Journal de la Société Française de Statistique
Vol. 158 No. 2 (2017)

Editorial of the special issue: Humanities and Statistics

Titre: Éditorial du numéro spécial : Humanités et Statistiques

Stéphane Lamassé[1] and Fabrice Rossi[2]

Digital humanities have combined successfully information technology with all the disciplines of the humanities. Early initiatives like the Text Encoding Initiative [3] led to the creation of massive digital text archives, followed by multimedia archives described by rich metadata. In addition to this digitizing movement, researchers started to apply methods from the humanities to analyze digital entities, such as computer mediated interactions of actors in online social networks. This trend is amplified by the open data movement which give access to very large corpora of data, in particular to those produced or collected by public bodies (governments, from local to federal). In the near future, digital humanities will handle "Big Data" in the original sense of massive data.

The availability of digital data describing either traditional humanity objects or digital objects has increased vastly the possibilities for conducting quantitative analysis in the humanities. Researchers can leverage tools from statistics, data mining, data visualization, and so on, to mine historical data, geographical data, legal data, etc. In the simpler cases, off-the-shelf techniques can be applied to standard data tables. This leads frequently to valuable insights on such data, proving that classical methods are still relevant to the humanities even in a digital context. Classical methods are generally included in the curricula of humanity scholars who remain this way autonomous in their research work. This is especially the case when those methods are available in standard statistical software.

While data are becoming ubiquitous, the humanities did not suddenly switch from a data less situation to the data deluge that somewhat plagues data mining. The increasing use of data in the humanities manifests itself both with an increase of the volume of data in some specific cases and with a generalized use of data in more and more academic disciplines. In this latter case, data are seldom available in large quantities. On the contrary, they frequently exhibit all the difficulties associated to sparse data or to small volume data. Even classical statistical methods can be ill suited for those situations.

In addition, systematic data collection produces more and more complex data that call for either specific methods or at the bare minimum for recent statistical methods. A natural consequence is the emergence of multidisciplinary collaborations in which humanity researchers, statisticians

[1] LaMOP UMR 8589, Université Paris 1 Panthéon Sorbonne, 1 rue Victor Cousin, 75232 Paris cedex 05.
E-mail : stephane.lamasse@univ-paris1.fr
[2] SAMM EA 4543, Université Paris 1 Panthéon Sorbonne, 90 rue de Tolbiac, 75013 Paris, France.
E-mail : fabrice.rossi@univ-paris1.fr
[3] http://www.tei-c.org/

 ISSN: 2102-6238

and data scientists are associated. In general new methods are designed in an iterative way that extends progressively the underlying model in order to account for all the complexity in the data. As a consequence those types of collaboration are more adapted to medium to long term research projects than to short term ones.

Fortunately, there are numerous opportunities for researchers from the humanities and from data science to meet. Papers presenting practical cases involving humanity data handled with statistical methods have been published on a regular basis since at least the nineteen sixties. In France for instance, one the earliest journal dedicated to such interaction was *Mathématiques et sciences humaines* [4] published by the *Centre d'analyse et de mathématiques sociales* from EHESS (CAMS [5]), and founded by Marc Barbut. More generally, all the disciplines of the humanities have similar journals and display a rich and lively scientific activity evolving statistics and demography, history, sociology, etc. Workshops and conferences on such subjects are also very frequent. For instance, the CAA conference (*Computer Applications & quantitative Methods in Archaeology* [6]) has been first organized in Birmingham in 1973 and is still organized nowadays on a yearly basis. It was motivated in part by the impact of preservation acts which induced a surge of excavations leading to an almost mandatory use of digital solutions for recording and analyzing the associated findings. This is a typical example of the way digitization creates both data and the need for a statistical analysis of them.

