UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Robust regression, MissForest and calibration combined with non-linear optimization with constraints to impute VAT turnover

Jacques Saliba (Federal Statistical Office, Switzerland)

jacques.saliba@bfs.admin.ch

I. INTRODUCTION

1. The value-added tax (VAT) is used as a main source to allocate a yearly turnover to business units in Switzerland. A direct link can be established for the majority of the business units. However, for more than 40% of business units, no direct link exists due to an exemption from paying the VAT or because these units are a member of a VAT group that pays the VAT for all its members. Therefore, turnover has been imputed taking into account the business' activity sector and its number of employees, to mention just a few.

2. In section II, we present the main notions and the main framework of the imputation and distribution model. In section III, a linear robust regression by activity sector is used to model the turnover using the number of employees and total wage as independent variables. This model is used to impute the missing values for the units with more than 20 employees. For units with less than 20 employees, a MissForest algorithm is used to impute the turnovers with the help of auxiliary variables related to the total wages and the activity sector of the units. In section IV, we introduce a calibration method enhanced by a non-linear optimization with constraints approach to show how the firstly imputed values are modified in order to impute a turnover to members of VAT groups, conditioned on the total turnover of the group. In section V, we finally present the results of the main different methods analysed by comparing the imputed turnovers to the ones of the yearly survey of the production and value added statistics.

II. Initial dataset

1. For the year 2018, we have a total of 695119 business units that are considered. Among them, 312649 have missing turnovers and 7540 of them are part of a VAT group. For the year 2019, we have a total of 696153 business units that are considered. Among them, 308104 have missing turnovers and 7617 of them are part of a VAT group.
We have, for each business unit, the following variables that are used in the imputation model:

- Number of employees,
- Number of full-time equivalents,

- The classification of economic activities (NOGA) for business units: It is derived from the NACE european classification, both classifications being identical up to level 4,

- Customs data (import, export) in CHF,
- Total wages based on the old-age and survivor's insurance.

Turnovers of business units can be classified into the following categories:

(1) Original: The turnover is known for the considered year thanks to the VAT that was paid by the business unit.
(2) Partial: The cumulated turnover obtained from paid VAT returns is strictly less than the true turnover of a business unit for the considered year. This is the case if the VAT returns are not complete for example or if some of the activities of the business unit are not subject to VAT.
(3) Missing turnover and the business unit is not part of a VAT group: This is the case when for example the turnover of a business unit does not exceed 100'000 CHF for a year.
(4) Missing and the business unit is a VAT group member: In this case, the total turnover of the VAT group is known (partially or completely) and has to be distributed among all its members.

We focus in this paper on the methods that were developed to impute completely missing turnovers, that will be adjusted afterwards for business units being part of a VAT group so that the sum of turnovers within a group equals the known total turnover of the group.

2. A robust regression between the original turnovers and their corresponding turnover from the survey of the production and value added statistics (WS) for 2018 has $R^2 = 0.697$. This shows a good correlation between turnovers of VAT and those of WS. The fact that R^2 of this regression is not equal or closer to 1 is mainly due to slightly different definitions of the turnover, which will not be outlined here. The robust method was chosen in order to reduce the influence of outliers on the regression parameters estimates. The turnovers of the WS will be nevertheless used as a source of comparison in order to evaluate the performance of the different imputation methods that are presented in section III.
For 2019, a robust regression between the original turnovers and their corresponding turnover from the survey of the production and value added statistics (WS) for 2018 has $R^2 = 0.702$. This shows, as for 2018, a good correlation between turnovers of VAT and those of WS.

III. Imputation of totally missing turnovers

1. We present in what follows two classes of imputation methods applied and tested with several different parameter settings.

A. Imputation using regression methods

1. The idea of the first imputation step is to model turnovers, in each NOGA, through a robust regression, with respect to total wages and number of employees. More precisely, we define imputation classes based on the NOGA levels 2 to 5 in the following way:

- Consider the total number of "model" business units in each NOGA2, for which the turnover, the total wage and the number of employees is known. If a NOGA2 contains more than 600 model business units, we split it into up to potentially 9 NOGA3 (for example, the NOGA 26 is potentially split into 8 NOGA 3-levels 261 to 268).
- This procedure is repeated up to NOGA level 5 if possible. In the end, small NOGAs obtained (less than 30 model units) are regrouped on the next higher level of the nomenclature (for

example, NOGAs 0115, 0116 and 0119 contain very few model business units in our case, so they were regrouped together, leaving 0111, 0112, 0113 and 0114 individually separated).

