UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert meeting on Statistical Data Editing
7-9 October 2024, Vienna

Full conditional distributions for handling restrictions in the context of automated statistical data editing

Christian Aßmann[1,2], Ariane Würbach[1], Younes Saidani[3], Florian Dumpert[3]

[1] *Leibniz Institute for Educational Trajectories, Bamberg, Germany;*
[2] *Chair of Survey Statistics and Data Analysis, Otto-Friedrich-Universität, Bamberg, Germany;*
[3] *Federal Statistical Office, Wiesbaden, Germany*

christian.assmann@lifbi.de; ariane.wuerbach@lifbi.de; younes.saidani@destatis.de; florian.dumpert@destatis.de

I. INTRODUCTION

1. Reported survey data typically contain ambiguities and inaccuracies due to respondent errors as reported values do not comply with logical restrictions or are missing. Statistical offices have often established edit-imputation routines following the Felligi-Holt paradigm to correct data and ensure data coherence, thereby employing an easily computable heuristic that does not necessarily use all information available in the observed data. In contrast, Bayesian methods for edit-imputation incorporate all available information in the full conditional distributions of implausible values and correctly reflect the uncertainty arising from the process of replacing erroneous values. While for categorical and continuous data, Bayesian approaches based on parametric models are available in the literature, this article lays out a method for specifying full conditional distributions using classification and regression trees (CART) while taking into account nested balance restrictions, i.e. nested restrictions involving several variables. Using the CART algorithm provides flexible univariate approximations to the full conditional distributions of the variables yet reduces the computational intensity of the overall Bayesian approach. The feasibility of the suggested approach is documented in terms of a simulation study and an empirical application based on insights into the data editing of a specific survey of the Federal Statistical Office of Germany. Simulation results suggest that compared to complete case analysis average mean square error of moment estimates can typically be reduced by 20 to 30 percent when using the non-parametric Bayesian approach and the corresponding specification of full conditional distributions using the CART algorithm.

2. Despite often legally bounding participation in official statistics' surveys, collected data may contain errors in relation with logical restrictions among variables. These errors render reported values as implausible. Therefore, most of the datasets collected for the purposes of official statistics undergo editing in the course of their production process. This occurs in particular in Sub-Processes 5.3 and 5.4 of the GSBPM [United Nations Economic Commission for Europe, 2019]. Editing is often semi-automated. In many cases, it is carried out using (i) defined editing rules to detect anomalies and (ii) largely manual correction by clerks of data records found to be possibly or definitely incorrect. The main weaknesses of this procedure are the extremely high manual effort and – where this effort cannot (or can no longer) be fully realised – possible quality losses. Even if financial or personnel-related reasons

would make comprehensive editing possible, the question of the time required arises. Official statistics are required to provide their results quickly. Editing that takes a long time, both in absolute terms and relative to the periodicity of the statistics, contradicts this. In view of the conflict of objectives between accuracy and timeliness – i.e. Principles 12 and 13 of the Quality Assurance Framework of the European Statistical System [European Statistical System, 2019] – a (partial) automation of editing appears to be imperative.

3. Statistical data editing is not a new field in official statistics. The UNECE workshops and expert meetings on this topic prove this with their numerous papers as well as the relevant standard works like De Waal et al. [2011] and Van der Loo and De Jonge [2018].

4. This paper adds to the literature on edit-imputation of microdata a Bayesian approach for handling of nested equality and inequality restrictions involving a diversity of different variable scaling including censoring and truncation, as well as categorical data. Key element is the specification of full conditional distributions for the implausible values taking the structure of the restrictions and the characteristics of the variables into account. We suggest to use classification and regression trees (CART) to account for the dependencies among the variables. The benefits are illustrated via a simulation study and an empirical application involving typical complexities.

5. The paper is structured as follows. Section II describes the methodological approach suggested within this paper for handling of implausible values occurring in official statistics' surveys. A simulation study is provided within Section III, where as an extension towards the complexities of an empirical illustration is provided in Section IV. Section V concludes.

II. HANDLING OF IMPLAUSIBLE VALUES

1. Methods to address the issue of statistical data editing aim at effectively assessing knowable quantities in the sense of Lewbel [2019], with distribution functions and functions thereof like expectations and quantiles typically summarize the information of interest. In a cross section data context with the number of observations N being larger than the number of variables P, the operationalisation of a multivariate density

$$f(X|\theta)$$

for a data matrix X of size $N \times P$ critically hinges on the characteristics of the considered random variables. These characteristics include scaling, e.g. categorical or numerical, as well as range restrictions in terms of truncations and censoring. With increasing number of variables involved, the specification of multivariate distributions becomes challenging per se.

2. In addition, the data distribution is also required to describe the nature or occurrence of the values labelled as implausible, i.e. among other things, the consideration of frequency distributions of the occurrence of implausible values. In the context of semi-automated machine learning methods, the way in which a joint density function can be constructed without full supervision is of particular interest. The discussion about the fundamental feasibility is closely linked to the possibilities of decomposing or factorising a joint distribution. There is the fundamental possibility to decompose a joint density function of $X = (X_1, \ldots, X_P)$ sequentially in any order, i.e.

$$f(X_1, \ldots, X_P|\theta) = f(X_1|\theta) f(X_2|X_1, \theta) \ldots f(X_P|X_1, \ldots, X_{P-1}, \theta).$$

Except for the context of time series data, any sequential ordering of the conditional distributions remains arbitrary to some extent. Of more general interest is to consider the decomposition of the joint distribution for a set of variables against the background of the Clifford-Hammersley theorem.

The basic statement of the Clifford-Hammersley theorem, see Robert and Casella [2004] for a basic presentation and further details, aims at showing that a joint distribution is

$$f(X^{\text{ipl}}, \theta | X^{\text{obs}}) \propto f(X^{\text{ipl}} | X^{\text{obs}}, \theta)\pi(\theta), \tag{1}$$

where θ denotes all relevant parameters describing the density of interest also including e.g. smoothing, tuning, or hyperparameters.

3. For assessing or setting up a multivariate density reflecting the dependencies and nested restrictions on the variables, the joint distribution according the Clifford-Hammersley theorem, see Robert and Casella [2004], is decomposed into a set of full conditional distributions. Hence,

$$f(\theta, X^{\text{ipl}} | X_{\text{obs}}) \propto \frac{f(\theta | \tilde{X}^{\text{ipl}}, X^{\text{obs}})}{f(\tilde{\theta} | \tilde{X}_{\text{ipl}}, X^{\text{obs}})} \frac{f(X^{\text{ipl}} | \theta, X^{\text{obs}})}{f(\tilde{X}^{\text{ipl}} | \theta, X^{\text{obs}})}, \tag{2}$$

where $\tilde{\cdot}$ denotes an arbitrarily chosen point of the indicated variables. Further, the full conditional distributions are subject to further decomposition based on an arbitrary ordering of the variables denoted as $[1], [2], \ldots, [P]$ yielding

$$\begin{aligned} f(X^{\text{ipl}} | \theta, X^{\text{obs}}) \quad \propto \quad & \frac{f(X^{\text{ipl}}_{[1]} | \theta, \tilde{X}^{\text{ipl}}_{[1]}, \ldots, \tilde{X}^{\text{ipl}}_{[P]}, X^{\text{obs}})}{f(\tilde{X}^{\text{ipl}}_{[1]} | \theta, \tilde{X}^{\text{ipl}}_{[1]}, \ldots, \tilde{X}^{\text{ipl}}_{[P]}, X^{\text{obs}})} \frac{f(X^{\text{ipl}}_{[2]} | \theta, X^{\text{ipl}}_{[1]}, \tilde{X}^{\text{ipl}}_{[2]}, \ldots, \tilde{X}^{\text{ipl}}_{[P]}, X^{\text{obs}})}{f(\tilde{X}^{\text{ipl}}_{[2]} | \theta, X^{\text{ipl}}_{[1]}, \tilde{X}^{\text{ipl}}_{[2]}, \ldots, \tilde{X}^{\text{ipl}}_{[P]}, X^{\text{obs}})} \frac{\ldots}{\cdots} \\ & \frac{\ldots}{\cdots} \frac{f(X^{\text{ipl}}_{[P]} | \theta, X^{\text{ipl}}_{[1]}, \ldots, X^{\text{ipl}}_{[P]}, X^{\text{obs}})}{f(\tilde{X}^{\text{ipl}}_{[P]} | \theta, X^{\text{ipl}}_{[1]}, \ldots, X^{\text{ipl}}_{[P]}, X^{\text{obs}})}. \end{aligned} \tag{3}$$