The goal of this special issue is to give a sample of recent research collaborations between humanity scholars and statisticians, in a way that illustrates the richness of those interdisciplinary approaches. As any small sample of a large population, this special issue cannot claim an exhaustive coverage of the field. However, its six articles give an idea of current practices (an advanced use of statistical methods for complex data) and of ongoing developments (new methods and new models). It shows both how recent statistical methods make their way into the common practices of humanity researchers and how completely new methods are designed to tackle problems that are specific to humanity data. All the articles from this special issue were reviewed by experts from both side of the "divide": in addition to the invited editors (an historian and a statistician) and to the editor-in-chief, each paper was reviewed by a statistician and by a researcher from the main humanity discipline. As a consequence, we believe that the resulting papers will be interesting both for statisticians and for humanity scholars.

The first article of this special issue, "Le chantier de la tour de Mutte à Metz : regards sur la production du fer en Lorraine à la fin du Moyen Âge. Fouille de données, analyses prédictives et traitement spatial des données. " by A. Disser, M. L'Héritier, P. Dillmann and A. Arles, provides a perfect example of an expert use of classical statistical methods by archaeologists. Those methods include principal component analysis, ascending hierarchical clustering and logistic regression. The archaeological problem under study is the history of European ferrous metallurgy, more specifically in Metz (Lorraine, France) and at the end of the Middle Ages. The authors base their work on measurements made on ferrous reinforcements and on their lead sealings used during the

[4] https://msh.revues.org/
[5] http://cams.ehess.fr/
[6] http://caaconference.org/

 ISSN: 2102-6238

building of the belfry of Metz. The study is conducted with a rather evolved methodology mostly based on unsupervised (or semi-supervised) methods rather than on supervised ones. This type of approach is quite popular among humanity scholars but also among researchers in information visualization, for instance. It is based on the general principle of putting the expert at the center of the analysis. As a concrete example, let us consider the case of the clustering of the lead sealings conducted in this article. While those sealings are characterized by their chemical composition and by their position on the facade of the belfry, the authors use only the chemical composition in the hierarchical clustering method. Then a posterior validation on the clustering is performed by the experts considering the positions of the clusters on the facade. The authors also reject a supervised approach in a latter part of the article in order to avoid problems that might be induced by biases in reference data. Among other aspects, this article emphasizes a major difficulty in statistics that is quite common in historical studies, for instance: data production and data collection are seldom under the complete control of the researchers. There are situations in medicine, agronomy, or more generally in industry, for instance, where experimental designs can be controlled (or even optimized). In history, on the contrary, reference data have been at least partially produced by an historical process. A naive application of standard statistical methods on such data can introduce some confusion or some blending between the data production and selection process on the one hand, and the historical phenomenon under study, on the other hand.

In "La mobilité inter-entreprises des migrants de Tunisie en région parisienne dans les Trente Glorieuses. Quelques outils statistiques au service d'une démarche historienne. ", A.-S. Bruno shows the importance of taking into account the statistical structure of the data when constructing a model. Multilevel mixed models are relatively standard in contemporary statistics. They are particularly useful in numerous applications which include repeated measurements, e.g. survival analysis. As repeated measurements appear in social sciences, history and more generally in the humanities, mixed models should be used to take advantage of all the available observations. This is the case in the present paper with an application in the field of biography analysis. The author studies career paths as a way to understand and quantify factors that favor inter-firm mobility. Data have an obvious hierarchical structure with repeated measurements: the core of the data analysis focuses on individual biographies in which the "survival" of a person in a given firm is modeled. In general, each biography contains several trajectories of a given individual in multiple firms. The multilevel mixed model points out a plurality of factors explaining inter-firm mobility, such as the activity segment of the firm.

This article and the previous one give an idea of how autonomous humanity scholars have become with respect to the use of statistical models. In this specific case, Bruno's article illustrates the evolution of research practices induced by the integration of recent models in humanity researchers' toolbox. In the early 2000s, such career path analyses were conducted with a standard Cox model. As a consequence, a given biography was generally reduced to a single trajectory in order to avoid the adverse effects of ignoring the repeated nature of the measurements. This reduced significantly the number of trajectories under study. The transition to multilevel models allows humanity scholars to include in their analysis entire data sets. This is particularly useful in disciplines where data remain scarce.