The final NOGA or group of NOGAs obtained at the end of this process is called an imputation class. For each imputation class, we test two different regression models for the turnover y_i of a business unit i as a function of the number of employees x_i and the total wages of the company s_i:

(1) Robust linear regression: The turnover is modeled as a linear combination of x_i and s_i in the following way:

$$y_i = \alpha_I + \beta_I x_i + \gamma_I s_i + \epsilon_i, \tag{1}$$

where α_I, β_I and γ_I are unknown model parameters estimated for each imputation class I and ϵ_i the residual of the regression. *Proc Robustreg* from SAS was used with the MM method and default parameters.

(2) Robust logarithmic regression. The turnover is modeled in each imputation class I as follows:

$$\log(y_i) = \alpha_I + \beta_I \log(x_i) + \gamma_I \log(s_i) + \epsilon_i, \tag{2}$$

where α_I, β_I and γ_I are unknown model parameters estimated for each imputation class I and ϵ_i the residual of the regression. The same procedure with the same parameters as for (1) were used to adjust the model (2).

B. Imputation using MissForest

1. In subsection A, we presented two different regression models that were tested to model the turnover as a function of the number of employees and the total wages of the company. The quality of these models varies depending on the imputation class and on the set of model units that is used to estimate the regression parameters in each imputation class. The MissForest algorithm described in this section was considered to tackle these limitations.

B.1. *MissForest algorithm.*

1. The input data set contains the following auxiliary variables that will be used in the MissForest imputation:

- Number of employees,
- Number of full-time equivalents,
- Size classes of the number of employees given by $[1,3), [3,5), [5,10), [10,15), [15,20), [20,30), [30,+\infty)$
- NOGA (50 modalities),
- Customs data (importations, exportations) in CHF,
- Total wages based on the old-age and survivor's insurance,
- Quantiles $0, 1, 5, 10, 25, 75, 90, 95, 99$ and 100 and the average of total wages, in each NOGA2.

2. Missing values in customs data are imputed to 0, whereas missing values of total wages are imputed using MissForest during the imputation of turnovers. The main steps of the MissForest algorithm used (Stekhoven and Bühlmann [2012], based on the random forest algorithm Breiman [2001]) are:

(1) Make initial guess for missing turnovers and missing wages,

(2) Since the total wages variable has less missings than turnovers in our case, the algorithm starts imputing the total wages variable first. From the set of observations with known total wages, draw n observations with replacement, where n is the total number of observations of the input data set,

(3) Create a decision tree with these drawn n observations: At each splitting, m auxiliary variables are randomly chosen as candidates to define the splitting criterion. The variable that divides the data in the most optimal[1] way is chosen to be used for the split. In our case, m has been set to 10,
(4) Repeat this procedure for a number of trees $ntree = 20$. The missing total wages of a unit is then imputed by taking the average of the average of total wages in each of the 20 leaves to which the unit belongs, one leaf for each decision tree,
(5) Apply the same procedure for missing turnovers, by fitting a random forest on the observed turnovers and using total wages (with the new imputed values) as well as the other auxiliary variables,
(6) Steps (2) to (5) are repeated until a stopping criterion is met (when the difference between the imputed dataframes of 2 successive iterations increases for the first time).