The advantage of this decomposition relates to the possibility to characterise the joint distribution of X^{ipl} in terms of univariate distributions. A flexible way to specify univariate full conditional distributions is to use classification and regression trees (CART) to characterise the dependencies, see Breiman et al. [1984], Burgette and Reiter [2010], and Doove et al. [2014], where as Aßmann et al. [2023] use CART in combination with a Bayesian estimation approach for handling missing values in a hierarchical regression model for binary and ordinal dependent variables.[1] The inherent flexibility of characterising the full conditional distributions via the CART algorithm also relates to consider arbitrary transformations of variables within the conditioning set. These transformation may include fractional or higher order moments of observed variables, as well as cross products to account for nonlinear relationships and stationary dependencies.[2]

4. Furthermore, the existence of the joint distribution implied by the decomposition is directly ensured in case the missing values relate to variables with finite sample space. If the sample spaces of the considered variables are not finite, existence may be indirectly granted when the full conditional distributions are characterised via the CART algorithm as the characterisation relies on measures of homogeneity and incorporates the restriction on the range implied by observed values. This is typically in line with the positivity constraint ensuring the existence of the joint distribution when for each point the positivity of the marginal distribution implies the positivity of the joint distribution. The central role of a joint distribution results from the possibilities to define quality parameters that characterise the quality of the automated procedure used and allow comparisons of different approaches to generate a

[1] Note that in order to map the structural nature of the set of variables under consideration adequately within the framework of a joint and thus necessarily multivariate density function, either a direct characterisation of the joint distribution can be carried out or an indirect characterisation can be used. In the case of an indirect characterisation, this can be complete or incomplete, i.e. based on conditional moments, for example. A complete indirect characterisation of the joint density function is usually performed using conditional density functions. Several approaches exist in the literature to analyse the properties of the different characterisations.

[2] Note that as a precondition, stationarity of variables is presumed, or variables are transformed suitably to exhibit stationary behavior.

joint density function. In addition, operationalisation options for error identification, error localisation and error correction are also directly based on a joint distribution.

5. The formulation of the problem by means of a density function opens up a variety of possibilities to check the modelling quality, to present it comprehensibly and to replace implausible ones statistically. An implausible value must be present if the observed value does not fulfil the specified restriction. This includes in particular the situation in which the restriction covers not only individual variables, but a subset of X. As an example, a restriction of the form $\sum_{p=1}^{P} X_{ip} = 0$ can be considered. All variable values that are covered by the restriction can be considered implausible in their entirety if the restriction is violated. Hence, all the variables related to the restrictions need to be sampled from the corresponding full conditional distribution.

6. Nevertheless, the violation of a restriction can have different impact on the full conditional distribution depending on the kind of restriction and the nesting structure. Consider the case of two nested equality restrictions as one of the involved variables is subject to both restrictions. In case both restrictions are violated all involved variables require replacing. However, in case only one of the two restrictions is violated, the variable occurring in both restrictions can be considered as validated and hence be used as conditional variable for the remaining variables in the violated restriction.

7. Next, consider the case of one equality and one inequality restriction. Again, if both restrictions are violated all corresponding variables require replacement. Further, if the equality restriction holds, the involved variables can be considered as validated and can serve as conditional variables for handling the remaining variables in the inequality restriction. However, in case the inequality restriction is not violated, this does not validate the involved variables in the violated equality restriction causing in turn that all variables within the restrictions require replacement. This illustrates the possibilities how restrictions inform and shape the set of conditional variables available for modelling.

8. The category of implausible values may further contain all values that do not violate any restrictions with regard to their permissible value range, but are classified as implausible per se. Although this kind of implausible values may be subject to a demarcation problem, they are subsumed as missing values as handling is similar to implausible values occurring from violated restrictions. In this sense, the suggested approach assumes that the causes of implausible values are not due to gross negligence or carelessness, but rather to oversights or carelessness. This includes typing errors, transposed figures, inadvertent use of incorrect reference periods or cut-off dates, but also the rounding of data to reduce the material and cognitive costs of data entry.[3]

III. SIMULATION STUDY AND SIMULATION RESULTS

1. In the simulation, we consider the following basic setup. Each data set contain $N = 10{,}000$ observations, where each observational unit refers to $J = 10$ variables. The variables for each unit $i = 1, \ldots, N$ are subject to three restrictions (A, B, and C), with two restrictions (A and B) being nested. The restrictions are described as $A : V_{i1} + V_{i2} + V_{i3} = V_{i4}$, $B : V_{i4} + V_{i5} = V_{i6}$, and $C : V_{i7} + V_{i8} + V_{i9} = V_{i10}$. The simulation refers to $S = 1{,}000$ repeatedly generated data sets based on the identical data generating process, i.e. each simulated data set is based on the same parameters as described below. With regard to violations of the restrictions, we consider the following combinations. For some units, all three restriction are violated causing a complete loss of information for this unit.

[3] It is beyond the scope of this paper to analyze the possible benefits for data quality using prompts during the data input phase informing data providers that the provided information is highly unlikely and possibly due to a typing error.

Next, for some units only restriction C is violated, while other units violate only A or B but not both, nor C. Finally, some units show violation of restrictions A and B but have intact restriction C. This simulation design results in five violation types, where each type corresponds to a specific situation to be modelled via a full conditional distributions, or, in other words required consideration within the estimation. Missing values implied by violations are generated completely at random, where each type shows approximately 5% of missing values with the related variables, see also Table 1.

2. The full conditional distributions handling the different types of restriction violation take then the following form,

(1) $f(V_{i7}, \ldots, V_{i9}(, V_{i10})|V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i5}(, V_{i6}))$,
(2) $f(V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i5}(, V_{i6})|V_{i7}, \ldots, V_{i9}(, V_{i10}))$,
(3) $f(V_{i1}, V_{i2}(, V_{i3})|V_{i4}, V_{i5}(, V_{i6}), V_{i7}, V_{i8}, V_{i9}(, V_{i10}))$,
(4) $f(V_{i5}(, V_{i6})|V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i7}, \ldots, V_{i9}(, V_{i10}))$,
(5) $f(V_{i1}, V_{i2}, V_{i3}, (, V_{i4}), V_{i5}(, V_{i6}), V_{i7}, V_{i8}, V_{i9}(, V_{i10}))$.

The specifications of the full conditional distribution thereby consider the identification issues inherent to the equality restrictions via dropping one variable (in parentheses) per equality restriction from the set of conditioning variables. The simulations setup based on the multivariate normal distribution offers several possibilities to specify the functional form of the full conditional distributions. First, we consider conditional multivariate normal distribution as implied by multivariate normal theory, see Mittelhammer [2013]. The corresponding distributions can be described as

$$f(x_{i,\text{mis}}|x_{i,\text{obs}}) \propto |\Sigma_{\text{mis}|\text{obs}}|^{-.5} \exp\left\{-\frac{1}{2}(x_{i,\text{mis}} - \mu_{\text{mis}|\text{obs}})'\Sigma^{-1}_{\text{mis}|\text{obs}}(x_{i,\text{mis}} - \mu_{\text{mis}|\text{obs}})\right\},$$

where

$$\mu_{\text{mis}|\text{obs}} = \mu_{\text{mis}} + \Sigma_{p,\text{mis},\text{obs}}\Sigma^{-1}_{\text{obs}}(x_{i,\text{obs}} - \mu_{\text{obs}}) \quad \text{und} \quad \Sigma_{\text{mis}|\text{obs}} = \Sigma_{\text{mis}} - \Sigma_{p,\text{mis},\text{obs}}\Sigma^{-1}_{\text{obs}}\Sigma'_{p,\text{mis},\text{obs}}$$

with

$$\mu = (\mu_{\text{mis}}, \mu_{\text{obs}})' \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{\text{mis}} & \Sigma_{p,\text{mis},\text{obs}} \\ \Sigma'_{p,\text{mis},\text{obs}} & \Sigma_{\text{obs}} \end{pmatrix}.$$

3. Alternatively, we use classification and regression trees to capture the dependencies among the variables. Classification and regression trees partition the data set and provide a set of similar values given the conditioning variables. This set can then be used to approximate the full conditional distribution, either by estimating parameters of suitable parametric density (CART-M) or in a non-parametric fashion (CART-B). With regard to the hyperparameters of the CART modelling approach, we specify a minimum of 5 observations in the final sets, where a minimum of 10 observations is required for splits to be eligible, while all other hyperparameters in terms of optimising fit are set to default values.