The rest of this special issue is dedicated to articles written by multidisciplinary teams including

 ISSN: 2102-6238

researchers from statistics and from the humanities. Methods presented in the following articles have been developed more recently as general purpose statistical techniques or, in some cases, have been specifically designed in order to solve specific problems associated to research questions from disciplines of the humanities.

A first illustration of this inter-disciplinary contributions is given by "Un problème clé de la paléodémographie : comment estimer l'âge au décès ? " authored by H. Caussinus, L. Buchet, D. Courgeau and I. Séguy. The objective of this work is to estimate the distribution of ages at death of a population using measurements made on skeletal remains (cranial measures in this article). On a statistical point of view, one has to estimate firstly the conditional distribution of the age at death given the cranial measurements (this conditional distribution is supposedly stable through long historical periods). This is a relatively simple problem as reference data are available: collections of cranial measures associated to true ages of death have been produced. As shown by the authors, those collections are large enough to provide acceptable estimates of the conditional distribution if one does not want to distinguish genders, but the quality of the estimates per gender is questionable. Once the conditional distribution has been estimated, it might seem that a simple application of the Bayes rule would lead to an acceptable estimate of distribution of ages at death given the observed distribution of the cranial measures. The authors show that this type of naive estimation gives unsatisfactory results and propose to replace it by a full Bayesian approach. As always in Bayesian approaches, the choice of the prior distribution (here of the age at death) is extremely important. This aspect is discussed in detail in the article. As a consequence, the article is a perfect illustration of the iterative collaborative design described above: the construction of the prior distribution is the result of a series of tests combining expertise from both fields (statistics and paleodemography).

In a similar way, P. Lanos and A. Philippe study dating problems with Bayesian approaches in "Hierarchical Bayesian modeling for combining dates in archeological context ". The research issue consists of combining different measurements in order to estimate the date of an archaeological event, especially when individual measurements provide date estimates via completely different methods (radiocarbon dating, luminescence dating and archaeomagnetic dating) with different calibration curves and noise levels. The authors introduce a complex hierarchical Bayesian model which takes into account noise at different levels: at date level (intrinsic noise), then at the calibration curve level (which models the relationship between a date and a physical measurement) and finally at the physical measurement level itself. As the model is complex, the authors use a rather sophisticated MCMC algorithm in order to estimate the posterior distribution of the parameters (in particular, conditional distributions cannot be simulated directly). To allow archaeologists to use this model in an autonomous way, the authors have implemented an open source multiplatform tool, ChronoModel[7]. The software completely hides the complexity of the inference for casual users (experts can still monitor the behavior of the MCMC algorithm via classical diagnostic plots). ChronoModel includes a graphical and interactive display of the results, a solution that seems particularly adapted in this context.

We would like to emphasize that the growing sophistication of statistical models used in the

[7] https://chronomodel.com/

 ISSN: 2102-6238

humanities, as illustrated in the previous two articles, might make them very difficult to master by humanity scholars. It seems quite unrealistic to assume that those researchers will be able in the near future to develop by themselves complex models as the one just presented (in fact MCMC algorithms are not mastered by all statisticians, for instance). However, as shown for instance in the article by A.-S. Bruno, the availability of modern methods in standard statistical software enables an autonomous use of those methods by humanity scholars (in Bruno's paper, a R package for multilevel Cox models was used). Standard but yet sophisticated models such as the multilevel Cox models are integrated in many statistical software, even ones that are easier to use than R. However, this is not the case of recent methods. As the consequence, software development initiatives such ChronoModel are extremely important: they lower substantially the effort needed to use very recent models by hiding the subtlety of the underlying statistical inference.