B.2. *Different imputation models using MissForest.*

1. We present in what follows 8 imputation models that were tested, using MissForest:

(1) RF_04: In the first step, we use a robust regression in each imputation class to model turnovers as explained in section A. For imputation classes with $R^2 > 0.4$, we use the regression model to impute missing turnovers. For the rest of imputation NOGAs, we use MissForest as described before in order to impute the missing values.
(2) RF_06: Same as RF_04, but with MissForest applied for imputation classes for which $R^2 \leq 0.6$ instead of ≤ 0.4.
(3) RF_log04 : Same as RF_04, but replacing the linear regression by a robust logarithmic regression as presented in section A.
(4) RF_log06 : Same as in 2), but replacing the linear regression by a robust logarithmic regression as presented in section A.
(5) RF_B3: The linear regression is applied only for business units with number of employees > 3. On the other hand, the missing turnovers of business units with number of employees ≤ 3 are imputed using MissForest.
(6) RF_B20 : Same as the previous model, but we set the threshold of number of employees at 20 to decide whether the missing turnover is imputed using the linear regression or the MissForest algorithm.
(7) RF_logB3 : Same as RF_B3, but applying a logarithmic regression instead of a linear regression.
(8) RF_logB20 : Same as RF_B20, but applying a logarithmic regression instead of a linear regression.

IV. Distribution of turnovers within VAT groups

1. Some business units are part of a VAT group so that the total VAT is paid by the group head unit for all the group members. In this case, only the total turnover of the group is known, and so we need to estimate the turnover of each member of the group, under the condition that the sum of these estimations equals the total turnover of the group.

2. The model for estimating turnovers of business units members of a VAT group can be divided into 4 steps:

[1] After determining an optimal threshold by minimizing the sum of squared residuals (or Gini impurity for categorical variables), the chosen variable is the one for which the optimal threshold has the least sum of squared residuals.

(1) In the first step, turnovers of business units and members of VAT groups are imputed as described in section B.2, considering these units as individual business units with missing turnover. The choice of the model used, among the 8 presented models, is discussed in section V. We write $y_i'^{(2)}$ for this initially imputed turnover of a business unit i.

(2) In the second step, a winsorization is applied to the imputed turnovers as follows:

$$\text{If } y_i'^{(2)}/x_i > prod_{99} \Longrightarrow y_i^{(2)} := x_i \times prod_{99},$$
$$\text{If } y_i'^{(2)}/x_i < prod_1 \Longrightarrow y_i^{(2)} := x_i \times prod_1,$$

where x_i is the number of employees, $prod_1$ and $prod_{99}$ are respectively the 1st and 99th quantiles of the productivity per employee (= turnover/number of employees) in the NOGA of level 2 corresponding to unit i, based on not imputed units that are not part of VAT groups, and $y_i^{(2)}$ the new imputed turnover for i.

(3) A calibration method (see Deville and Särndal [1992]) using a linear truncated distance is used in order to modify the already imputed turnovers of group members so that the sum of the modified turnovers equals the total turnover of the group. This calibration method can be summarized as follows: Assign initial weights equal 1 to each business unit of a VAT group G and let $z^{(1)}$ be the total turnover of the group and $z^{(2)}$ the sum of the imputed turnovers of the group members. We denote by $r := \frac{z^{(1)}}{z^{(2)}}$ the ratio between these two quantities. The aim of this method is to find weights g_i's as close as possible to the initial weights (= 1 in our case) such that

$$\sum_{i \in G} g_i y_i^{(2)} = z^{(1)}. \tag{3}$$

In this case, $g_i \times y_i^{(2)}$ is the new imputed turnover for the unit i. More precisely, for a VAT group G, the goal is to get g-weights satisfying (3), and minimizing

$$\sum_{i \in G} D(g_i, 1), \tag{4}$$

where the pseudo-distance $D(.,.)$ is a non-symmetrical function that is given by

$$D(a,b) = \begin{cases} \frac{(a-b)^2}{2b} & if\ Lb < a < Hb, \\ \infty & otherwise, \end{cases}$$

where L and H are lower and upper bounds that are fixed, that guarantee that $Lb_i < a_i < Hb_i$ for all i. Since in our case, initial weights $b_i = 1$, we deduce that the weights g_i will satisfy (3) while minimizing the sum (4) and they will be bounded between L and H.
Equation (3) cannot be satisfied if all g_i are strictly smaller or strictly larger than r, the ratio between the total turnover of a VAT group and the sum of the imputed turnovers of its members. For this reason, the lower and upper bounds of the calibration for each group are chosen such that $\{1, r\} \in [L, H]$. The choice of these bounds for each VAT group is as follows:

$$\text{If } r < 1 \text{ then } L = \min(0.01, r^5) \text{ and } H = \max(100, 1/r^5),$$
$$\text{if } r \geq 1 \text{ then } L = \min(0.01, 1/r^5) \text{ and } H = \max(100, r^5).$$

The distributed turnover of a business unit i member of a VAT group after this calibration step is then given by $y_i^c := g_i \times y_i^{(2)}$.

