Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Wage Imputation with Deep Learning in the French Labor Force Survey

Prepared by Damien Babet, Insee, France

I. Introduction

A. Non-response to the wage question in the LFS

1. The French Labor Force Survey (LFS) surveys about 100000 in about 60000 dwellings each quarter. The LFS is a rotating panel: we survey dwellings for six consecutive quarters, and one sixth of the sample is renewed each quarter. Like other national LFS, it is the source for the ILO unemployment rate, and for numerous other publications on employment and working conditions. Since 1982 in the French LFS, employees are asked for their monthly wage, first in income bands up until 1989, then for the exact amount since 1990. Self-employed still answer with (annual) income bands categories. We collect wage and income for each individual on the first interview (usually the first quarter a dwelling is in the sample) and on the last dwelling interview.

2. These variables have obvious drawbacks. Non-response is high, amounts are generally rounded, all respondents might not equally understand the exact concept of wage used in the survey, and there is a strong suspicion of underreporting of higher wages. Still, we can check variable quality at aggregate level and at an individual level through file matching with fiscal and administrative data, and the results are reassuring.

3. While the primary sources in France draw from administrative registers for wage statistics and from fiscal data for income statistics, econometric studies on wages often need to rely on the numerous LFS variables missing from other sources. Notably, studies of wage inequalities and discriminations by gender, age or ethnicity usually make use of such LFS variables as households' composition, childcare, social and national origins of respondents and their parents, diplomas and experience [Aeberhardt et al, 2010, Boutchenik, Lê, 2017]. It is thus of great relevance to enhance the quality of the published wage variables, and more specifically to implement the best possible imputation method.

4. The wage question is the one that draws this highest rate of non-response in the French LFS. About 19% (and rising) of respondents do not give an amount, including blank answers and explicitly coded refusal and "do not know" answers. There is then a follow-up question asking for an answer in income bands categories: a bit less than 4% (and rising) of respondents do not answer by income bands either. For this study, we do not use income bands answers for exact amount imputation; however, they do provide a direct indication of the link between non-response and wage amount: income bands answers are more frequent for respondents in higher wage categories. Consequently, non-response is a likely source of bias, and a good imputation method seems necessary. Currently, about 20% of wages values in the French LFS published files are imputed (total non-response, income bands answers and outlier correction) thanks to a composite method using traditional wage equations for outlier detection and total non-response imputation, and an additional step of random draw in a conditional distribution to take into account income bands answers.

B. Machine learning could be useful for imputation

5. Machine learning (ML) methods could improve on the current imputation method. We define ML as all predictive algorithms whose predictive performance cannot be rigorously computed but has to be empirically assessed on a test data set, distinct from the train or learning data set [Breiman, 2001]. ML algorithms can learn much more complex representations of the data-generating process than traditional models, starkly reducing the modelling error (or "bias" in the variance / bias dilemma framework), at a cost. They tend to be less mathematically tractable, prone to overfitting, greedy in data and computing time, costly to fine-tune, and not "explainable". For this study, deep learning (DL) methods are tested. They are described below (§ 19). Among ML methods, they show the most potential in important fields: computer vision, natural language processing (NLP) and complex games problems. The general drawbacks of ML are also particularly strong with DL.

6. There are several reasons in favour of using ML for wage imputation. First, the problem of imputation being purely predictive, the lack of explainability of ML is less of an obstacle. Second, wage is a non-linear phenomenon, with a strong exponential link with variables such as years of study and experience, but also non-linear relations and interactions with individual characteristics (gender and number of children, occupation and industry, etc.). Third, if imputed wages are used in econometric estimations, it is necessary (but not sufficient) that the imputation model is richer than the econometric model. If not the case, the imputation model will artificially generate spurious correlations. Finally, there is an abundance of relevant variables in the LFS that are seldom used, including for imputation, and are easier to include in an ML method than in a traditional model.

7. On the other hand, ML might fail to outperform a more traditional imputation method here for two main reasons: lack of data, and the fact that wage equations are a well-known field of research with good predictive results. On the contrary, we show that DL improves on traditional, Mincer type wage equations on purely predictive grounds, and that a large-scale survey such as the LFS provide enough data for DL to train.

8. In the second part of this paper, we describe the wage imputation model currently in production in the French LFS, which we use as a benchmark. The third part introduces nomenclature vectorization, a treatment of high cardinality qualitative variables with DL that allows for including in the imputation model more detailed variables. The fourth part describes the setting, tuning and training of the DL imputation model *per se*. The fifth part compares predictive performance of the various models on two main test datasets: the first quarter of 2018, and the group of all respondents born in March. Finally, the sixth part concludes.

