this source, it is therefore possible to study the health status of the entire male population and to obtain aggregative anthropometric measures.

[2] During medical examinations the following information was registered: measures of height, chest circumference and in some cases weight, illnesses or malformations, occupation, literacy (whether conscripts knew how to read and write).

The data analysed in this paper relates to some of the military districts of the region of Friuli[3]. In particular, the enrolment lists compiled between 1867 and 1910 have been considered, which regard the cohorts 1846-1890[4]. The database holds information for around 90,000 conscripts[5].

Even if Italy is very advantaged in terms of the richness of its military records, it also faces problems of missing data. Figure 1 shows, for the area analysed, the trend in mean height at age twenty[6] and the percentage of missing data for cohorts 1846-1890. Variations throughout time in the reasons why some conscripts were not measured can explain the large fluctuations observed in the percentage of missing data. In particular, two different periods can be distinguished: 1846-54 and 1855-90.

Figure 1: Mean height at age 20 and percentage of missing data for the region of Friuli



[3] Military districts of Ampezzo, Cividale, Gemona, Latisana, Moggio, San Daniele, San Pietro, Tolmezzo.
[4] For the districts of Ampezzo, Cividale, Gemona and Latisana there is no data for cohorts 1868 and 1874, since the relative registers are not available.
[5] For this study only data relative to the first medical examination of each conscript was considered.
[6] Mean age at measurement was different for each cohort. Three reasons determined these variations. The first regards the effect of the date of birth; for example, when measurements were taken a conscript born in January would have been nearly a year older than one born in December. Secondly, the military medical examination did not always take place at the same time of the year, implying that age varied for the different cohorts. Lastly and most importantly, different cohorts were called to the examination at different ages. In fact, males born in 1846-50 were called for conscription after the 21st birthday, but later cohorts had to attend the medical examination at a younger age (for example, the mean age at measurement for young males born in 1890 was 19.5). Considering that in the past, or more precisely in populations with modest living standards, the growth period in some cases continued even up to age 26 (Hulanicka, Kotlarz 1991, 429-434), the variations through time in mean height can only be studied after having obtained estimates of measurements at a precise age (in this case twenty). For an in-depth description of the method used to obtain these estimates, see Fornasin, Quaranta (2006). Further improvements to this method will be introduced in the forthcoming Breschi, Fornasin, Quaranta (2007).

2.1.Missing heights of Friulian conscripts of cohorts 1846-54

Initially males who were exempted from military service were not measured. The heights of 33% of conscripts born in the period 1846-54 are missing for this reason. Conscripts could be exempted, for example, for being orphans, the only son, the first-born child of a widowed mother, or for having a brother who was doing the military service at the time or who had died in the army. Instead, from the cohort 1855 onwards, exemptions were treated in a different manner. Conscripts had to first attend the medical examination in the same way as almost all other males, and it was only possible to request exemptions afterwards. For exempted males born in 1855-90 height measurements are therefore available.

It is important to observe whether heights of exempted males could have been different than statures of other conscripts. If differences did exist, measures of the mean height of the population would have been incorrect.

Various studies have demonstrated that, on average, first-born males have smaller heights than younger siblings (M. Hermanussen, *et al.* 1988; Al-Omair 1991). The proportion of first-borns was lower amongst exempted males, implying that the mean height of this group was probably higher than that of other conscripts. Preliminary models were constructed using the data of the cohorts 1855-90, which have confirmed this hypothesis.

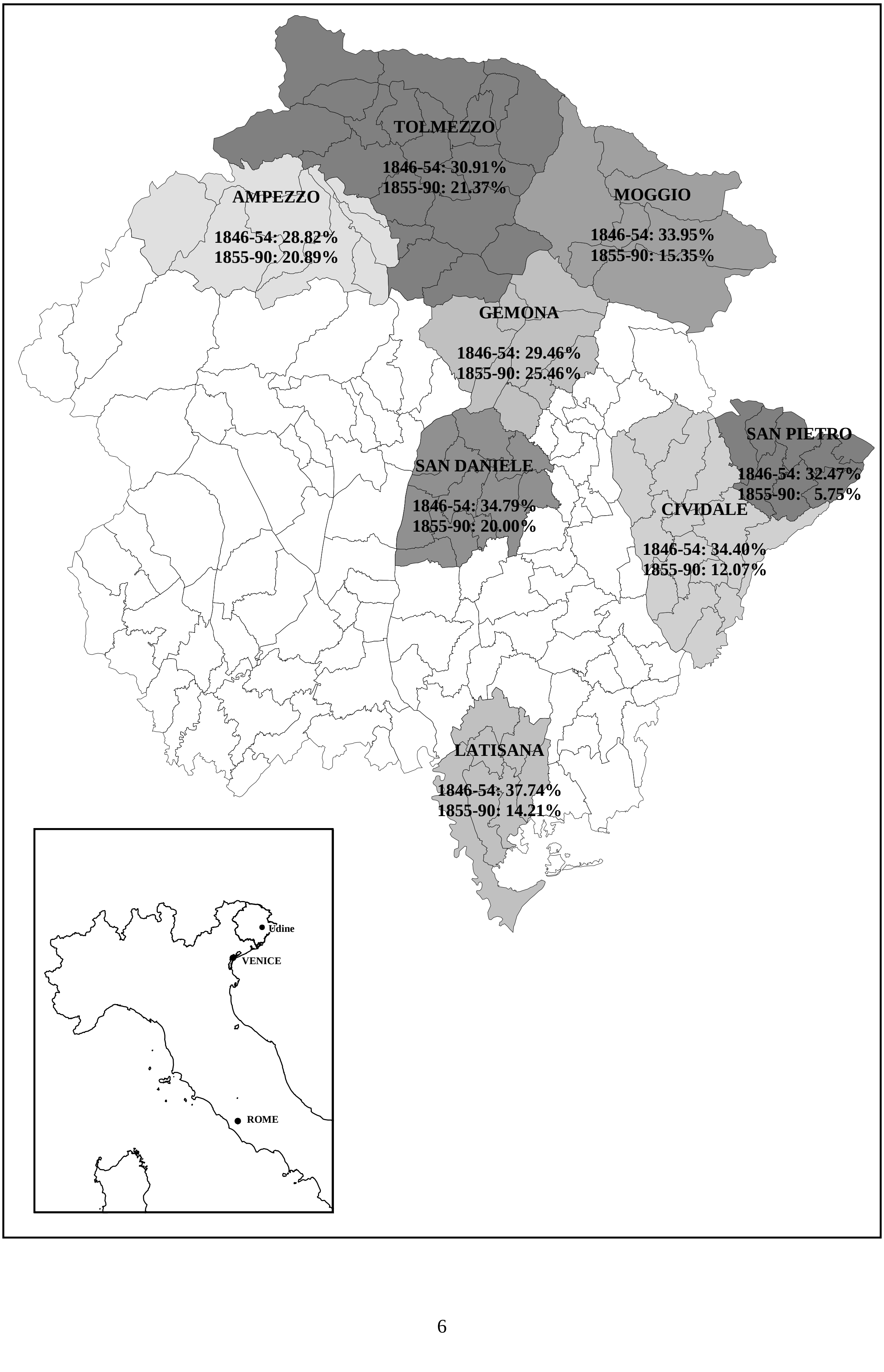
Map 1 shows the military districts of the region of Friuli which were analysed and the respective percentages of missing heights. Differences in the percentages of missing data can be observed. It was chosen to carry out estimates using district as a covariate and considering the entire dataset.