(4) The aim of the last step is to adjust the already distributed turnovers in order to satisfy some productivity lower and upper bounds. For this, we compute first the pairs of percentiles $\{1, 99\}, \{2, 98\}, \{3, 97\}, \{4, 96\}$ and $\{5, 95\}$ of the productivity of units belonging to the same NOGA level 2 and the same size class of the number of employees.

Using the package NlcOptim in R, we test whether it is possible to modify the calibration weight g_i of a unit i into a weight g_i' such that

$$p_5 \leq \frac{g_i' \times y_i^{(2)}}{x_i} \leq p_{95},$$

where x_i is the number of employees of the unit i and p_5, p_{95} are respectively the 5% and 95% percentiles of the productivity of all units belonging to the same NOGA2 and the same size class of employees as the unit i. This is applied to all group members of a VAT group and we test if a solution exists such that the sum of the modified distributed turnovers is still equal to the known turnover total of the group, denoted by $z^{(1)}$.
If no solution could be found, the method is tested again with percentiles pair $\{4, 96\}$ as lower and upper productivity bounds. This procedure is reiterated (as long as no solution is found) for productivity percentiles pairs $\{3, 97\}, \{2, 98\}$ and $\{1, 99\}$.
If a solution is found for a VAT group, the resulting g''s will be used instead of the g-weights of the calibration, to distribute total turnover among the members of the group, that is, the new distributed turnover of a unit i will be $g_i' \times y_i^{(2)}$.
It is in general not possible to find a solution for all the VAT groups. Whenever a solution is not found for a particular group, the distributed turnover of a unit i that is member of this group is given by $g_i \times y_i^{(2)}$ as calculated in step 3, hence ignoring the productivity boundaries.

V. Results

1. We present in this section the results of the different models of section B.2 and the procedure for distributing total turnover of groups among their members described in section IV.

2. As mentioned beforehand, in the case of MissForest, the number of trees is fixed at 20 and the number of auxiliary variables randomly drawn at each splitting to decide the splitting criterion is fixed at 10. The linear and $\log - \log$ models without MissForest are simply denoted by lin and $\log - \log$ in what follows.

3. We denote by Old-imp the basic imputation model where a robust linear regression model is applied for NOGA2 only (or groups of NOGA2 when a NOGA2 is too small) using only the number of employees as auxiliary variable and where the distributed turnovers in VAT groups are obtained by multiplying the imputed turnovers by the ratio r between the total turnover of the group and the sum of imputed turnovers of its members, that was introduced in section IV. We also denote Old2345 the same model as Old-imp except that levels 3, 4 and 5 of NOGAs are considered in the regression model, as explained in A. Comparisons between these two old models and the new versions that use calibration are presented thereafter for the year 2018 using the total wages of 2018 as an additional auxiliary variable (Table 1), for turnovers of the year 2019 with the total wages of 2019 (Table2) and finally for the turnovers of 2019 with the total wages of 2018 as auxiliary variable (Table3). The two first situations give us an idea of the results in the ideal case where the total wages of the same year are available and are used in the imputation model, whereas the third case gives an idea of the expected results for production, as the total wages are only available with a lag of about 15 months.

4. These comparisons are made between original, imputed and distributed VAT turnovers and their corresponding turnover from the survey of the production and value added statistics (WS) for the years 2018 and 2019. The WS is a yearly survey that concerns around 14000 business units and its definition of turnovers is slightly different from that of the VAT, but is nevertheless strongly correlated to it as shown in section II. The tables Table1, Table2 and Table3 show the R^2 of robust regression between the VAT turnover (original, imputed or distributed) and the turnover of the surveys WS2018 or WS2019 (depending on the year considered for the VAT turnovers).