4. The data generating mechanism resembled within the simulation study refers to the multivariate normal distribution. It is designed to provide insights into the functioning of the suggested approach via comparing different approaches towards handling the implausible values due to violated restrictions. The multivariate normal distribution allows for setting up benchmark estimators that are available given the considered linear restrictions and allow for gauging the relative benefits of the suggested approach. Hence, data are generated via a 7-dimensional multivariate normal distribution, i.e.

$$V_r = (V_{i1}, V_{i2}, V_{i3}, V_{i5}, V_{i7}, V_{i8}, V_{i9})' \sim \mathcal{MVN}(\mu_r, \Sigma_r),$$

where $\mu_r = (100, 100, 100, 200, 100, 100, 100)'$ and $\Sigma_r = 20(.5\text{diag}(\iota_7) + .5\iota_7\iota'_7)$ with (ι_7 denoting a 7×1 vector of ones.

The considered structure of the restrictions allows to derive the variables V_4, V_6 and V_{10} as linear combinations (L) of the seven other variables, i.e.

$$V = LV_r \sim \mathcal{MVN}(\mu = L\mu_r, \Sigma = L\Sigma_r L').$$

Accordingly, the corresponding moments and quantities of interest like expected values μ, covariance Σ, and quantiles $q_x = \mu + \text{diag}(\Sigma)^{0.5}\Phi^{-1}(x)$ with $\Phi^{-1}(\cdot)$ denoting the inverse cdf of the standard normal distribution can be used for benchmarking the precision of estimates for these quantities. However, not for all quantities of interest analytical expressions are available. With regard to minimum and maximum, i.e. first and last order statistic, no closed form expression exist with regard to expected value and higher order moments thereof. However, expected values can be readily assessed via the simulation setup, where the expected value can be approximated as[4]

$$\text{E}[\min\{V_i\}] \approx \frac{1}{S}\sum_{s=1}^{S} \min_i \left(V_i^{(s)}\right) \quad \text{and} \quad \text{E}[\max\{V_i\}] \approx \frac{1}{S}\sum_{s=1}^{S} \max_i \left(V_i^{(s)}\right).$$

5. These quantities then can be used to assess mean square errors and relative mean square errors of alternative estimators, i.e.

$$\text{rel. MSE} = (1-)\frac{\frac{1}{S}\sum_{s=1}^{S}(\theta_s^* - \theta_{(s)})^2}{\frac{1}{S}\sum_{s=1}^{S}(\tilde{\theta}_s - \theta_{(s)})^2},$$

where $\theta_s^* = \frac{1}{G}\sum_{g=1}^{G}\theta_{s,g}^*$ (G = 1,000) and $\theta_{(s)}$ refers to the parameter value of the data generating process (not varying with s), or, the estimator resulting from complete data (before deletion). For benchmarking, we consider before deletion, complete case, and the suggested approach using full conditional distributions for estimating the structural parameters of the involved data generating process. Each of these approaches is applied to each of the simulated data sets resulting in S estimators denoted as $\{\tilde{\theta}_s\}_{s=1}^{S}$. Denoting the corresponding value of the data generating process as θ, (relative) mean square errors can be calculated. This is done for each quantity of interest, thus θ involves μ, vec(Σ), $q_{10\%}$, $q_{25\%}$, $q_{75\%}$, $q_{90\%}$, all 10 minima, and all 10 maxima and hence in total 125 parameters.

6. Results of the corresponding simulation study are provided in Table 2 for mean square errors. The results show that mean square errors relative to the estimator before deletion are smallest for the approach handling implausible values via full conditional distributions based on the multivariate normal distribution. This results from the instantaneous relation between the data generating mechanism and the approach for handling implausible values. The relative mean square error of the CART based approaches are slightly larger when coupled with a parametric normal distribution and again increased when referring to the non-parametric approximation of the full conditional distribution. The reason for these differences between the approaches can be best illustrated with the results for the minima and maxima. The results show negligible relative mean square errors for the CART based estimators. In connection with the relatively low missing rates induced by the simulation design, this implies that on average the true extreme value is within the completed distribution of each variable. The relative advantage of the semi-parametric approach is relative to the possibility of smoothing the available information towards the observed data distribution. However, this advantage pays off only in this specific context and comes at the cost of reduced flexibility that will typically prevail in empirical application context.

[4]The alternative calculation via numerical integration would involve extra steps, whereas the simulation based approximation is directly available.

IV. EMPIRICAL ILLUSTRATION

1. For the empirical illustration, we consider a data set that is larger, shows more complexity with regard to restrictions, and with regard to modelling of variables thereby illustrating the potential of the suggested approach.
The implementation is based on the following information. First, a $N \times P$ matrix indicating which values of variables $p = 1, \ldots, P$ are missing per observation $i = 1, \ldots, N$, i.e. implausible values are already removed. This matrix contains information about the available complete cases and the distribution of missing values per observation $i = 1, \ldots, N$. This information is relevant for initialising the missing values via initial conditional distributions based on the initially available information and then proceeding to conditional distributions based on augmented information. Some descriptives of the variables are given in Table 3: first, the support of the variables is listed, followed by some summary statistics (mean, sd, minima, maxima, curtosis). It should be noted that for the majority of variables the support is restricted to positive continuous values ranging between zero and infinity. A few variables are counting variables, and only one variable is categorical. A possible zero-inflation for the continuous variables is shown in column "Zeros", and the last two columns inform about the shares of ones albeit one is not in the support of a continuous variable.
Next, information about the restrictions is required. Figures 1 and 3 provide the observed restrictions, how they are related to each other by common variables, and restrictions that are commonly violated. For Figures 2 and 4 the perspective is on the variables, i.e. variables related to each other by restrictions, and variables that are jointly subject to violation. Hence, Table 4 and Figure 2 provide the involvement of the variables within the restrictions.

2. The initial sequence for filling in missing values is based only on the completely observed cases (N_{obs}) to characterise the full conditional distributions. Thereby, the missing values are handled per observation unit i, with the units ordered according to increasing number of missing values values per unit including a check, whether a missing pattern prevails for several units. Then, for each missing pattern, with maximum number of missing patterns limited from above by the number of units with missing values (N_{mis}), with $N = N_{\text{obs}} + N_{\text{mis}}$. Given a specific missing pattern, the features of the involved restrictions need to be considered in setting up the full conditional distributions.[5] The sequence of full conditional distribution can be set up based on the information provided in Tables 5 and 6. The consecutive sequence of variables is decided upon ordering the variables according to increasing numbers of missing values implying a decomposition of the multivariate distribution as described in Equation 3.

3. The Tables 5 and 6 inform about the dependencies occurring in terms of the available set of conditioning variables and the sequence that should be followed in order to ensure sampling of missing values in line with the restrictions. The intended joint distribution of missing values $f(X_{i,\text{mis}}|X_{\text{obs}}, \psi)$ is then decomposed into univariate full conditional distributions while taking the prevalent restrictions into account. With regard to taking the restrictions on the variables into account, we apply the following reasoning: The prevailing restrictions can be classified as (i) equality restrictions, or (ii) inequality restrictions. With regard to nested restrictions, we follow Kim et al. [2015] and adapt a bottom-up strategy applied where applicable. However, nested restrictions given the involved identification problem also influence the set of conditioning variables available for modelling the full conditional distributions. A typical decomposition takes the form

$$f(x_{i,\text{mis}}|X_{\text{obs}}) = f(x_{i,\text{mis}}|X_{\text{obs}} \setminus \{x^{(1)}, \ldots, x^{(P)}) f(x_{i,\text{mis}}|X_{\text{obs}} \setminus \{x^{(2)}, \ldots, x^{(P)}) \cdots f(x_{i,\text{mis}}|X_{\text{obs}} \setminus \{x^{(P)}\}).$$

[5] In the absence of restrictions, the conditional joint distribution of missing values $f(X_{i,\text{mis}}|X_{\text{obs}})$ could be decomposed into a sequence of univariate conditional distributions.