2.2. Missing heights of Friulian conscripts of cohorts 1855-90

In subsequent years, migration flows, which were particularly important in the Alpine region of Friuli, were the main cause of missing data[7]. The economy of this area was, in fact, very weak and the only resource available in excess was population. Some migrants left the country definitely while others seasonally. This second flow mainly concerned certain members of the population: construction and forestry workers (who left in spring and returned in late autumn) and also textile workers and house-servants (who were away from autumn until spring). Seasonal migrants were primarily directed to Austria and Germany (Renzulli 1978, 133-152).

[7] Some young males were not measured because they were draft evaders. It is possible to hypothesise, however, that many of them were not informed about conscription laws since they had migrated at a very young age and, therefore, that they ignored that conscription was a requirement for all males (Arcaleni 1998).

Map 1: Percentage of missing heights in the districts of Friuli analysed. Cohorts 1846-54 and 1855-90



If a young male had not regularized his military position he was not allowed to exit the country. For this reason, many conscripts who intended to migrate asked to be enrolled to the military service before the date of the medical examination. It was also possible for another member of the family of the young male to enrol him if he had already left the country. Furthermore, conscripts could also present themselves to an Italian embassy or consulate in their country of residence instead of returning to Italy for their medical examination (Illari, 1999). Diplomatic authorities only indicated, for cohorts 1855-88, whether the person was suitable for military service, without registering height measures nor any other information.

For cohorts 1855-90, 18% of heights were missing. As Figure 1 shows, percentages were much higher than this mean value for young males born in 1874-88 (31%), as a result of increases in migration flows. Cohorts 1889-90 evidenced a drop in missing data, since heights of conscripts who had migrated started to be registered[8].

Map 1 also shows the territorial variations in the percentage of missing data for these cohorts. It can be seen that the southern territories and those confining with Slovenia had lower percentages of missing data. A possible explanation for these variations are differences in labour specialization of each area and therefore in the type of economy.

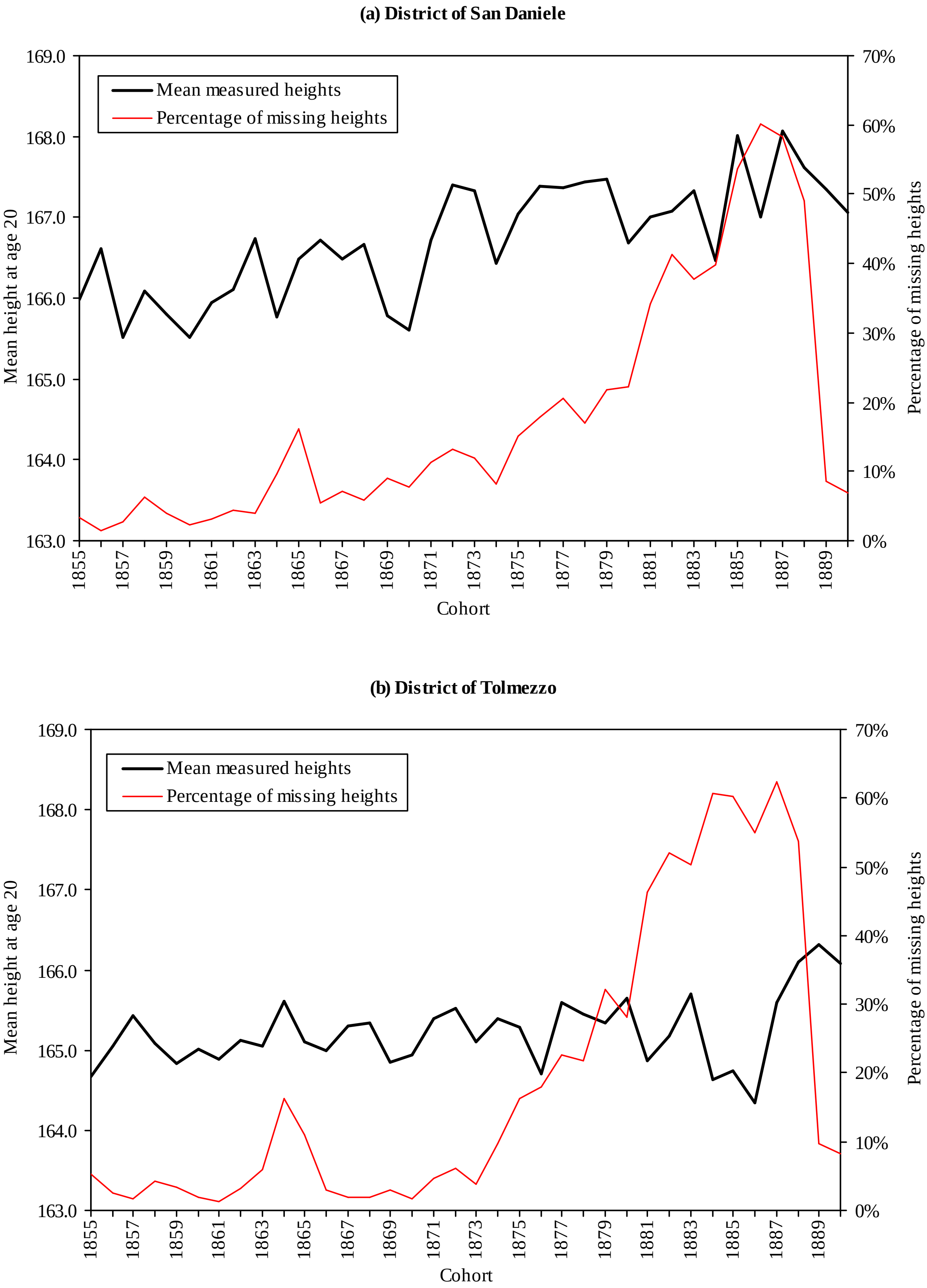
It is important to consider whether missing data might have determined biases when calculating the mean stature for each district. Some studies (Arcaleni 1998, 50) have assumed that migrants are often the most robust members of the population[9].