II. Current imputation model

9. The current imputation model is called SALRED, for "salaire redressé" or "corrected wage". It applies a first correction depending on the gross or net value the respondent chose to declare, and the annual or quarterly bonuses respondents mentioned in other questions. Then, for each quarter, the population is divided into 5 groups (white and blue collars versus intermediate and managerial occupations, crossed with men versus women, and a fifth group for atypical job contracts, precarious occupations and minors), and a log-wage equation is fitted for each group. The log-wage equations include the following variables: sex, hours usually worked, occupation, industry branch, experience on the job, type of employer and number of employees, age, diploma, bonuses, living in Paris, multiple jobs, occasional job and public sector. 107 dummies encode these variables, without interactions.

10. Values with a residual above 9 times the mean square error of the fitted equation are deemed outliers and put to blank, before a new fit of the equation. Missing values (outliers, non-response and income bands answers) are then simulated with the generation of a normal residual conditional on the min and / or max values of the (log) income band, when it is known. When there are missing values in the explanatory

variables, the current model imputes the log-wage as either the middle of the log income band (when known) or the mean log-wage of the group.

11. The method is currently implemented in SAS and runs in production since 2014 in its current form. It computes in several minutes for the quickest version, depending on the simulation method for the conditional law. There is a trade-off between numerical accuracy, especially when the income band is far from the mean of the estimated distribution, and speed.

12. To our knowledge, this model's predictive performance was never tested. We estimate it on the first quarter of 2018 through cross-validation. We randomly divide the quarter file in 5 equal dataset, taking only observations for which we know the wage amount. Each of these groups' wages are put to blank in turn and the four remaining are used to fit the equations and run the SALRED imputation algorithm. We then compute the R2 of the imputed log-wage as a predictor to the log-wage actually declared by respondents. This yields a result of 0.43. When computed on wages (and not log-wages), the R2 goes down to 0.19. These values are our benchmark.

III. Nomenclature vectorization

13. One obvious way forward for the imputation model is to include classifications that are more detailed. For instance, the occupation variable in the French nomenclature has about 500 categories. Only 24 dummies aggregate these categories in the current SALRED model. Other high-cardinality nomenclatures include industry branch, level of diploma, field of diploma, legal status of the employer firm, region of the job and region and country of birth. These categories could be directly included in a wage equation as dummies, but the high number of variables would increase the risk of overfitting, and some sort of regularization might be called for. Variable selection methods are one possibility, even if the hypothesis of sparsity for such variables is not intuitive. Moreover, interaction effects between these variables are highly likely, and the number of dummies would explode if all possible interactions were included.

14. It is possible to include a very high number of dummies as inputs for most ML algorithm, and usually not necessary to explicitly code for interactions. ML algorithm usually allow for interaction effects between the inputs. For DL, however, a big input vector significantly increases computational costs (since the number of parameters is roughly quadratic in the number of neurons on each layers) and dummies are not efficient inputs for DL training.

15. Encoding these high-cardinality qualitative variables into a smaller number of real numbers (or "vectorization") gives a usable input for both wage equations and DL. Depending on the encoding strategy, it likely conveys information on relations between categories. For wage equations, it reduces the risk of overfitting. For DL, it increases speed of computation and training [Cerda, Varoquaux, 2019].

16. Many vectorization strategies are possible, starting with principal component analysis (PCA). In the same family of dimension reduction, we used word2vec [Mikolov et al., 2013], a NLP DL method. It uses the co-occurrence of words in the same sentences as input data to encode each word by training a neural network (with only one hidden layer and one non-linearity) to predict neighbouring words. Here, we use nomenclature categories as words, and nomenclatures as vocabularies, and we build the "sentences" as, for instance, co-occurrence in the same households, or between parents and children, or siblings, or spouses, or during the course of the six quarters of interviews. This has the benefit of encoding additional information about the relations between the categories that comes from the dependencies between observations and is lost when we consider each observation in isolation.

17. We used three encoded nomenclatures in this study: occupation, industry branch, and region (there is ongoing work to encode other variables). For occupation, we built "sentences" using three strategies: co-occurring occupations in the same household (when there is more than one), couples of current and last occupation of individuals (when they differ), and different occupations occurring in the course of the 6 interviews for each individual (when it changes). Aggregating data since 2013, we finally used 180,000 sentences counting 370,000 raw words, from a vocabulary of 486 words (ie. occupations). We plan to build a bigger training set by further exploiting the LFS. We used the Python package Gensim [Rehurek, Sojka, 2010] to implement word2vec, finally keeping a light encoding of dimension 4 **(Figure 1)**. For (French)

regions, we used all co-occurring region of birth of respondents of the same household and their parents as sentences, yielding 3,190,000 raw words in 530,000 sentences, from a vocabulary of 111 regions, encoded in 3 dimensions. Finally, we encoded industry branches in 10 dimensions, using a corpus from 2009 onward of 1,750,000 raw words in 870,000 sentences, from a vocabulary of 905 activity sectors.