4. With regard to sampling, the set of values revealed by the CART algorithm provides an empirical distribution approximating the full conditional distribution of interest. Sampling can hence be performed using a Bayesian bootstrap, or, the set of values can be used to characterise parameters of a suited parametric distribution, see Tables 8 for details on estimation and parametric density applying to the variables in the empirical illustration. It will be illustrated that combining the CART characterisation of dependencies with parametric approximations may benefit statistical efficiency for some estimators depending on the overall amount of missing data and total information available. In detail, several variables can be characterised in form of a censored truncated[6] normal distribution given as

$$f_{\mathcal{CTN}}(x|D) = p(D)\mathcal{I}(x=0) + (1-p(D))\frac{\phi\left(\frac{x-\mu(D)}{\sigma(D)}\right)}{1-\Phi\left(\frac{-\mu(D)}{\sigma(D)}\right)}\mathcal{I}(0<x<\infty),$$

where $\phi(\cdot)$ and $\mathcal{I}(\cdot)$ denote the standard normal density function and indicator function respectively, whereas $p(D)$ denotes the probability of the variable to equal zero and $\mu(D)$ and $\sigma(D)$ denote the location and scale parameters estimated from the conditioning data D. The parameters can be derived from the set of donor elements characterised via the CART algorithm with

$$p(D) = \frac{1}{n_{\text{donor}}}\sum_{j=1}^{n_{\text{donor}}}\mathcal{I}(x_j^{\text{donor}}=0),$$

whereas $\mu(D)$ and $\sigma(D)$ can be established via the moment conditions

$$\begin{pmatrix} f_\mu(\mu,\sigma) = m(x|D) \\ f_\sigma(\mu,\sigma) = s^2(x|D) \end{pmatrix},$$

where

$$f_\mu(\mu,\sigma) = \mu(D) + \sigma(D)\frac{\phi(-\frac{\mu(D)}{\sigma(D)})}{1-\Phi(-\frac{\mu(D)}{\sigma(D)})} \quad , \quad f_{\sigma^2}(\mu,\sigma) = \sigma(D)^2\left(1-\left(\frac{\phi(-\frac{\mu(D)}{\sigma(D)})}{1-\Phi(-\frac{\mu(D)}{\sigma(D)})}\right)^2\right)$$

and

$$s^2(x|D) = \frac{1}{n_{\text{donor}} - \sum_{j=1}^{n_{\text{donor}}}\mathcal{I}(x_j=0)}\sum_{j=1}^{n_{\text{donor}}}(x_j - m(x|D))^2\mathcal{I}(0<x_j<\infty),$$

with

$$m(x|D) = \frac{1}{n_{\text{donor}} - \sum_{j=1}^{n_{\text{donor}}}\mathcal{I}(x_j=0)}\sum_{j=1}^{n_{\text{donor}}}x_j\mathcal{I}(0<x_j<\infty).$$

The system of equations can be solved iteratively using a Taylor approximation of first order.[7] While the Bayesian bootstrap directly accounts for the parameter uncertainty, the semi-parametric approach can account for parameter uncertainty via setting up suitable priors. However, if the full conditional

[6] We refer to censoring and truncation in the following way. Censoring occurs when a perceived continuous random variables has probability mass at one specific point that routinely would have a probability mass of zero. Truncation occurs when the range of the random variable is restricted to a range of values being element of an open interval. In the empirical setting considered here, the censoring can be conceptualized in form of a mixture distribution.

[7] The system of equations can be described as

$$\begin{pmatrix} f_\mu(\mu,\sigma) \\ f_{\sigma^2}(\mu,\sigma) \end{pmatrix} - \begin{pmatrix} m(x|D) \\ s^2(x|D) \end{pmatrix} = 0.$$

Using a first order Taylor series approximation

$$\begin{pmatrix} f_\mu(\mu,\sigma) \\ f_{\sigma^2}(\mu,\sigma) \end{pmatrix} \approx \begin{pmatrix} f_\mu(\mu_0,\sigma_0) \\ f_{\sigma^2}(\mu_0,\sigma_0) \end{pmatrix} + H(\mu_0,\sigma_0)\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

this results in

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = H^{-1}(\mu_0,\sigma_0)\left[\begin{pmatrix} m(x|D) \\ s^2(x|D) \end{pmatrix} - \begin{pmatrix} f_\mu(\mu_0,\sigma_0) \\ f_{\sigma^2}(\mu_0,\sigma_0) \end{pmatrix}\right]$$

distributions are characterised in terms of a discrete density function, sampling can be conducted via inversion of the empirical distribution function.

5. The data set considered within the empirical contains $N = 17{,}286$ observations on $P = 48$ variables, where a total of 61 restrictions does apply. The most common restriction prevailing excludes a variable to take specific values.

V. CONCLUSION

1. The paper illustrates handling of missing values due to violations of restrictions placed on the involved variables. The proposed methods resembles a Bayesian approach via iterative sampling from the set of full conditional distributions providing an approximation towards the joint distribution of the observed sample data. To cope with the different dependencies among the variables and the different scaling types, classification and regression trees are used to characterise the full conditional distributions. Sampling from the full conditional distributions is either performed based on the empirical distributions using a Bayesian bootstrap or via sampling from an adapted parametric distribution. While the simulation results showing the possibilities to gains statistical efficiency in terms of reduced bias and mean square error, the empirical illustration focuses on implementation issues and adaption towards the complexities of empirical application.

Acknowledgements

The authors thank the participants of the UNECE Expert Meeting on Statistical Data Editing 2024 for helpful comments and suggestions and thank Katja-Verena Bürk (Federal Statistical Office, Germany) for her support.

References

C. Aßmann, C. Gaasch, and D. Stingl. A Bayesian Approach towards Missing Covariate Data in Multilevel Latent Regession Models. *Psychometrika*, 88(4):1495–1528, 2023. doi: https://doi.org/10.1007/s11336-022-09888-0.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

L. F. Burgette and J. P. Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172:1070–1076, 2010. doi: https://doi.org/10.1093/aje/kwq260.

T. De Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation*. John Wiley & Sons, 2011.

---

with

$$
\begin{aligned}
H(\mu_0,\sigma_0) &= \begin{pmatrix} \frac{\partial}{\partial\mu} f_\mu(\mu,\sigma) & \frac{\partial}{\partial\mu} f_\mu(\mu,\sigma) \\ \frac{\partial}{\partial\sigma} f_\sigma(\mu,\sigma) & \frac{\partial}{\partial\sigma} f_\sigma(\mu,\sigma) \end{pmatrix} \\
&= \begin{pmatrix} 1+\sigma\frac{\phi'(-\frac{\mu}{\sigma})(-\frac{1}{\sigma}(1-\Phi(-\frac{\mu}{\sigma})-\phi^2(-\frac{\mu}{\sigma})\frac{1}{\sigma}}{\left(1-\Phi(-\frac{\mu}{\sigma})\right)^2} & -2\frac{\phi(-\frac{\mu}{\sigma})}{1-\Phi(-\frac{\mu}{\sigma})}\frac{\phi'(-\frac{\mu}{\sigma})(-\frac{1}{\sigma}(1-\Phi(-\frac{\mu}{\sigma})-\phi^2(-\frac{\mu}{\sigma})\frac{1}{\sigma}}{\left(1-\Phi(-\frac{\mu}{\sigma})\right)^2} \\ \left(\frac{\phi(-\frac{\mu}{\sigma})}{1-\Phi(-\frac{\mu}{\sigma})}\right)+\sigma\frac{\phi'(-\frac{\mu}{\sigma})(-\frac{\mu}{\sigma^2}(1-\Phi(-\frac{\mu}{\sigma})+\phi^2(-\frac{\mu}{\sigma})\frac{\mu}{\sigma^2}}{\left(1-\Phi(-\frac{\mu}{\sigma})\right)^2} & -2\frac{\phi'(-\frac{\mu}{\sigma})(-\frac{1}{\sigma}(1-\Phi(-\frac{\mu}{\sigma})-\phi^2(-\frac{\mu}{\sigma})\frac{1}{\sigma}}{\left(1-\Phi(-\frac{\mu}{\sigma})\right)^2} \end{pmatrix}.
\end{aligned}
$$

L.L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014. doi: https://doi.org/10.1016/j.csda.2013.10.025.

European Statistical System. Quality Assurance Framework of the European Statistical System (Version 2.0), 2019. URL https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf.

H.J. Kim, L.H. Cox, A.F. Karr, J.P. Reiter, and Q. Wang. Simultaneous edit-imputation for continuous microdata. *Journal of the American Statistical Association*, 110(511):987–999, 2015. URL https://www.jstor.org/stable/24739700.

A. Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, December 2019. doi: 10.1257/jel.20181361. URL https://www.aeaweb.org/articles?id=10.1257/jel.20181361.

R.C. Mittelhammer. *Mathematical Statistics for Economics and Business*. Springer New York, 2013. ISBN 978-1-4614-5021-4 978-1-4614-5022-1. doi: 10.1007/978-1-4614-5022-1. URL https://link.springer.com/10.1007/978-1-4614-5022-1.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer New York, New York, NY, 2004. ISBN 978-1-4419-1939-7. doi: https://doi.org/10.1007/978-1-4757-4145-2.

United Nations Economic Commission for Europe. Generic Statistical Business Process Model (GSBPM), 2019. URL https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

M. Van der Loo and E. De Jonge. *Statistical data cleaning with applications in R*. John Wiley & Sons, 2018.

Tables

TABLE 1. Missing Data Structure within Simulation Study

lfd. Nr.	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
1	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	m	m	m	m
3	x	x	x	x	x	x	x	x	x	x
4	m	m	m	m	m	m	x	x	x	x
5	x	x	x	x	x	x	x	x	x	x
6	m	m	m	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x
8	x	x	x	x	m	m	x	x	x	x
9	x	x	x	x	x	x	x	x	x	x
10	m	m	m	m	m	m	m	m	m	m
11	x	x	x	x	x	x	x	x	x	x
⋮										
N	x	x	x	x	x	x	x	x	x	x
# ≈ complete cases	85%	85%	85%	90%	85%	85%	90%	90%	90%	90%
					75%					

TABLE 2. Results of Simulation Study – RMSE

	DGP			DGPt			data		
expectations	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**
Min.	0,0078	-0,0008	-0,0398	0,0077	-0,0001	-0,0398	0,1797	-0,0103	-0,1268
1st Qu.	0,0246	0,0144	-0,0200	0,0246	0,0142	-0,0201	0,2322	0,0623	-0,0500
Median	0,0367	0,0350	0,0001	0,0367	0,0350	-0,0003	0,2910	0,1791	0,0143
Mean	0,0391	0,0342	0,0023	0,0390	0,0341	0,0022	0,3089	0,1672	0,0527
3rd Qu.	0,0508	0,0391	0,0089	0,0508	0,0389	0,0088	0,3586	0,2214	0,0926
Max.	0,0786	0,0890	0,0840	0,0786	0,0888	0,0835	0,5604	0,4461	0,4330
	DGP			**DGPt**			**data**		
minima	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**
Min.	0,1256	-0,4465	-0,4529	0,1256	-0,4465	-0,4529	0,0887	-4,8852	-4,3840
1st Qu.	0,1349	0,1000	0,0000	0,1349	0,1000	0,0000	0,1039	-0,2001	-0,0026
Median	0,1423	0,1735	0,0000	0,1423	0,1735	0,0000	0,1307	-0,0740	0,0000
Mean	0,1637	0,1063	-0,0406	0,1637	0,1063	-0,0406	0,1603	-0,5681	-0,4182
3rd Qu.	0,1986	0,2071	0,0064	0,1986	0,2071	0,0064	0,1721	0,0403	0,0000
Max.	0,2235	0,2324	0,0261	0,2235	0,2324	0,0261	0,3762	0,2243	0,2221
	DGP			**DGPt**			**data**		
maxima	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**
Min.	0,1337	-0,4458	-0,8819	0,1337	-0,4458	-0,8819	0,0784	-5,9917	-8,0131
1st Qu.	0,1354	0,0961	0,0000	0,1354	0,0961	0,0000	0,1243	-0,2220	-0,0042
Median	0,1592	0,1692	0,0000	0,1592	0,1692	0,0000	0,1766	-0,1807	0,0000
Mean	0,1671	0,1032	-0,0843	0,1671	0,1032	-0,0843	0,1766	-0,6827	-0,7872
3rd Qu.	0,1961	0,2094	0,0043	0,1961	0,2094	0,0043	0,1959	0,0317	0,0000
Max.	0,2190	0,2311	0,0216	0,2190	0,2311	0,0216	0,3678	0,3243	0,1574
	DGP			**DGPt**			**data**		
covariances	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**
Min.	0,1042	-1,6477	-0,9647	0,1044	-1,6323	-1,0887	0,5311	-6,6447	-3,9709
1st Qu.	0,1500	0,0457	0,0109	0,1505	0,0617	-0,0295	0,5964	0,1749	-0,0023
Median	0,1625	0,1036	0,0924	0,1625	0,1094	0,0867	0,6520	0,4454	0,3355
Mean	0,1608	0,0146	0,0282	0,1608	0,0242	0,0058	0,6395	0,0640	0,0364
3rd Qu.	0,1756	0,1342	0,1404	0,1746	0,1369	0,1382	0,6892	0,5378	0,4553
Max.	0,1987	0,1762	0,1877	0,1987	0,1783	0,1816	0,7278	0,7104	0,6792
	DGP			**DGPt**			**data**		
quantiles	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**	**MVN**	**CART(N)**	**CART(B)**
Min.	0,0681	-0,5086	-0,3859	0,0679	-0,4835	-0,3979	0,1157	-3,6348	-2,5414
1st Qu.	0,1011	0,0572	-0,0423	0,1011	0,0548	-0,0429	0,1716	-0,1085	-0,3704
Median	0,1329	0,0721	-0,0258	0,1328	0,0745	-0,0282	0,2053	-0,0212	-0,2885
Mean	0,1367	0,0397	-0,0452	0,1368	0,0405	-0,0478	0,2278	-0,2448	-0,4246
3rd Qu.	0,1583	0,0912	-0,0092	0,1581	0,0869	-0,0167	0,2497	0,0485	-0,1963
Max.	0,2492	0,1421	0,0521	0,2489	0,1306	0,0521	0,4849	0,3347	0,2039

Notes: relative MSE (1-rel. MSE) in comparison to benchmark estimator, aggregation of parameters for each group of parameters, i.e. means, covariance, quantiles, minima, and maxima.