Analyses carried out using data of cohorts 1889-90, where heights of migrants had been registered, have evidenced different patterns for each district. In some cases, in fact, conscripts who migrated were, on average, taller than those who did not, while in other districts the opposite situation was observed.

[8] These changes could have resulted from various reasons. One might have been the anticipation of the date of the medical examination, which might have allowed conscripts who migrated seasonally to attend the examination before leaving the country. The examinations for the cohorts 1889 and 1890 were, in fact, carried out, respectively, at the beginning of March and February, therefore before spring, which was when many seasonal migrants departed. The drop in the percentage of unmeasured heights could have also resulted from the introduction of new conscription laws, which have changed how the number of conscripts that had to be enrolled in the army was decided. This might have represented an incentive to record measurements for all young males.
A slight decrease in migration flows was also evidenced, which could perhaps be explained by the Bosnian Crisis of 1908-1909, which was caused by the annexation of Bosnia and Herzegovina by Austria-Hungary in October 1908.

[9] Danubio, Amicone and Vargiu (2005) demonstrated that southern Italian men who had emigrated to the USA between 1908-28 and 1960-70 were taller than southern Italian recruits. These authors have stated that these differences were probably due to the lack of accurateness of self-reported male heights, since in this case migrants were mostly unskilled workers. Even if misreporting had occurred, however, height differences were probably also due to the higher likelihood of healthier members of the population to migrate.

Figure 2: Percentage of missing data and mean heights for the districts of San Daniele and Tolmezzo, cohorts 1855-90



(a) District of San Daniele

(b) District of Tolmezzo

Figure 2, which regards the districts of Tolmezzo and San Daniele, evidences these two opposing patterns. It can be seen from the graph that in Tolmezzo mean height was negatively related to the percentage of missing data. Cohorts 1882-87 show very low mean stature and very high percentages of missing data (more than 50%). On the contrary, conscripts born in 1888-90 had the highest mean heights, while the percentages of missing data were very low (less than 10%). It can therefore be concluded that migrants of Tolmezzo were the taller members of the population. The district of San Daniele, instead, shows a positive correlation between mean height and percentage of missing data, meaning that in this case migrants were probably shorter than other members of the population[10].

By observing the figure it is possible to conclude that mean heights do not correctly describe the population as a whole if the percentages of missing measurements are high. It is therefore necessary to try to estimate this missing data, in this case using different estimation models for each District.

3. The application of Multiple Imputation to estimate military records

Height is determined by a very complex interaction between genetic and environmental factors, which takes place along the entire growth period of an individual (Tanner 1989, 119). It is impossible, however, to account for all these factors nor to measure them. Estimates of missing statures of conscripts can therefore only be based on the information contained in military registers.

It is now possible to explain how Multiple Imputation can be used to estimate missing heights. Considering that data was not always missing for the same reason, it is clear that estimates should be carried out separately for conscripts born before or after 1855. However, since occupation was not registered for males born before 1859, estimates were made using the following two groups of cohorts: 1846-58 and 1859-90.

[10] In the district of Tolmezzo the majority of both migrants and non-migrants worked in the construction sector. A different pattern was observed in San Daniele, where most migrants were kiln-men, while the majority of non-migrants were involved in the agricultural sector. Agricultural workers of San Daniele were, on average, taller than kiln-men of the same district.

3.1. Estimation of missing heights for cohorts 1846-58

The method adopted to estimate missing heights of cohorts 1846-58 can now be described. The construction of preliminary regression models allowed to identify which variables were highly correlated with height and, therefore, which information to use to obtain estimates. It is also important to consider how these variables were treated in the models. The covariates considered in this case were age, year of birth, district of residence, health status, and indicators of whether males resided abroad and of whether they had been exempted.

Not all of these covariates presented complete data, and it was therefore important to implement a procedure that would also estimate them. Table 1 shows variables which had missing data and the respective percentages. An advantage of the use of Multiple Imputation is that several variables can be estimated simultaneously.

To carry out estimates, the entire dataset was used (cohorts 1846-90), in order to observe height measurements for exempted males, which were only registered for cohorts 1855-90. After the estimation process was completed, only data imputed at this stage for conscripts born in the years 1846-58 was considered. More accurate estimates could in fact be obtained for individuals born in the years 1859-90 by adopting a procedure which also took into account occupation.

The first variable which can be described is the year of birth. It should be considered when estimating stature, since people of the same cohort were likely to have experienced similar living conditions and, therefore, similar factors might have influenced their growth. No estimations were needed for this variable, since data was complete.

Age was not directly registered, but it could be calculated precisely using the dates of birth and of the medical examination, which were both available. However, the date of birth of more than 80% of conscripts born in 1847-58 had to be estimated due to missing data. Information was more complete for cohort 1846.

The procedure adopted consisted in estimating the day of the year (comprised between 1 and 365) in which conscripts were born, considering district of residence as a covariate. In this case, it was incorrect to assume that the seasonality of births was normally distributed. A bootstrap sample was therefore used to estimate $\hat{\beta}_*$ (Royston 2005a, 232). Another specification made was that the day of birth was interval censored between 1 and 365[11].

[11] The imputed value of *y* therefore followed a truncated normal distribution with bounds 1 and 365. Mean and variance for *y* were estimated using interval regression.

Table 1: Patterns of missing data – cohorts 1846-58

Variable	Number of conscripts with missing data	Percentage of conscripts with missing data
Age	18190	84.5%
Health Status	1510	7.0%
Height	5070	23.6%

After having imputed the day of the year in which conscripts were born, it was possible to calculate their age at measurement, an operation that required two steps. The complete date of birth was first calculated, taking into consideration the year in which each young male was born. Using this date and the date of the medical examination, the age of conscripts was obtained. In this case missing values did not have to be imputed, but only calculated from the already imputed days of birth[12].

Another factor which highly influences height and growth is health status (Livi 1905; Tanner 1989; Fornasin, Quaranta 2006). Military registers contained information on the illnesses and malformations that affected conscripts, which corresponded to the reason why some young males where reformed from the military service. Pathologies were first classified into groups, according to the typologies or to the body part that they affected and, more importantly, to the effects they had on height and growth (Fornasin, Quaranta 2006). These effects were observed through preliminary regression models.

Information on the health status of conscripts was included in models using dummy variables. The groups of pathologies considered were: anaemia, goitre, height under 154 cm, general physical disharmony (extreme weakness or slimness), other illnesses which can affect growth (for example bone problems, tuberculosis, tumours), and other illnesses which probably do not affect growth (for example skin and eye problems, deafness, muteness).