18. Nomenclature vectorization is a promising road on its own right, not only as pre-processing for imputation algorithms. It can be used for text variables (we tested it successfully on first names, on field collection files of the LFS), it provides a data exploration and visualization tool, allows for an easier computation of distance between observations, etc. More generally, it speaks to the adaptation of ML methods to survey data. The context of ML development is often one of raw, homogenous, real-numbered big data, whereas survey data is smaller, cleaned, structured, and highly heterogeneous. Vectorization could be one tool to bridge these two worlds.

IV. DL setting and tuning

19. Deep learning is the name of neural networks models with hidden layers [Goodfellow, Bengio, Courville, 2016]. A "neuron" in a neural network is a simple, usually non-linear function (or "activation function") of a linear combination of the input variables. The parameters of the neural network are the coefficients of this linear combination. A typical dense layer comprises many neurons independently connected to the same input, each with its own parameters. When layers are stacked in a simple, feed-forward neural network, the output of all neurons of one layer are the input of each neuron of the next layer. The first layer receives the input data; the last layer outputs the result, and intermediate layers are dubbed "hidden layers". For this study, we only used feed-forward neural networks with dense layers and one to eight hidden layers.

20. The neural network is trained via backpropagation of stochastic gradient descent, meaning a loss function of the neural network prediction and ground truth is computed for randomly selected batches of training data. For each batch and before the next, the gradient of this loss is computed and then propagated backwards through the network as a gradient of each parameter. Each parameter moves in the direction of the reduction of the loss, by a quantity depending on the learning-rate. The entire training dataset is used during an "epoch", and training might need many such epochs. The process is stochastic, and offers no convergence certainty.

21. This quick presentation shows there are many hyper-parameters to tune when setting up neural networks. To name a few: the number and size of layers, the activation function, the loss function, the learning rate, the batch size, the number of epochs, and more elaborate aspects of the algorithm (regularization parameters and dropout, dynamic learning rate functions, etc.). Although there are a few rules of thumb and some preliminary results in the literature, tuning these hyper-parameters remains an open question. At scale, DL tuning usually relies on computationally heavy automated exploration of the hyper-parameter space. In a more experimental setting, like the present one, this exploration is done by hand.

22. Our data comprises 1,220,000 observations drawn from the LFS between 1993 and the first quarter of 2018. Included variables are sex, date, date of birth, place of birth and place of birth of both parents, immigration status, education level, date of higher diploma, experience on the job, occupation and occupation of both parents, industry branch, weekly work hours (usually and during the reference week), part-time, type of contract, type of employer, existence of bonuses. We consider missing data as one more category for categorical variables, and impute with the mean for quantitative variables, with a dummy signalling the missing data [Josse et al. 2019]. The wage variable is similar to the SALRED reference variable, corrected for gross or net declaration and annual bonuses. We further correct older wage data for a specific error: typing errors of refusal and "does not know" codes appearing as wage such as 999,999 or 99,999 euros (or francs). This problem is absent from the 2018 data and does not affect the SALRED benchmark. The final training variable is the log-wage.

23. We then extract two test datasets: people born in March, and respondents from the first quarter of 2018. We further randomly split the remaining set during each training between a training set of 995,000 lines and a validation set of 110,000 lines (90%/10%). After pre-processing, the input vector is of dimension 215. Our preferred DL model has 7 hidden dense layers of sizes [500, (dropout), 100, 50, 20, 8, 4, 2], with

one dropout layer with a 0.5 dropout rate. It uses the SELU activation function [Klambauer et al. 2017] – except for the output layer that is linear, a mean squared error loss, the Adam optimizer [Kingma, Ba, 2014], a batch size of 200, and 50 epochs of training. One epoch typically needs about 40 seconds to compute with our settings (one working station, no parallelization), which impedes exploration of the hyper-parameter space.

V. Model comparisons

24. The Mincer equation [Mincer, 1974] is one of the most widely used model in empirical economics in its classical form:
$logy = logy_0 + rS + \beta_1 X + \beta_2 X^2$
Where y is earnings, S the years of schooling and X the potential years of labor market experience. Its success might come from several factors: human capital theoretical foundations for the use of the log-wage and the years of schooling variable, and a parsimonious model of the effects of age and experience fitting well with empirical data. It remains highly accurate in most datasets, even in its simplest form [Lemieux, 2003]. We used this foundation as a first step for improving on the SALRED benchmark, by introducing the potential experience and squared potential experience that are not in the SALRED model.