TABLE 3. Variable Descriptives

No.	Variable	Scaling/ Unconstrained Support	Levels	Mean	SD	Minimum	Median	Maximum	Curtosis	Zeros	Ones	One allowed
1	$v^{(4)}$	nominal/{4,7,8}	3							0.000	0.000	TRUE
2	$v^{(21)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		4.22	37.76	0	1.0	2,177	0.085	0.455	0.261	TRUE
3	$v^{(22)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		389.07	12,629.52	0	103.0	1,643,477	0.023	0.001	0.003	TRUE
4	$v^{(24)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		32.89	174.84	0	10.0	9,784	0.131	0.057	0.041	TRUE
5	$v^{(25)}$	transformation		19.96	114.23	0	5.0	6,283	0.131	0.057	0.058	TRUE
6	$v^{(27)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		393.24	12,629.89	0	107.0	1,643,528	0.023	0.001	0.000	TRUE
7	$v^{(28)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		1.01	9.40	0	0.0	540	0.107	0.820	0.124	TRUE
8	$v^{(29)}$	quantitative count/$\{0, 1, 2, 3, \dots\}$		71.98	352.56	0	23.0	21,128	0.139	0.024	0.018	TRUE
9	$v^{(35)}$	quantitative censored/$\{0\} + (0, \infty)$		82,458,491.68	801,037,516.93	0	15,049,206.5	46,653,852,983	0.084	0.001	0.003	FALSE
10	$v^{(37)}$	quantitative censored/$\{0\} + (0, \infty)$		18,378,395.64	486,369,099.79	0	0.0	45,186,695,043	0.038	0.555	0.004	FALSE
11	$v^{(38)}$	quantitative censored/$\{0\} + (0, \infty)$		38,540.64	1,177,544.70	0	0.0	107,769,875	0.033	0.959	0.000	FALSE
12	$v^{(39)}$	quantitative censored/$\{0\} + (0, \infty)$		2,903,498.60	44,379,909.33	0	0.0	2,358,319,533	0.065	0.537	0.000	FALSE
13	$v^{(40)}$	quantitative censored/$\{0\} + (0, \infty)$		103,778,926.56	1,231,765,641.27	0	16,700,783.5	91,980,543,264	0.071	0.001	0.000	TRUE
14	$v^{(41)}$	quantitative censored/$\{0\} + (0, \infty)$		10,571,346.41	110,784,753.75	0	918,193.0	6,224,862,988	0.087	0.149	0.000	FALSE
15	$v^{(42)}$	quantitative censored/$\{0\} + (0, \infty)$		10,100,826.11	103,903,756.20	0	900,964.5	6,502,970,000	0.089	0.147	0.000	FALSE
16	$v^{(43)}$	transformation		-470,449.97	43,147,595.17	-5,013,884,175	0.0	1,071,222,119	-0.011	0.149	0.000	TRUE
17	$v^{(44)}$	quantitative censored/$\{0\} + (0, \infty)$		320,276.85	8,275,723.12	0	0.0	755,941,405	0.039	0.814	0.000	FALSE
18	$v^{(46)}$	quantitative censored/$\{0\} + (0, \infty)$		103,628,753.44	1,226,981,865.72	0	16,680,362.5	91,600,793,087	0.071	0.001	0.000	TRUE
19	$v^{(50)}$	quantitative censored/$\{0\} + (0, \infty)$		4,757,362.54	36,819,768.40	0	815,400.5	3,270,778,338	0.107	0.059	0.004	FALSE
20	$v^{(51)}$	quantitative censored/$\{0\} + (0, \infty)$		4,661,819.76	39,247,657.04	0	815,937.5	3,525,209,212	0.098	0.060	0.003	FALSE
21	$v^{(52)}$	quantitative censored/$\{0\} + (0, \infty)$		42,099,125.02	486,193,289.26	0	5,238,628.0	34,090,655,897	0.076	0.042	0.001	FALSE
22	$v^{(53)}$	quantitative censored/$\{0\} + (0, \infty)$		42,194,667.80	484,484,354.95	0	5,292,565.5	34,090,659,506	0.076	0.024	0.001	TRUE
23	$v^{(55)}$	quantitative censored/$\{0\} + (0, \infty)$		1,690,797.28	12,650,484.41	0	228,971.5	999,332,936	0.116	0.081	0.003	TRUE
24	$v^{(56)}$	quantitative censored/$\{0\} + (0, \infty)$		1,470,845.45	32,203,852.27	0	0.0	2,481,052,456	0.046	0.648	0.003	FALSE
25	$v^{(57)}$	quantitative censored/$\{0\} + (0, \infty)$		1,474,293.05	32,073,248.62	0	0.0	2,470,663,900	0.046	0.651	0.002	FALSE
26	$v^{(58)}$	quantitative censored/$\{0\} + (0, \infty)$		14,813,556.16	415,621,808.77	0	0.0	37,766,591,675	0.036	0.577	0.002	FALSE
27	$v^{(59)}$	quantitative censored/$\{0\} + (0, \infty)$		14,810,108.56	416,500,951.50	0	0.0	37,993,813,089	0.036	0.555	0.003	TRUE
28	$v^{(60)}$	quantitative censored/$\{0\} + (0, \infty)$		16,368,964.88	142,832,544.73	0	3,841,535.5	8,575,855,799	0.088	0.001	0.000	FALSE
29	$v^{(61)}$	quantitative censored/$\{0\} + (0, \infty)$		2,863,216.48	22,269,449.22	0	741,513.0	1,459,615,671	0.095	0.008	0.000	FALSE
30	$v^{(62)}$	quantitative censored/$\{0\} + (0, \infty)$		870,043.74	10,806,246.54	0	34,845.0	633,828,747	0.077	0.183	0.000	FALSE
31	$v^{(63)}$	quantitative censored/$\{0\} + (0, \infty)$		680,487.33	5,862,294.16	0	24,651.5	544,349,874	0.112	0.416	0.000	FALSE
32	$v^{(64)}$	quantitative censored/$\{0\} + (0, \infty)$		2,355,753.51	28,547,972.85	0	7,014.5	1,935,703,643	0.082	0.480	0.000	FALSE
33	$v^{(65)}$	quantitative censored/$\{0\} + (0, \infty)$		1,921,595.30	19,156,825.79	0	259,580.5	1,692,343,235	0.087	0.023	0.000	FALSE
34	$v^{(66)}$	quantitative censored/$\{0\} + (0, \infty)$		1,370,337.17	12,736,115.49	0	307,966.5	1,135,495,850	0.083	0.042	0.000	FALSE
35	$v^{(67)}$	quantitative censored/$\{0\} + (0, \infty)$		536,486.76	9,627,847.98	0	0.0	1,127,689,569	0.056	0.528	0.000	FALSE
36	$v^{(68)}$	quantitative censored/$\{0\} + (0, \infty)$		10,867,763.89	118,363,996.70	0	1,221,285.5	7,670,141,025	0.081	0.001	0.000	FALSE
37	$v^{(69)}$	quantitative censored/$\{0\} + (0, \infty)$		279,849.93	2,581,569.73	0	71,318.5	173,236,547	0.081	0.001	0.006	FALSE
38	$v^{(71)}$	quantitative censored/$\{0\} + (0, \infty)$		2,623,440.43	94,746,048.39	0	17,635.5	7,431,901,017	0.028	0.043	0.000	FALSE
39	$v^{(72)}$	quantitative censored/$\{0\} + (0, \infty)$		2,297,722.73	94,491,660.23	0	0.0	7,420,334,674	0.024	0.967	0.000	FALSE
40	$v^{(74)}$	quantitative censored/$\{0\} + (0, \infty)$		3,591,277.23	60,231,381.10	0	400,307.0	5,400,002,120	0.053	0.036	0.002	FALSE
41	$v^{(75)}$	quantitative censored/$\{0\} + (0, \infty)$		790,017.38	17,262,908.95	0	28,024.5	1,770,356,435	0.044	0.214	0.000	FALSE
42	$v^{(78)}$	quantitative censored/$\{0\} + (0, \infty)$		44,301,954.10	394,013,832.22	0	8,561,257.0	23,053,878,516	0.091	0.001	0.000	TRUE
43	$v^{(80)}$	quantitative censored/$\{0\} + (0, \infty)$		58,484.80	1,169,379.70	0	0.0	113,985,767	0.050	0.865	0.000	FALSE
44	$v^{(82)}$	quantitative censored/$\{0\} + (0, \infty)$		7,430,381.65	72,748,185.40	0	1,489,051.0	5,900,441,845	0.082	0.040	0.000	FALSE
45	$v^{(83)}$	quantitative censored/$\{0\} + (0, \infty)$		12,566,182.82	426,825,860.42	0	1,377,768.5	54,220,350,000	0.026	0.040	0.000	FALSE
46	$v^{(84)}$	quantitative censored/$\{0\} + (0, \infty)$		486,993.28	8,441,914.47	0	22,422.5	760,131,228	0.055	0.297	0.000	FALSE
47	$v^{(86)}$	quantitative censored/$\{0\} + (0, \infty)$		3,942,170.18	89,568,339.77	0	0.0	6,408,681,445	0.044	0.678	0.001	FALSE
48	$v^{(87)}$	quantitative censored/$\{0\} + (0, \infty)$		20.57	276.60	0	0.0	18,150	0.074	0.696	0.030	TRUE