When considering the health status of exempted conscripts, it should be questioned whether these males would have been declared able or rejected if they had attended the medical examinations in the normal manner; i.e. if they were or not healthy. The *Regio Decreto* n. 5100, passed on June 26 1869, and which therefore regarded cohorts 1848-90, partly answers this question. It states that a conscript who had the right to be exempted but who also believed to have physical defects which rendered him unable to the military service could first ask to be reformed, and if this request was denied he could then request to be exempted.

[12] The `passive()` command can be used when attempting to obtain a value for a certain variable, not by estimating it, but by passively imputing it based on values of other variables (which can either already exist, or be imputed) (Royston 2005, 190-191). This command can be used, for example, when it is necessary to calculate multiplicative interactions between variables.

Figure 3: Percentage of exempted and of reformed conscripts and total percentage of missing heights

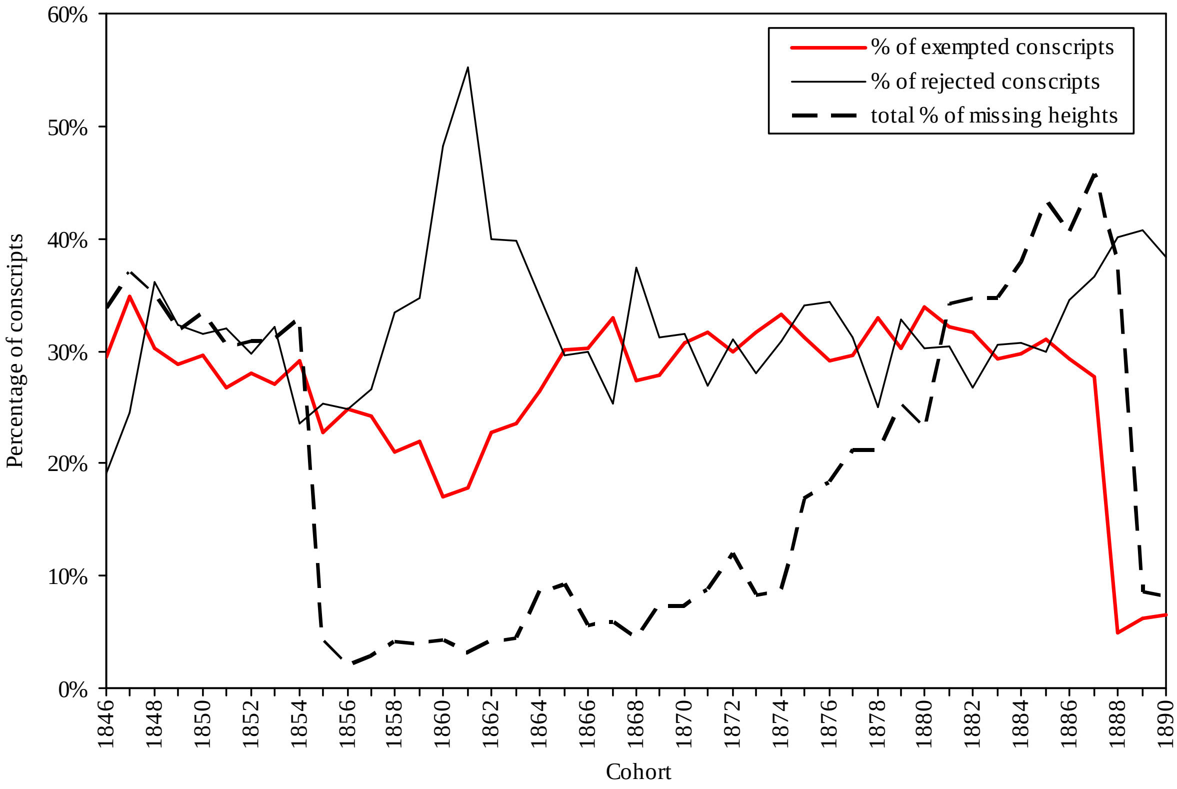


Figure 3 evidences, for each cohort, the percentage of exempted and rejected males and the total percentage of missing heights. Cohorts 1846-47, which attended the medical examination in the years prior to the described law, show much lower percentages of rejected males than other cohorts, evidencing that some exempted conscripts would have probably been reformed if they had attended the examination normally. For cohorts 1849-59 it was, instead, possible to assume that all exempted males were healthy. The percentages of reformed males in fact respect the trends of successive years.

To obtain accurate calculations of height, it was therefore important to estimate whether exempted males of cohorts 1846-47 would have been declared able or rejected from the military service. Estimates were based on the following covariates: district of residence, an indicator of whether the conscript resided abroad, and a variable representing either the percentage of able conscripts of the relative cohort (for young males born in the years 1848-58, period in which the percentages of rejected males probably represented reality) or the mean percentage of able conscripts of all cohorts (for young males born in the years 1846-47, where the percentages of rejected males were lower than expected).

In cases where the estimated decision was “rejected”, an illness (anaemia, goitre, height under 154 cm, general physical disharmony, other illnesses which can affect growth, other illnesses which probably don’t affect growth), or in other words the reason for rejection, also had to be estimated. Conditional imputation was employed, in order to avoid assigning an illness to conscripts considered able[13]. The covariates used were year, district of residence and height.

3.2. Estimation of missing heights for cohorts 1859-90

For the cohorts 1859-90, and for each district separately, heights were estimated by considering year of birth, age, occupation, health status, and indicators of whether the conscript resided abroad and of whether he was exempted. Table 2 shows which of these variables presented missing data and the respective percentages.

Table 2: Patterns of missing data – cohorts 1859-90

Variable	Number of conscripts with missing data	Percentage of conscripts with missing data
Age	825	1.3%
Exemption	3796	5.8%
Health Status	4107	6.3%
Height	12577	19.2%
Occupation	14889	22.7%

Once again, complete information on the year of birth was available and no estimates were therefore needed. Missing age (calculated precisely) data was estimated using the same procedure described earlier, with the exception that cohort was used as a covariate in estimates of the day of the year (1-365) in which conscripts were born.

Preliminary observations should also be made regarding health status. Figure 3, shows that some of the conscripts who were not visited (in this case migrants) would have probably been rejected if they were visited like other males. In fact, percentages of rejected conscripts were lower for those born before 1889 than for males of cohorts 1889-90. Moreover, a direct look at the 1889-90 data shows that 36.5% of migrants were rejected from the military service

[13] The `conditional()` option of ICE allows to carry out estimates of a variable *x* conditional on the value of variable *z*. In this case *z* represents two subpopulations (able and rejected), and therefore *x* can vary with respect to which subpopulation the individual belongs to. All able conscripts were set to be “healthy” when estimating the health status, while all rejected males were given any value other than “healthy” (one of the six different types of illnesses).

