From Census Records to People: Reconstructing the Population of France in the SocFace Project

Jérôme Bourdieu, PSE - Paris School of Economics & EHESS - Ecole des Hautes Etudes en Sciences Sociales
Yannick Dupraz, Université Paris Dauphine-PSL, PSE & INED - Institut national d'études démographiques, France
Christopher Kermorvant, TEKLIA, France
Lionel Kesztenbaum, INED & PSE

Extended abstract – Preliminary, do not quote

The *listes nominatives du recensement* are the surviving individual records of the French censuses. Produced every five years from 1836 to 1936 and organized spatially (municipality; wards, hamlets, or streets; houses; households), they summarize census information, listing each individual with some of her characteristics, e.g., name, year of birth, or occupation. The sheer number of listes (around 20 million images from 1836 to 1936, corresponding to approximately 500 million individual records) and their dispersion over space (these records are kept in a hundred different archive repositories) have limited their use until now. The SocFace project aims at making them accessible for all French censuses from 1836 to 1936 at the individual level by developing AI-based tools to automatically transcribe handwritten entries. Taking advantage of Deep Learning, SocFace vastly reduces the financial and ethical costs of manual transcription: 300 million people manually transcribed offshore would cost an estimated €45 million, compared with less than 1% of that cost for our automated approach. It provides open access to one of France broader and wider historical source.

But automation adds another layer between users and the source material. Historian Carlo Ginzburg famously described historical sources as a "distorted glass" through which one sees the world. AI-generated transcriptions contribute to that distortion, but in a peculiar (and novel) way. In fact, one could almost say that automated transcriptions add a new glass, complementary but distinct from the original one: they contain errors (on average 15% at the word level) that are unevenly distributed, e.g., common names like "Marie" are transcribed accurately, whereas less frequent ones are more distorted. These challenges concern both the treatment of documents and the use by social scientists of the extracted information. On the one hand, extracting complex information from a very large array of archival sources requires a workflow as simple and consistent as possible. On the other hand, normalizing and distributing such a vast amount of data requires clear processes and documentation. On both sides, there is some tension arising from simply scaling up standard technologies. We quantify these issues in the case of the SocFace project and describe the various tools we have developed to assess data quality and implement corrections post-processing. One important feature is that these tools need to be designed with the intended use of the dataset in mind: they won't be the same for linking individuals across censuses; studying

social mobility; or assessing fertility by cohorts. So, our tests are end-user oriented and differ depending on which problem we want to tackle.

Assessing quality over millions of HTR records

Despite the tremendous increase in the quality of automated text recognition to extract information from historical documents, challenges remain regarding the processing of very large sets of documents (hundreds of millions of records): some individuals might be hallucinated by the model; most text fields are very noisy; some pages may have been omitted in the processing, etc. This is a major issue when scaling up a project such as the processing of census lists for France over 100 years. Not only is it inconceivable to manually correct all issues, but even identifying and assessing them need to develop specific tools. These tools must be adapted to variations in the original forms (20 in total); specific habits and uses of each writer (e.g., abbreviations, internal references); temporal and spatial variations in the very meaning of words.

Indeed, without automated transcription, French historical census information would not be available to historians. But at the same time, using these methods results in a tremendous amount of data of uneven quality. Moving forward, we have developed various tools to assess data quality and implement corrections post-treatment. Importantly, such tools need to be designed with the intended use of the dataset in mind. For instance, for linking individuals it doesn't matter if 'Arnaud' (a first name) is spelled 'Arnaut' or 'Ernaud'; but it does matter if the age is 86 instead of 16. So, our tests are end-user oriented and differ depending on which problem we want to tackle.

A first set of tools is related to the images themselves, without text recognition or entity extraction: since there is a fixed set of lines (individuals) on each page, bar the last (around 30 individuals on a page), we can estimate the population of a given municipality for a given year and compare it to the official figure produced at the time by the administration. Large discrepancies indicate an issue with the images: some municipalities were merged, due to a mistake in metadata; some images were missed in the collecting process; some images are missing from the original archives; etc. This tool helps to pinpoint issues related to the sorting of the original source that hinder automated recognition. Simultaneously, it provides a complete picture of the image collection and its limitations, something of great interest for social scientists wishing to work with the final dataset.