TABLE 4. Variable Restrictions

No.	Variable	Restrictions
1	$v^{(4)}$	$\mathcal{I}(v_i^{(4)} \in \{1,2\}, v_i^{(21)} = 0) + \mathcal{I}(v_i^{(4)} = 3, v_i^{(21)} > 0)$
2	$v^{(21)}$	$\mathcal{I}(v_i^{(21)} \geq v_i^{(28)})\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})[\mathcal{I}(v_i^{(4)} \in \{1,2\}, v_i^{(21)} = 0) + \mathcal{I}(v_i^{(4)} = 3, v_i^{(21)} > 0)]$
3	$v^{(22)}$	$\mathcal{I}(v_i^{(22)} \neq 0)\mathcal{I}(v_i^{(22)} \geq \max\{v_i^{(24)}, v_i^{(25)}, v_i^{(29)}, v_i^{(87)}\})\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})\mathcal{I}(v_i^{(22)} \geq 1, v_i^{(60)} > 1)\mathcal{I}(v_i^{(87)} \leq v_i^{(22)})$
4	$v^{(24)}$	$\mathcal{I}(v_i^{(24)} \leq v_i^{(22)})[\mathcal{I}(0 < v_i^{(25)} < v_i^{(24)} \geq 2) + \mathcal{I}(v_i^{(24)} = v_i^{(25)} = 0)]$
5	$v^{(25)}$	$\mathcal{I}(v_i^{(25)} \leq v_i^{(22)})[\mathcal{I}(0 < v_i^{(25)} < v_i^{(24)} \geq 2) + \mathcal{I}(v_i^{(24)} = v_i^{(25)} = 0)]$
6	$v^{(27)}$	$\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})\mathcal{I}(v_i^{(27)} \geq 1, v_i^{(60)} > 1)$
7	$v^{(28)}$	$\mathcal{I}(v_i^{(21)} \geq v_i^{(28)})$
8	$v^{(29)}$	$\mathcal{I}(v_i^{(29)} \leq v_i^{(22)})$
9	$v^{(35)}$	$\mathcal{I}(v_i^{(35)} > 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
10	$v^{(37)}$	$\mathcal{I}(v_i^{(37)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})[\mathcal{I}(v_i^{(37)} = v_i^{(59)} = 0) + \mathcal{I}(v_i^{(59)} \geq 1, v_i^{(37)} \geq 1)]$
11	$v^{(38)}$	$\mathcal{I}(v_i^{(38)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
12	$v^{(39)}$	$\mathcal{I}(v_i^{(39)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
13	$v^{(40)}$	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
14	$v^{(41)}$	$\mathcal{I}(v_i^{(41)} \neq 1)\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
15	$v^{(42)}$	$\mathcal{I}(v_i^{(42)} \neq 1)\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
16	$v^{(43)}$	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
17	$v^{(44)}$	$\mathcal{I}(v_i^{(44)} \neq 1)\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})$
18	$v^{(46)}$	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})$
19	$v^{(50)}$	$\mathcal{I}(v_i^{(50)} \neq 1)\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$
20	$v^{(51)}$	$\mathcal{I}(v_i^{(51)} \neq 1)\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$
21	$v^{(52)}$	$\mathcal{I}(v_i^{(52)} \neq 1)\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$
22	$v^{(53)}$	$\mathcal{I}(v_i^{(53)} \geq v_i^{(55)})\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$
23	$v^{(55)}$	$\mathcal{I}(v_i^{(55)} \leq v_i^{(53)})$
24	$v^{(56)}$	$\mathcal{I}(v_i^{(56)} \neq 1)\mathcal{I}(v_i^{(56)} - v_i^{(57)} + v_i^{(58)} = v_i^{(59)})$
25	$v^{(57)}$	$\mathcal{I}(v_i^{(57)} \neq 1)\mathcal{I}(v_i^{(56)} - v_i^{(57)} + v_i^{(58)} = v_i^{(59)})$
26	$v^{(58)}$	$\mathcal{I}(v_i^{(58)} \neq 1)\mathcal{I}(v_i^{(56)} - v_i^{(57)} + v_i^{(58)} = v_i^{(59)})$
27	$v^{(59)}$	$\mathcal{I}(v_i^{(56)} - v_i^{(57)} + v_i^{(58)} = v_i^{(59)})[\mathcal{I}(v_i^{(37)} = v_i^{(59)} = 0) + \mathcal{I}(v_i^{(59)} \geq 1, v_i^{(37)} \geq 1)]$
28	$v^{(60)}$	$\mathcal{I}(v_i^{(60)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})\mathcal{I}(v_i^{(22)} \geq 1, v_i^{(60)} > 1)\mathcal{I}(v_i^{(27)} \geq 1, v_i^{(60)} > 1)$
29	$v^{(61)}$	$\mathcal{I}(v_i^{(61)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
30	$v^{(62)}$	$\mathcal{I}(v_i^{(62)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
31	$v^{(63)}$	$\mathcal{I}(v_i^{(63)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
32	$v^{(64)}$	$\mathcal{I}(v_i^{(64)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
33	$v^{(65)}$	$\mathcal{I}(v_i^{(65)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
34	$v^{(66)}$	$\mathcal{I}(v_i^{(66)} \neq 1)\mathcal{I}(v_i^{(66)} \geq v_i^{(67)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
35	$v^{(67)}$	$\mathcal{I}(v_i^{(67)} \neq 1)\mathcal{I}(v_i^{(67)} \leq v_i^{(66)})$
36	$v^{(68)}$	$\mathcal{I}(v_i^{(68)} \neq 1)\mathcal{I}(v_i^{(68)} \geq v_i^{(69)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
37	$v^{(69)}$	$\mathcal{I}(v_i^{(69)} > 1)\mathcal{I}(v_i^{(69)} \leq v_i^{(68)})$
38	$v^{(71)}$	$\mathcal{I}(v_i^{(71)} \neq 1)\mathcal{I}(v_i^{(71)} \geq v_i^{(72)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
39	$v^{(72)}$	$\mathcal{I}(v_i^{(72)} \neq 1)\mathcal{I}(v_i^{(72)} \leq v_i^{(71)})$
40	$v^{(74)}$	$\mathcal{I}(v_i^{(74)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
41	$v^{(75)}$	$\mathcal{I}(v_i^{(75)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
42	$v^{(78)}$	$\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
43	$v^{(80)}$	$\mathcal{I}(v_i^{(80)} \neq 1)$
44	$v^{(82)}$	$\mathcal{I}(v_i^{(82)} \neq 1)$
45	$v^{(83)}$	$\mathcal{I}(v_i^{(83)} \neq 1)\mathcal{I}(v_i^{(83)} \geq v_i^{(84)})$
46	$v^{(84)}$	$\mathcal{I}(v_i^{(84)} \neq 1)\mathcal{I}(v_i^{(84)} \leq v_i^{(83)})$
47	$v^{(86)}$	$\mathcal{I}(v_i^{(86)} \neq 1)[\mathcal{I}(v_i^{(87)} = v_i^{(86)} = 0) + \mathcal{I}(v_i^{(87)} \geq 1, v_i^{(86)} > 1)]$
48	$v^{(87)}$	$\mathcal{I}(v_i^{(87)} \leq v_i^{(22)})[\mathcal{I}(v_i^{(87)} = v_i^{(86)} = 0) + \mathcal{I}(v_i^{(87)} \geq 1, v_i^{(86)} > 1)]$

TABLE 5. Precondition of Sampling Trajectories

Trajectory	Precondition	P-Value	Validation	V-Value
T_1	$\mathcal{I}(v_i^{(80)} \neq 1)$	0	-	
T_2	$\mathcal{I}(v_i^{(82)} \neq 1)$	0	-	
T_3	$\mathcal{I}(v_i^{(83)} \neq 1)\mathcal{I}(v_i^{(84)} \neq 1)\mathcal{I}(v_i^{(83)} \geq v_i^{(84)})$	0	-	
T_4	$\mathcal{I}(v_i^{(50)} \neq 1)\mathcal{I}(v_i^{(51)} \neq 1)\mathcal{I}(v_i^{(52)} \neq 1)\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$	0	-	
T_5	$\mathcal{I}(v_i^{(53)} \geq v_i^{(55)})$	0	$\mathcal{I}(v_i^{(50)} \neq 1)\mathcal{I}(v_i^{(51)} \neq 1)\mathcal{I}(v_i^{(52)} \neq 1)\mathcal{I}(v_i^{(50)} - v_i^{(51)} + v_i^{(52)} = v_i^{(53)})$	1
T_6	$\mathcal{I}(v_i^{(56)} \neq 1)\mathcal{I}(v_i^{(57)} \neq 1)\mathcal{I}(v_i^{(58)} \neq 1)\mathcal{I}(v_i^{(56)} - v_i^{(57)} + v_i^{(58)} = v_i^{(59)})$	0	$\mathcal{I}(v_i^{(37)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$	1
T_7	$\mathcal{I}(v_i^{(41)} \neq 1)\mathcal{I}(v_i^{(42)} \neq 1)\mathcal{I}(v_i^{(44)} \neq 1)\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$	0	$\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$	1

Notes:

TABLE 6. Full Conditional Model Specifications (Semi-Parametric)

Trajectory / Preconditon	
T_1	$\mathcal{CTN}(v^{(80)})\|S)$
T_2	$\mathcal{CTN}(v^{(82)})\|S)$
T_3	$\mathcal{CTN}(v^{(83)})\|S) \searrow$
	$\quad \mathcal{CTN}(v^{(84)})\|S)$
T_4	$\mathcal{CTN}(v^{(50)})\|S) \searrow$
	$\quad \mathcal{CTN}(v^{(51)})\|S) \searrow$
	$\quad\quad \mathcal{CTN}(v^{(52)})\|S) \searrow$
	$\quad\quad\quad \mathcal{CTN}(v^{(53)})\|S) \searrow$
	$\quad\quad\quad\quad \mathcal{CTN}(v^{(55)})\|S)$
T_5	$\mathcal{CTN}(v^{(55)})\|S)$
T_6	

Notes: D denotes all data except the considered variable; val. refers to a validated variable value in case an equality restriction involving the variable is valid.