(either temporarily or definitively) because they were affected by one of the illnesses or malformations described earlier.

Estimates of whether conscripts would have been declared able or rejected if they had attended the medical examination in the normal manner were therefore obtained. The covariates considered were occupation, an indicator of whether the young male resided abroad, and a variable representing either the percentage of able conscripts of the relative cohort (for those born in the years 1859-68 or in the years 1889-90, period in which percentages of missing data were low) or the mean percentage of able conscripts of all cohorts (for young males born in the years 1869-88, where the percentages of missing data were highest). For young males considered "rejected", an estimate of an illness was obtained, using year and height as covariates.

Data on whether conscripts were exempted from the military service was missing for 5.8% of young males. This variable was estimated considering cohort and height as covariates.

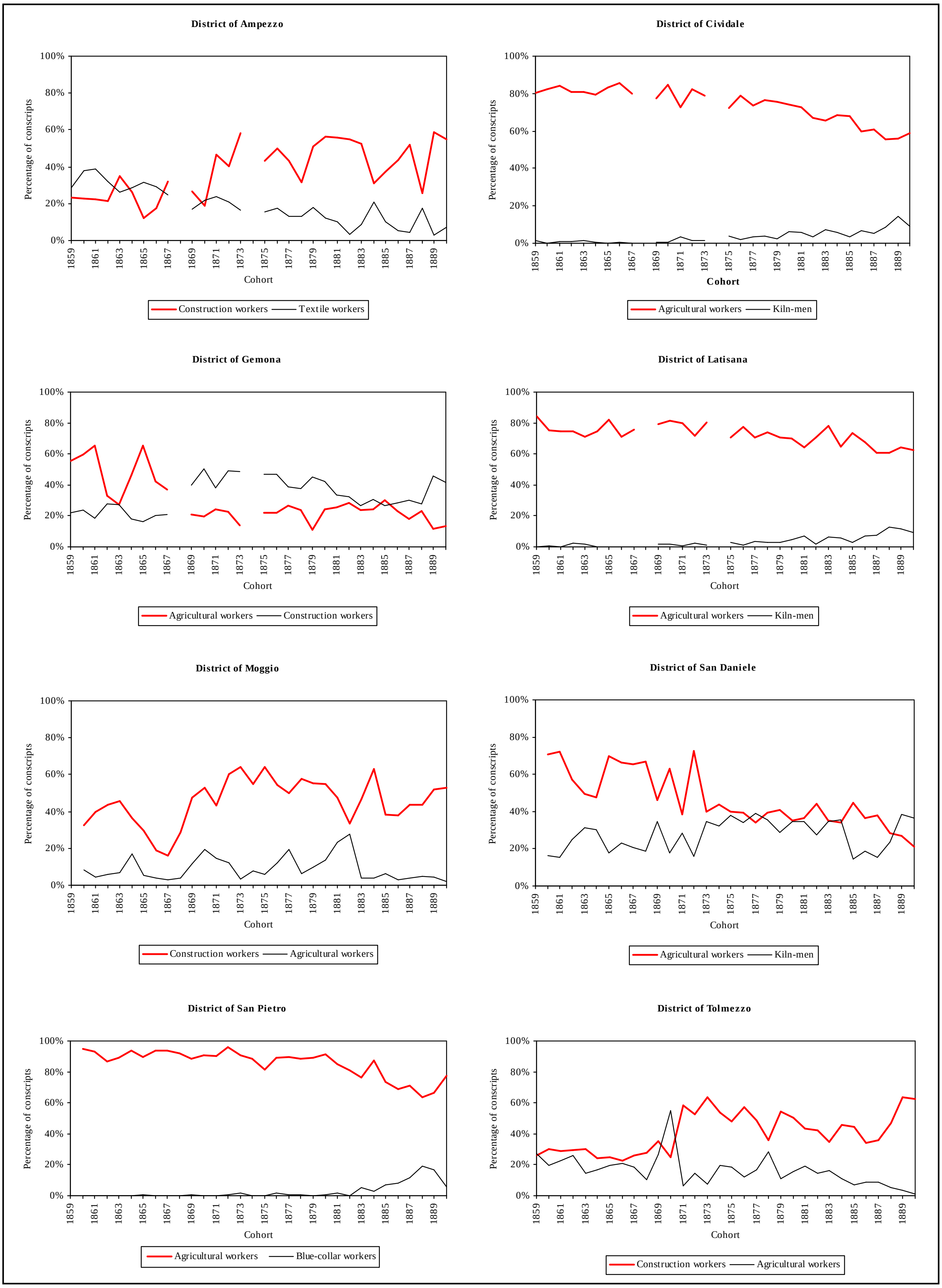
The most complicated variable to estimate was occupation, which was missing for 22.7% of conscripts (Table 2). A classification into 12 categories was first made[14]: agricultural workers; craftsmen; textile workers; merchants; blacksmiths; carpenters; kiln-men; blue-collar workers; construction workers; forestry workers; elite; other. The elite category grouped together high-paying occupations and also those which required many years of studies. Some examples are students, teachers, land-owners, etc.

Figure 4 shows, for each district, the trend of the percentages of conscripts involved in the two most diffused occupations or in those that evidenced, in the observed time period, the largest variations in importance. The first aspect that can be noted is that males born in 1859-70 presented a different distribution of labour than cohorts 1871-90.

For estimates to be accurate, these temporal changes had to be respected. It was, therefore, necessary to adopt different estimation models for the two groups without, however, dividing the dataset, which would have led to less accurate estimates of other variables. Occupation was thus estimated twice, following methods appropriate to the two periods. By interacting these two estimates with a variable that indicated whether the conscript was born in 1859-70 or in 1871-90, one single estimate for each individual was afterwards selected.

[14] These categories were selected taking into consideration the HISCO classification (Breschi, Fornasin, Manfredini, Marzona 2006; van Leeuwen, Maas, Miles 2002; 2004).

Figure 4: Percentage of conscripts involved in the two most diffused occupations



Note: Registers for cohorts 1868 and 1874 for the districts of Ampezzo, Cividale, Gemona and Latisana are not available. In Moggio, San Daniele and Gemona occupation started to be registered for males born after 1859.

The method adopted for cohorts 1859-70 was rather simple, since the percentages of missing data were low. Occupation was estimated based on the covariates cohort, height, and a variable that indicated whether the conscript resided abroad.