The main statistical innovation in the previous two articles was the use of a Bayesian approach. The Bayesian approach remains underrepresented in statistics, where frequentist methods tend to dominate, especially in curricula designed for non specialists. On the contrary, in "Markov and the Duchy of Savoy: segmenting a century with regime-switching models ", J. Alerini, M. Olteanu and J. Ridgway use a standard frequentist inference to estimate the parameters of two original generative models. Those models are conceived as ways to analyze integer valued time series with numerous null terms. The time series under study is obtained by counting the number of legal texts about military logistics published by month from 1559 to 1661 in the Duchy of Savoy. It is known from previous historical studies that this series should display a modification in its general behavior around 1610 when the Duchy enters into a cycle of wars. However, classical change point detection methods and regime switching methods do not provide satisfactory results on this time series. The limitations of those classical models motivate the co-design of new regime switching models by statisticians and historians. Here, the specificity of the data manifests by the excess of null values as compared to more classical data: the models are specifically designed to take this excess into account. The results obtained with those new models are particularly relevant on a historical point of view, especially as they provide complementary insights: one model emphasizes the gradual change in the Duchy of Savoy while the other pinpoints some specific historical periods.

This special issue concludes with "Génération de graphes aléatoires par échanges multiples d'arêtes " by L. Tabourier, J.-P. Cointet and C. Roth. While all the other papers focus on historical or archaeological data, this article originates from a collaboration with sociologists. It concerns social interactions modeled by mathematical graphs (social networks in the proper sense as opposed to social web sites). Graphs are complex and rich mathematical structures. As a consequence, it remains difficult to decide whether a particular (sub)structure observed in a graph is significant and gives important insights on the interactions modelled by the graph or if it should be considered as a consequence of well known structural constraints. Indeed simple structural constraints (parenthood constraints for instance) can "generate" substructures that could be attributed to some other causes while they are in fact present (to some degree) in all graphs that respect the structural constraints. The authors propose in the paper a general MCMC like method that can generate random graphs which fulfill arbitrary structural constraints. This can be used

 ISSN: 2102-6238

both to identify (sub)structuring constraints and to conduct empirical tests. The method is applied with success to the analysis of scientific collaboration networks (co-authoring networks).

The panorama offered by this special issue is by essence limited. For instance, it covers a very limited selection of the disciplines of the humanities as its main focus is on historical sciences (with an extension to sociology). It concerns mostly temporal problems and numerical data. To give a single example of missing topics, among a very long list, we may mention text data analysis which is completely absent from this special issue while it represents a very important subject for humanities related data. It might have been interesting, for instance, to show how relatively new models such as the topic models [8] are gradually integrated into the standard practice of humanity scholars.

Despite its limitations, this special issue illustrates several general aspects of the quantification movement in the humanities. We shall first point out that none of the articles of this special issue deals with large scale data (and thus none of them is even remotely connected to the "big data" problems). On the contrary, one of the main problems faced by humanity scholars is the relative scarcity of the data. This is particularly crucial in historical sciences where acquiring new data is frequently impossible. As a consequence Bayesian approaches are particularly relevant. In addition, one should use all the available data, even when they exhibit statistical dependencies: multilevel mixed models are very important in this context.

We shall also emphasize the crucial role played by software implementations as necessary (but not sufficient) conditions for autonomous research practices by humanity scholars. Indeed data from the humanities need complex and recent statistical models. In the early phases of their development, those models can only originate from inter-disciplinary collaborations. In the best cases, those collaborations produce easy to use software (such as ChronoModel already mentioned) which will in turn allow humanity researchers to apply these sophisticated model relatively autonomously.

We thank the authors for their contributions to this special issue and for their patience. Editing this special issue took us more time than we naively expected, as we vastly underestimated the constraints associated to an inter-disciplinary work. We thank also the reviewers for their very useful work as well as for the additional efforts induced by the inter-disciplinary nature of the papers. Finally we thank the editor-in-chief, Gilles Celeux, for his constant monitoring, his advises and his support in the production of this special issue.

[8] http://www.cs.columbia.edu/~blei/topicmodeling.html

 ISSN: 2102-6238