25. Two other Mincer-type models were tested. One that introduces social origin as supplementary explanatory variables that are not in SALRED (immigration status, father and mother profession, and a more detailed diploma variable), and another one that includes vectorised nomenclatures. This last model is arguably already a DL model, since vectorization uses a (shallow) DL model, word2vec. A Mincer-type model including these complete nomenclatures as dummies would be an interesting point of comparison, but it proves too heavy for our computational resources and will need further work.

26. All Mincer-type models were estimated on the same training dataset than the DL model, and the same two test sets were used: the first quarter of 2018 (allowing for a direct comparison with SALRED) and the set of all people born in March in observations dating back from 1990. For each test sets, we provide the R2 results for the wage and the log-wage. All models are fitted on the log-wage.

27. **Figure 2** details the results. Mincer-type equations are a clear improvement on the benchmark (the R2 jumping from 0.4 to 0.6) and do not differ much with one another, suggesting the potential experience and quadratic potential experience are doing most of the work here. The DL model further enhances the prediction performances to 0.7. All models have inferior performance on the "born in March" test set, presumably for two reasons: the smaller 2018 test set might have less outliers, and the older data is of lower quality (because of missing explanatory variables, mostly). Finally, SALRED is not a good imputation model. While we are not yet sure why exactly, one early suspect was the outlier correction. On the contrary, it appears that without it, the model further underperforms, suggesting that outliers have a strong impact on the estimation. Anyhow, it shows that sophisticated, classical imputation models need to be evaluated on test sets and compared with alternative strategies.

VI. Discussion and future work

28. This imputation experiment confirms the pros and cons of ML, and especially of DL methods. Their additional modelling power allows them to learn more from the data, even in a problem space as well known as wage equations. The increase in predictive performance compared with models that are more classical is high enough to be relevant, considering the use of imputed LFS wage data in econometric works. However, the known problems of ML are also clear here: the algorithm takes some work to code and fine-tune. Its training, and its results, are stochastics, meaning it can take several tries to land on a well-trained model (on the validation set performance metrics). Finally, the model is not easily explainable. This lack of explainability is not obviously a problem, though. First, imputation does not necessarily need explanations, or interest parameters estimation. Second, the current imputation model SALRED, while classical in its idea, is too convoluted to be easily explainable in a practical sense. Third, to our knowledge, no one even tried. We suspect this is common for imputation models in production in public statistics settings. On the contrary, the implementation of ML methods forces us to evaluate performances on test datasets, a methodological improvement over the status quo.

29. In further work, we plan to add more variables to the models, and take into account the income bands answers. Obvious candidates for additional explanatory variables are the household composition variables, the detailed diploma level and field nomenclatures (vectorised), countries of origin for immigrants and descendant of immigrants (vectorised), detailed variables on working hours (such as work at night and on weekends, etc.), employment history in the last 12 month (a good addition to job seniority when it is short). More generally, all LFS variables could be included. We are currently working with a group of student on using auto-encoders to allow the computing of such a high number of variables, and partially automate the process. This is somewhat similar to recent work done by the British Office for National Statistics [Kaloskampis, 2019].

30. Income bands answers, when they exist, would likely greatly improve imputation performances. It remains to be tested if ML algorithms are able to learn enough from these bands so that they always predict a value inside the correct band, and, if not, what is to be done. Such a work would benefit from a more systematic comparison of various ML models, in addition to DL, and a measure of the variability of performance results, through cross-validation. Other ML models might also prove easier to fine-tune. Yet further work might include an exploration of explainable ML, for instance through importance measures [Boelaert, Ollion, 2018]. Another way forward is the growing literature on ML use for econometrics. Here, a predictor of wages is a stepping-stone for more ambitious models aiming at identifying specific causal effects. One such possibility would be to implement the classical month-of-birth instrumental variable to identify the causal effect of diploma on wages, with an ML predictor in the first stage

VII. References

Aeberhardt, Romain, et al. "Wages and employment of French workers with African origin." *Journal of population economics* 23.3 (2010): 881-905.

Boelaert, Julien, and Étienne Ollion. "The Great Regression." *Revue française de sociologie* 59.3 (2018): 475-506.

Boutchenik, Béatrice, and Jérôme Lê. "Les descendants d'immigrés maghrébins: des difficultés d'accès à l'emploi et aux salaires les plus élevés." *Emploi, chômage, revenus au travail* (2017): 21-33.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical science* 16.3 (2001): 199-231.