A more complicated procedure had to be adopted for estimates of cohorts 1871-90, since percentages of missing data were much higher. The fact that migration mainly concerned particular professions implied that the distribution of labour of non-migrants (observed data) was probably rather different than the distribution for migrants (non-observed data). Cohorts 1889-90 represented an advantage in the analysis, since occupations of migrants were also registered. This information could therefore be used in the estimates of males born in other years. It was, in fact, possible to assume that the distribution of labour of migrants of cohorts 1871-88 followed patterns similar to conscripts born in the years 1889-90.

Estimates were obtained through a series of steps. Taking into consideration, for each district, the two occupations shown in Figure 4, a variable with three categories was first created (occupation1, occupation2, other occupation). For conscripts with missing data this variable was estimated considering as covariates height, an indicator of whether the conscript resided abroad, and a variable which stated the period of reference. This allowed to estimate, for conscripts of cohorts 1871-88 whose data was missing, a value for the three level categorical following the same distribution of occupation presented by data of cohorts 1889-90.

The use of this variable as a covariate, together with height and year, permitted to obtain more accurate estimates of occupation for conscripts born in 1871-90. In fact, the three level categorical variable represented an expected value for the unobserved data. When estimating occupation and also the three level categorical variable, bootstrap samples were employed.[15].

The models adopted to estimate each variable for cohorts 1859-90 have now been all described. A note can be made on how imputation equations were selected. In order to choose between different possible models, the accurateness of estimates had to be tested. This was done by obtaining estimates for individuals whose data was complete, and by later comparing real with calculated heights. Data of the District of Tolmezzo was used for this procedure and a new dataset was created[16].

[15] Furthermore, since occupation is a categorical variable, when including it's value in models to estimate stature, the `substitute()` and `passive()`options of ICE were used (Royston 2005, 190-191).

[16] To create a new dataset, at a first stage only conscripts with complete data were selected. Secondly, missing data was artificially created by deleting information on occupations, health status, heights and date of births of some young males. This deletion was not, however, done randomly. Data of cohorts 1889-90 evidenced that a very high portion of migrants from the District of Tolmezzo were construction workers. In order to respect this pattern, only data for construction workers were, therefore, deleted.

The best model was selected by observing the mean error. In this case by *error* is meant the difference, for each conscript, between estimated and measured stature. In fact, even if it was important to select an appropriate estimation model for each variable, the main parameter of concern was the accurateness of height.

The mean error of the model selected and described equalled 0.06 cm. A *ttest* (*p-value* 0.72) showed that it was possible to accept the hypothesis that, on average, measured heights were equal to estimated heights. Two further aspects observed when choosing models were the distribution of errors, which proved to be normal both in graphs and in distributional tests, and the trends in the mean values, for each cohort, of real and estimated heights.

4. Results

Having described the methods adopted to impute missing variables, the results obtained can now be presented. Figure 5 shows the trend in mean height at age twenty (Fornasin, Quaranta 2006) for exempted, non-exempted and all conscripts. It can be observed that mean estimated heights of exempted males (cohorts 1846-54) respect the values shown by exempted males who were measured (1855-90), except for cohorts 1846-47, which had lower mean heights than successive cohorts.

Figure 5: Trend in height at age twenty of exempted, non-exempted and all conscripts

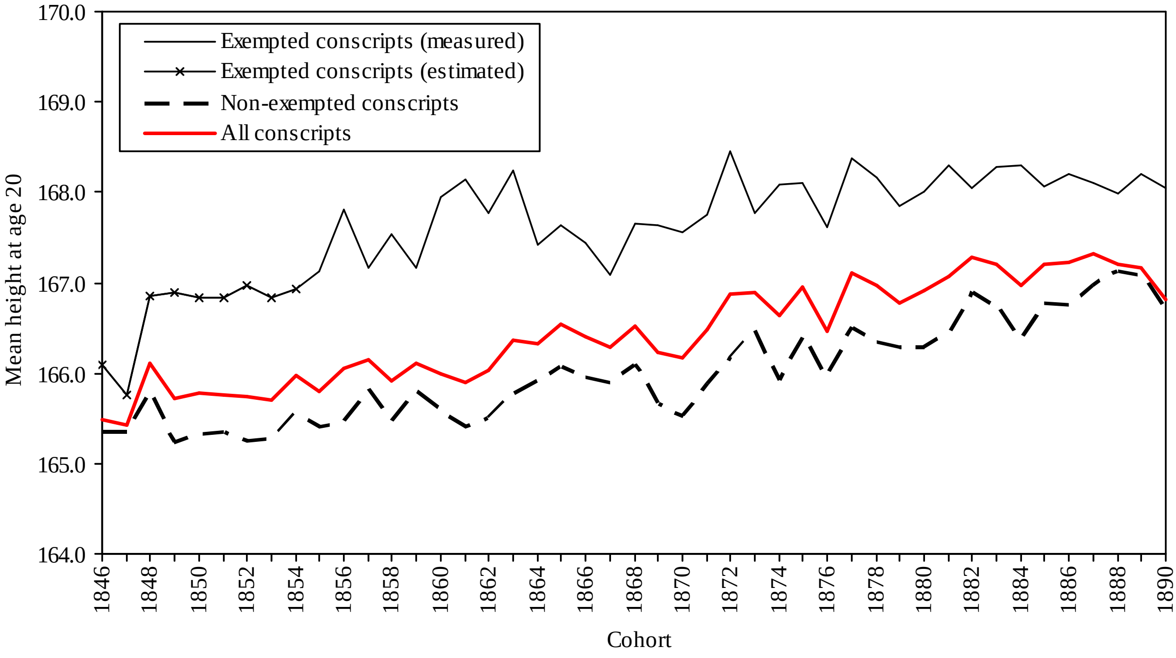
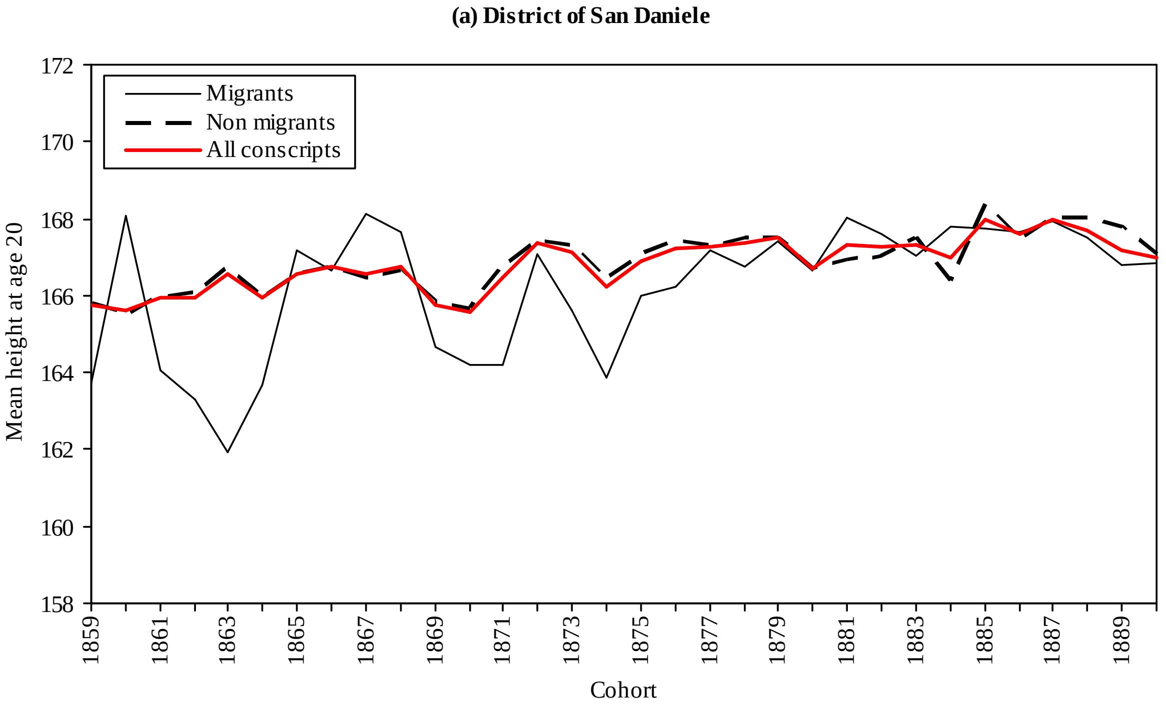