TABLE 1. R^2 between distributed/imputed/original VAT turnovers and WS turnovers for the year 2018 with total wages of 2018

	distributed	imputed	original	ratio-distributed	ratio-imputed
Old-imp	0.279	0.236	0.697	0.400	0.339
Old2345	0.270	0.297	0.697	0.387	0.426
Lin	0.364	0.305	0.697	0.522	0.438
Log-Log	0.358	0.359	0.697	0.514	0.515
RF_0.4	0.320	0.263	0.697	0.459	0.377
RF_0.6	0.350	0.240	0.697	0.502	0.344
RF_B3	0.366	0.336	0.697	0.525	0.482
RF_B20	0.364	0.315	0.697	0.522	0.452
RF_LogB3	0.384	0.352	0.697	0.551	0.505
RF_LogB20	0.355	0.239	0.697	0.509	0.343
RF_Log0.4	0.358	0.261	0.697	0.514	0.374
RF_Log0.6	0.327	0.238	0.697	0.469	0.341

TABLE 2. R^2 between distributed/imputed/original VAT turnovers and WS turnovers for the year 2019 with total wages of 2019

	distributed	imputed	original	ratio-distributed	ratio-imputed
Old-imp	0.283	0.255	0.702	0.403	0.363
Old2345	0.281	0.344	0.702	0.400	0.490
Lin	0.366	0.355	0.702	0.521	0.506
Log-Log	0.363	0.395	0.702	0.516	0.562
RF_0.4	0.333	0.291	0.702	0.475	0.414
RF_0.6	0.371	0.296	0.702	0.529	0.422
RF_B3	0.372	0.343	0.702	0.529	0.489
RF_B20	0.386	0.362	0.702	0.550	0.516
RF_LogB3	0.373	0.340	0.702	0.531	0.485
RF_LogB20	0.373	0.273	0.702	0.531	0.388
RF_Log0.4	0.355	0.327	0.702	0.505	0.466
RF_Log0.6	0.335	0.319	0.702	0.477	0.455

TABLE 3. R^2 between distributed/imputed/original VAT turnovers and WS turnovers for the year 2019 with total wages of 2018

	distributed	imputed	original	ratio-distributed	ratio-imputed
Old-imp	0.283	0.255	0.702	0.403	0.363
Old2345	0.281	0.344	0.702	0.400	0.490
Lin	0.367	0.338	0.702	0.522	0.481
Log-Log	0.364	0.3829	0.702	0.518	0.545
RF_0.4	0.344	0.312	0.702	0.489	0.444
RF_0.6	0.367	0.314	0.702	0.523	0.447
RF_B3	0.372	0.317	0.702	0.529	0.452
RF_B20	0.385	0.337	0.702	0.549	0.480
RF_LogB3	0.387	0.334	0.702	0.552	0.476
RF_LogB20	0.393	0.232	0.702	0.560	0.331
RF_Log0.4	0.365	0.302	0.702	0.519	0.431
RF_Log0.6	0.334	0.300	0.702	0.476	0.427

5. The quality of imputed turnovers increases as NOGAs of level 3, 4 and 5 are considered in the robust regression step and also after the use of total wages as an auxiliary variable. The method RF_B20 gives consistent good results for 2018 and 2019 for both imputed and distributed turnovers. This is also shown in the following plots that compare the robust regression with WS for the imputed and distributed turnovers for $Old - imp$ and RF_B20.

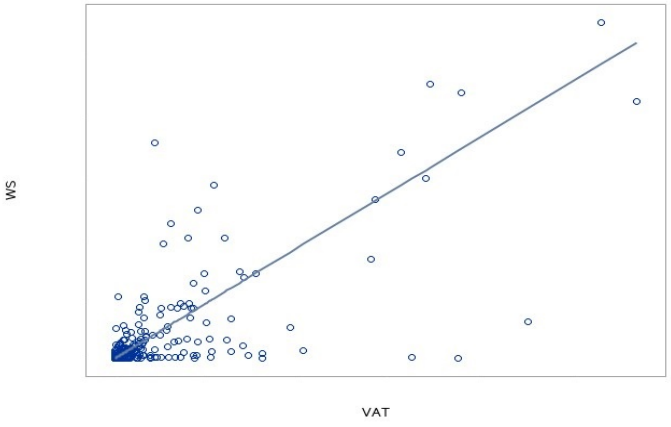

FIGURE 1. Robust regression between WS and the VAT distributed turnovers of Old-imp