Cerda, Patricio, and Gaël Varoquaux. "Encoding high-cardinality string categorical variables". 2019. ⟨hal-02171256v4⟩

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Josse, Julie, et al. "On the consistency of supervised learning with missing values." *arXiv preprint arXiv:1902.06931* (2019).

Kaloskampis Ioannis, *Synthetic data for public good*, Office for National Statistics data science campus, 2019, https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

Klambauer, Günter, et al. "Self-normalizing neural networks." *Advances in neural information processing systems*. 2017.

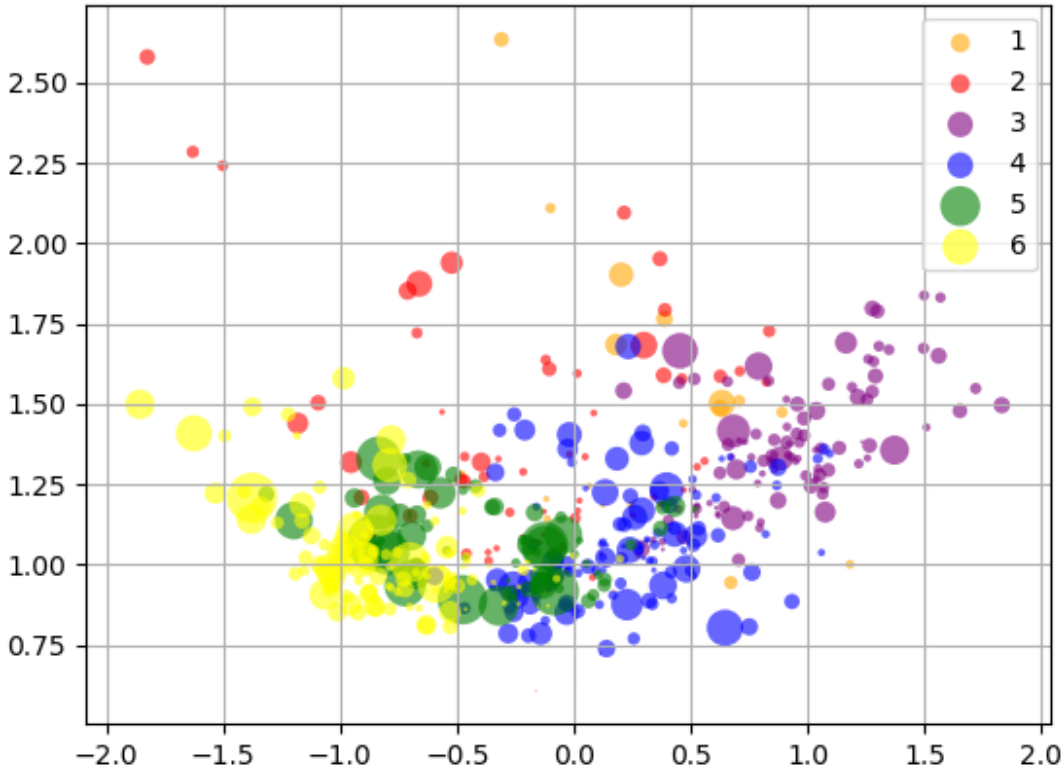
Lemieux, Thomas. "The “Mincer equation” thirty years after schooling, experience, and earnings." *Jacob Mincer a pioneer of modern labor economics*. Springer, Boston, MA, 2006. 127-145.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Mincer, Jacob. "Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2." (1974).

Rehurek, Radim, and Petr Sojka. "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 2010.

Figure 1: 486 occupations projected on dimensions 1 and 4 of a 4-dimensions encoding



Professional groups: 1. Farmers 2. Craftsmen, tradesmen, entrepreneurs 3. Managers and intellectual occupations 4. Intermediate occupations 5. Shop and service employees 6. Manual workers
Dimensions have no explicit meaning, the two dimensions shown are an arbitrary choice. Area is proportional to the number of observations for each occupation

Figure 2: Comparative R2 of wage imputation algorithms, on wage and log-wage, on two test sets

		Train set, log wage	Test set: Q1 2018		Test set: people born in march	
			Log wage	wage	Log wage	wage
Classic models	SALRED with no outlier correction	///	0.344	0.103	///	///
	SALRED	///	0.432	0.186	///	///
	Mincer equation with SALRED variables	0.545	0.597	0.327	0.549	0.217
	Mincer equation with additional variables	0.548	0.603	0.320	0.552	0.218
DL models	Mincer equation with vectorised nomenclatures	0.557	0.608	0.333	0.561	0.228
	DL		0.696	0.381	0.662	0.270