Figure 6: Trend in height at age twenty for migrants, non-migrants and all conscripts for the Districts of San Daniele and Tolmezzo



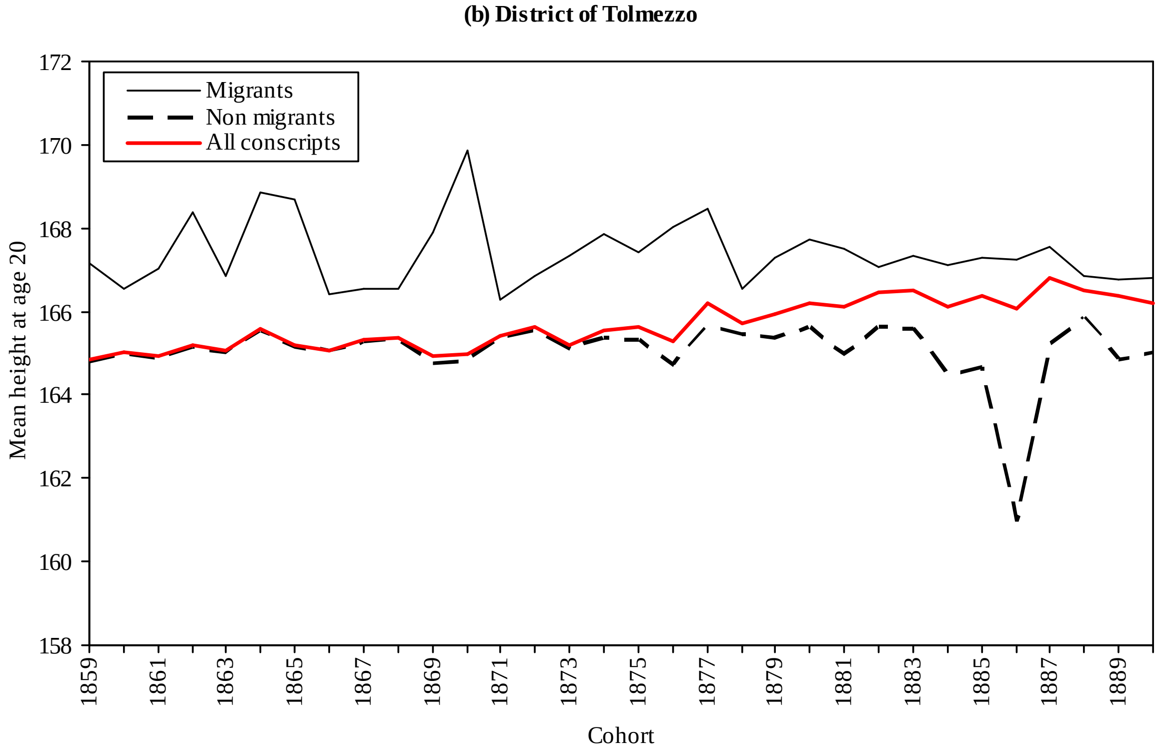


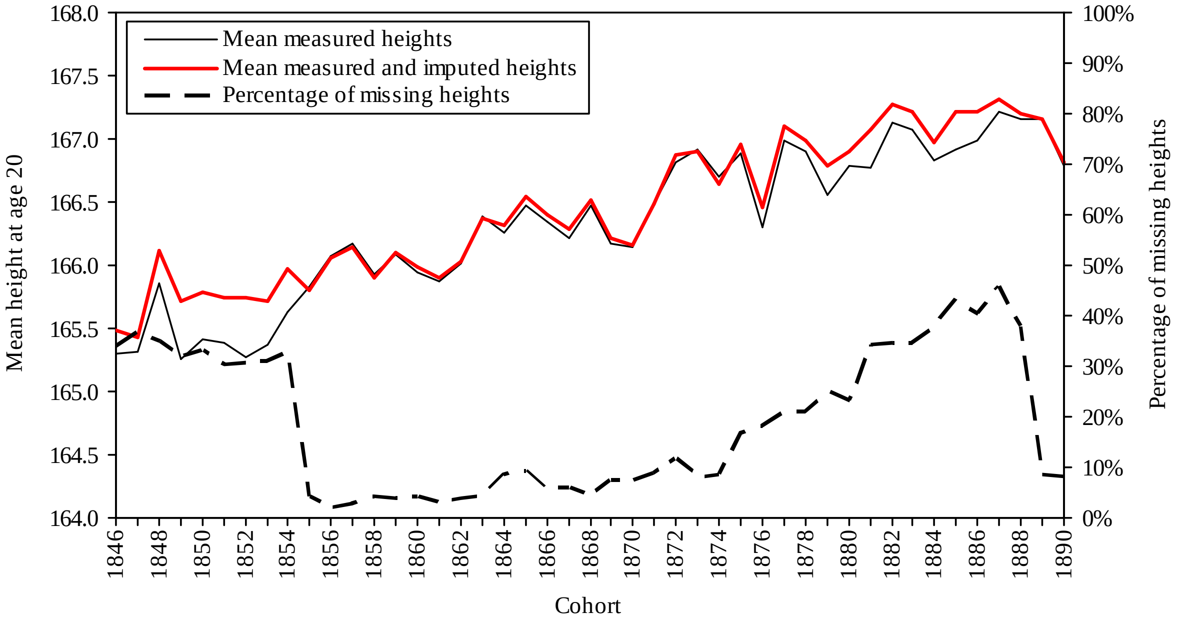
Figure 6 shows the trend in mean height at age twenty for migrants, non-migrants and all conscripts of the Districts of San Daniele and Tolmezzo, for cohorts 1859-90. Less importance should be posed on data of migrants born in the years 1859-74, where the migration percentages were quite low.