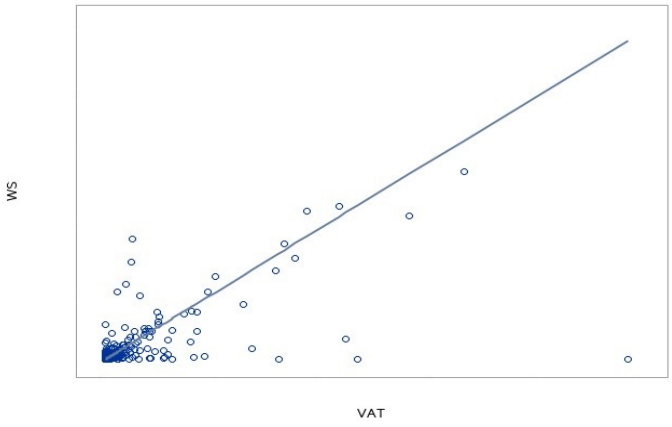

FIGURE 2. Robust regression between WS and the VAT distributed turnovers of RF_B20

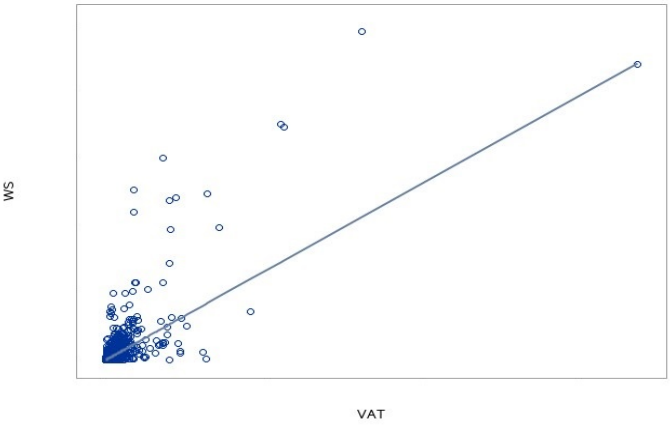

FIGURE 3. Robust regression between WS and the VAT imputed turnovers of Old-imp

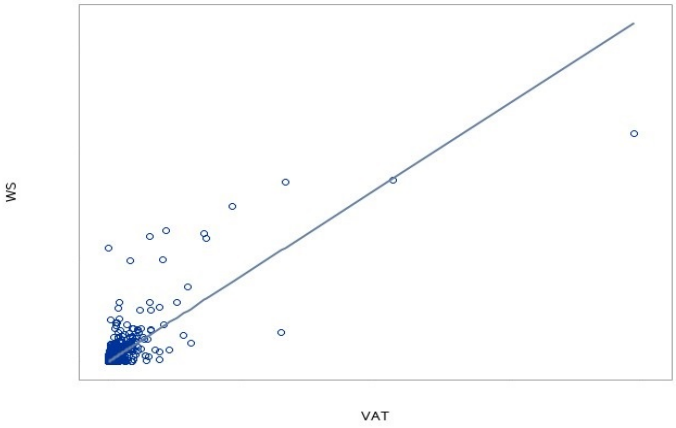

FIGURE 4. Robust regression between WS and the VAT imputed turnovers of RF_B20

VI. Conclusions

1. As shown in section V, the use of NOGAs of level 3, 4 and 5, the use of total wages as an additional auxiliary variable as well as the use of MissForest algorithm to impute a part of missing turnovers seems to enhance the quality of the imputation model. In the case of VAT groups, a calibration method is used to adjust in a minimal way the already imputed turnovers in order to obtain the desired total turnover of the VAT group ensuring equality with the distributed turnovers of its members. This step is further enhanced by adjusting the distribution weights, when possible, in order to satisfy productivity bounds. This procedure of distributing the total turnover within a VAT group to its members seems to be of higher quality that the simpler method of giving the same weight $r = \frac{z^{(1)}}{z^{(2)}}$ (as explained in V) to all members. Several potential improvements to the imputation and distribution procedure can still be explored, by adding further potential auxiliary variables to the regression and the MissForest steps, tuning the MissForest parameters and sharpening the choice of units for which this algorithm will be applied and by using potentially past years VAT data to improve the distributed turnovers in particular.

References

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Daniel J. Stekhoven and Peter Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.