A second set of tools is directly related to estimating model quality, at various levels (lines, pages, documents). We compute a set of binary indicators for internal consistency at the line level (e.g., nobody can be 134 years old; someone with a female first name cannot have a male position in the household). At this stage, we are not concerned with the source of the error: if Julie is a son, it's either because the first name (Julien) or the position (daughter, given that fils and fille are very close in French) was misread. What matters here is only that there is an error in the information that will be used by social scientists (for instance the sex ratio of children might be biased). As of now, we have 12 different tests for each line, that we average at the page level. In the end, a quarter of pages do not have any issue, many have a few issues, while a few have many issues. Systematic use of these tests helps us to locate problematic pages, identify their characteristics, e.g., years or municipalities; position in the document (is it more frequently the first or last pages?); type of municipality (are large cities harder to decipher?). In the future, we will build more complex consistency tests at the household level (is there more than one wife of the household head? does the age difference between the head of household and his children make sense?)

Granular data to assess transformation of the social space

Eventually, SocFace will produce a complete database of all individuals who lived in France between 1836 and 1936. In addition, to make the most of the richness of the lists, records between censuses will be linked using automated methods. This linkage also needs to take into account the quality control built during the extraction project. This database could also be exploited as a basis

for new research in combination with other databases and sources: in the same way that the census today serves as a basis for many other surveys, the dataset of historical censuses could be mobilized to develop quantitative historical research on France in many directions.

Censuses provide a series of nearly exhaustive snapshots of the French population. These censuses have been digitized over nearly a century, allowing, among other things, the study of social phenomena at the finest possible scale, at the individual level. While it might be thought that each snapshot is itself a sample from the theoretical set of all possible French populations at that point in time, the sample we have is not just any sample: it is the actually observed and realized sample, which makes it a unique sample, whose potential singular properties are important in themselves. Above all, due to their very size, these snapshots allow for the description of phenomena with an extremely high level of observational diversity. In a certain sense, in terms of generic variables, these data provide a very impoverished perception of the social world: it is limited to 5 or 6 variables: name, first name, sex, age, place of birth, place of death, occupation. But the internal variability of each of these variables is immense and, by definition, unparalleled, since it is an exhaustive statistic.

Take occupation for instance: the data obtained through SocFace allow for a quantitative evaluation of professional diversity in France over space and time and as a function of simple individual characteristics such as age and sex. Thus, various phenomena can be observed that relate to the functioning of employment and its evolution over time. What is observed are the transformations of the occupational space during the process of industrialization. This is of course assuming the issues of transcription quality have been addressed, which, importantly, does not mean fully solved: for some individuals (we need to describe whom) the profession cannot be identified; for others, the professions retained are dependent on prior processing. Furthermore, the mention of profession is very heterogeneous, notably in its precision.

In this framework, two main hypotheses can be considered, which are partly contradictory and partly cumulative. The first one suggests that what is observed is the homogenization of the labor market over time as statistical knowledge and practices increase, the state create standard nomenclatures, regional designations of professions that are otherwise almost identical (vaisselier/ménager/cultivateur) are progressively abandoned. This might also be the translation of a real phenomenon (so not related only to its measurement), that is the process of de-skilling and homogenization of the workforce inherent to industrialization. All unskilled industrial workers become laborers. This mechanism can also be observed in the greater or lesser prevalence of precision regarding a profession: worker versus mason worker. There is, of course, the risk of attributing real meaning to variations that are due only to variable administrative practices (in areas where the number of census agents is low relative to the population, they do not have the time to be detailed). The study of the professional space must take into account the fact that some observed variations are an effect of the source and the treatment of the source. The second hypothesis, on the contrary, would be the manifestation of the continuation of the process of social division of labor and the appearance of new professions related both to new production techniques and to new modes of work organization.