Lastly, the trends in mean measured and mean measured and imputed heights are shown in Figure 7. It is possible to observe a clearer trend when the entire population is considered, that is, when mean heights are calculated using both measured and imputed data. In most cases mean height increases when estimates of missing data are included. As expected, wider differences are seen in periods where the percentages of missing data were highest.

For cohorts 1846-54, the higher mean heights obtained when estimates of exempted males are also considered can be explained by the fact that these conscripts were taller than other males for two reasons. One is that the proportion of first-born (therefore of shorter males) amongst this group was probably lower than in the rest of the population. Secondly, for cohorts 1848-54, as described earlier, all exempted males were probably healthy and would have therefore been declared able if they would have been examined.

With regards to cohorts 1855-90, the higher mean heights obtained can be explained by the fact that, even if each district presented different migration characteristics, the overall pattern demonstrates that conscripts who were not measured as a result of migration were, in mean, taller than males who remained in the territory and attended the medical examination normally.

Figure 7: Trend of mean measured and imputed heights for all military districts



5. Conclusions

The aim of this paper was to evidence the problems that can arise when analysing datasets which large proportions of missing information. The initial paragraphs described the problem of missing data and one method which can be used to solve it: Multiple Imputation. In particular, its application through the ICE package of Stata was introduced.

Successively, the importance of the use of military records when studying living standards was described, also evidencing how incorrect conclusions can be made in cases where measurements for certain sections of the population are not available. In the region of Friuli two reasons for missing data were identified. The first regarded cohorts 1846-54, where a large portion of conscripts had been exempted from the military service without being measured. The second concerned males born in the years 1855-90, were the majority of migrants (seasonal or definitive) were not measured.

In both cases, heights were estimated based on other information available on military registers: age, year of birth, health status, occupation, and indicators of whether the conscript resided abroad and of whether he had been exempted. Some of these variables (health status, age and occupation) also presented missing data and therefore had to be estimated as well. One advantage of Multiple Imputation is that values for different variables can be obtained through the same procedure.

The results have evidenced that mean heights varied if imputed values for missing data were also considered, showing, in most cases, an increase. This study demonstrates that analyses carried out using incomplete datasets can be biased and that better results can be obtained when missing data is appropriately estimated.

References

B. A'Hearn 2004, *A restricted maximum likelihood estimator for truncated height samples*, «Economics and Human Biology», 2, 5-19.

A.O. Al-Omair 1991, *Birth order, socioeconomic status and birth height of Saudi infants*, «Journal of the Royal Society of Health», 111(6).

E. Arcaleni, 1998, *La statura dei coscritti italiani delle generazioni 1854-1976*, «Bollettino di Demografia Storica», 29, 23-59.

M. Breschi, A. Fornasin, M. Manfredini, A. Marzona, *Occupations and the rise of migration in Friuli (North-eastern Italy) in the second half of the 19th century*, Presented at the XIV International Economic History Congress 2006, Helsinki.

M. Breschi, A. Fornasin, L. Quaranta 2007, *Heights of twenty years old males of Friuli (Italy) born between 1854 and 1890.* Forthcoming.

M.E. Danubio, E. Amicone, R. Vargiu, *Height and BMI of Italian inmigrants to the USA, 1908-1970*, «Economics and Human Biology», 2005(3).

R. Floud, K. Wachter, A. Gregory 1990, *Height Health and History. Nutritional Status in the United Kingdom, 1750-1980*, Cambridge: Cambridge University Press.

A. Fornasin, L. Quaranta 2006, *La statura degli italiani nati dal 1854 al 1890. Prime ipotesi per la costruzione di una nuova serie storica*, Working Paper N.8, Dipartimento di Statistica dell'Università degli Studi di Udine.

M. Hermanussen, B. Hermanussen, J. Burmeister 1988, *The association between birth order and adult stature*, «Annals of Human Biology», 15(2).

B. Hulanicka, K. Kotlarz, *The final phase of growth in height*, «Annals of Human Biology», 10 (1991) 429-434.

V. Ilari, *Storia del servizio militare in Italia*, vol. II, *La «Nazione armata» (1187-1918)*, Roma, Centro militare di studi strategici, 1990.

J. Komlos 2004, *How to (and how not to) analyse deficient height samples*, «Historical Methods», 37(4), 160-173.

R. Little, D. Rubin 2002, *Statistical Analysis with Missing Data*, New York: John Wiley & Sons.

R. Livi 1905, *Antropometria militare*, parte 2, *Dati demografici e biologici*, «Giornale medico del Regio Esercito», Roma 1905.

G. Renzulli 1978, *Economia e società in Carnia fra '800 e '900*, Istituto Friulano per la Storia del Movimento di Liberazione, Udine, Arti Grafiche Friulane.

P. Royston 2004, *Multiple imputation of missing values*, «Stata Journal», 4(3): 227-241.

P. Royston 2005a, *Multiple imputation of missing values: update*, «Stata Journal», 5(2):188-201.

P. Royston 2005b, *Multiple imputation of missing values: Update of ice*, «Stata Journal», 5(4):527-536.

D. B. Rubin 1987, *Multiple Imputation for Non-response in Surveys*, New York: John Wiley & Sons.

J.M. Tanner 1989, *Foetus into Man. Physical Growth from Conception to Maturity*, Cambridge Ma, Harvard University Press.

S. van Buuren, H. C. Boshuizen, D. L. Knook 1999, *Multiple imputation of missing blood pressure covariates in survival analysis*, «Statistics in Medicine», 18 (6).

M. van Leeuwen, I. Maas, A. Miles 2002, HISCO. *Historical International Standard Classification of Occupations*, Leuven, Leuven University Press.

M. van Leeuwen, I. Maas, A. Miles 2004, *Creating a Historical International Standard Classification of Occupations: An Exercise in Multinational, Interdisciplinary Cooperation*, «Historical Methods», 37: 186–97.