TABLE 7. Full Conditional Model Specifications (Non-Parametric)

No.	Variable	Order of Sampling ($\rightarrow$) / Derivation ($\Rightarrow$)	Involved Joint Distribution (non-parametric)
1	$v^{(4)}$	quali. / categ.	
2	$v^{(21)}$	quant. / count	$\mathcal{I}(v_i^{(21)} \geq v_i^{(28)})\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})$
3	$v^{(22)}$	quant. / count	$\mathcal{I}(v_i^{(22)} \neq 0)\mathcal{I}(v_i^{(22)} \geq \max\{v_i^{(24)}, v_i^{(25)}, v_i^{(29)}, v_i^{(87)}\})\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})\mathcal{I}(v_i^{(22)} \geq 1, v_i^{(60)} > 0)$
4	$v^{(24)}$	quant. / count	$\mathcal{I}(v_i^{(24)} \leq v_i^{(22)})[\mathcal{I}(0 < v_i^{(25)} < v_i^{(24)} \geq 2) + \mathcal{I}(v_i^{(24)} = v_i^{(25)} = 0)]$
5	$v^{(25)}$	quant. / transformed	$\mathcal{I}(v_i^{(25)} \leq v_i^{(22)})[\mathcal{I}(0 < v_i^{(25)} < v_i^{(24)} \geq 2) + \mathcal{I}(v_i^{(24)} = v_i^{(25)} = 0)]$
6	$v^{(27)}$	quant. / count	$\mathcal{I}(v_i^{(21)} + v_i^{(22)} = v_i^{(27)})$
7	$v^{(28)}$	quant. / count	$\mathcal{I}(v_i^{(21)} \leq v_i^{(28)})$
8	$v^{(29)}$	quant. / count	$\mathcal{I}(v_i^{(29)} \leq v_i^{(22)})$
9	$v^{(35)}$	quant. / float	$\mathcal{I}(v_i^{(35)} > 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
10	$v^{(37)}$	quant. / float	$\mathcal{I}(v_i^{(37)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
11	$v^{(38)}$	quant. / float	$\mathcal{I}(v_i^{(38)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
12	$v^{(39)}$	quant. / float	$\mathcal{I}(v_i^{(39)} \neq 1)\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
13	$v^{(40)}$	quant. / float	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})\mathcal{I}(v_i^{(35)} + v_i^{(37)} + v_i^{(38)} + v_i^{(39)} = v_i^{(40)})$
14	$v^{(41)}$	quant. / float	$\mathcal{I}(v_i^{(41)} \neq 1)\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
15	$v^{(42)}$	quant. / float	$\mathcal{I}(v_i^{(42)} \neq 1)\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
16	$v^{(43)}$	quant. / transformation	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})\mathcal{I}(v_i^{(42)} - v_i^{(41)} = v_i^{(43)})$
17	$v^{(44)}$	quant. / float	$\mathcal{I}(v_i^{(44)} \neq 1)\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})$
18	$v^{(46)}$	quant. / float	$\mathcal{I}(v_i^{(40)} + v_i^{(43)} + v_i^{(44)} = v_i^{(46)})$
19	$v^{(50)}$		
20	$v^{(51)}$	}	$(1-\mathcal{I}(\text{val}.v^{(53)}))\text{CART}(v^{(50)}\|D\setminus\{v^{(51)}, v^{52}, v^{(53)}\})\text{CART}(v^{(51)}\|D\setminus\{v^{52}, v^{(53)}\})\text{CART}(v^{(52)}\|D\setminus\{v^{(53)}\})\text{CART}(v^{(55)}\|D\|_{\leq v^{(53)}})+$
21	$v^{(52)}$	$v^{(50)} \rightarrow v^{(51)} \rightarrow v^{(52)} \Rightarrow v^{(53)} \rightarrow v^{(55)}$	$\mathcal{I}(\text{val}.v^{(53)})\text{CART}(v^{(55)}\|D\|_{\leq v^{(53)}})$
22	$v^{(53)}$		
23	$v^{(55)}$		
24	$v^{(56)}$	}	$\mathcal{I}(\text{val}.v^{(37)} \geq$
25	$v^{(57)}$	$v^{(56)} \rightarrow v^{(57)} \rightarrow / \Rightarrow v^{(58)} \Rightarrow v^{(59)}$	$1)\text{CART}(v^{(56)}\|D \setminus \{v^{(57)}, v^{(58)}, v^{(59)}\}\|_{v^{(37)}\geq 1})\text{CART}(v^{(57)}\|D \setminus \{v^{(58)}, v^{(59)}\}\|_{v^{(37)}\geq 1})\text{CART}(v^{(58)}\|D \setminus \{v^{(59)}\}\|_{v^{(37)}\geq 1})+$
26	$v^{(58)}$	$(\rightarrow v^{(35)} \rightarrow v^{(37)} \rightarrow v^{(38)} \rightarrow v^{(39)} \Rightarrow$	$\mathcal{I}(\text{val}.v^{(37)} = 0)\text{CART}(v^{(56)}\|D \setminus \{v^{(57)}, v^{(58)}, v^{(59)}\}\|_{v^{(37)}=0})\text{CART}(v^{(57)}\|D \setminus \{v^{(58)}, v^{(59)}\}\|_{v^{(37)}=0})+$
27	$v^{(59)}$	$v^{(40)})$	$\mathcal{I}(\text{val}.v^{(37)})\text{CART}(v^{(56)}\|D\setminus\{v^{(57)}, v^{(58)}, v^{(59)}\})\text{CART}(v^{(57)}\|D\setminus\{v^{(58)}, v^{(59)}\})\text{CART}(v^{(58)}\|D\setminus\{v^{(59)}\})$ and continue with $v^{(35)}$
28	$v^{(60)}$	quant. / float	$\mathcal{I}(v_i^{(60)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})\mathcal{I}(v_i \geq 1, v_i \geq 1)]$
29	$v^{(61)}$	quant. / float	$\mathcal{I}(v_i^{(61)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
30	$v^{(62)}$	quant. / float	$\mathcal{I}(v_i^{(62)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
31	$v^{(63)}$	quant. / float	$\mathcal{I}(v_i^{(63)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
32	$v^{(64)}$	quant. / float	$\mathcal{I}(v_i^{(64)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
33	$v^{(65)}$	quant. / float	$\mathcal{I}(v_i^{(65)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
34	$v^{(66)}$	quant. / float	$\mathcal{I}(v_i^{(66)} \neq 1)\mathcal{I}(v_i^{(66)} \geq v_i^{(67)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
35	$v^{(67)}$	quant. / float	$\mathcal{I}(v_i^{(67)} \neq 1)\mathcal{I}(v_i^{(67)} \leq v_i^{(66)})$
36	$v^{(68)}$	quant. / float	$\mathcal{I}(v_i^{(68)} \neq 1)\mathcal{I}(v_i^{(68)} \geq v_i^{(69)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
37	$v^{(69)}$	quant. / float	$\mathcal{I}(v_i^{(69)} > 1)\mathcal{I}(v_i^{(69)} \leq v_i^{(68)})$
38	$v^{(71)}$	quant. / float	$\mathcal{I}(v_i^{(71)} \neq 1)\mathcal{I}(v_i^{(71)} \geq v_i^{(72)})\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
39	$v^{(72)}$	quant. / float	$\mathcal{I}(v_i^{(72)} \neq 1)\mathcal{I}(v_i^{(72)} \leq v_i^{(71)})$
40	$v^{(74)}$	quant. / float	$\mathcal{I}(v_i^{(74)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
41	$v^{(75)}$	quant. / float	$\mathcal{I}(v_i^{(75)} \neq 1)\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
42	$v^{(78)}$	quant. / float	$\mathcal{I}(v_i^{(60)} + v_i^{(61)} + v_i^{(62)} + v_i^{(63)} + v_i^{(64)} + v_i^{(65)} + v_i^{(66)} + v_i^{(68)} + v_i^{(71)} + v_i^{(74)} + v_i^{(75)} = v_i^{(78)})$
43	$v^{(80)}$	$v^{(80)}$	$\text{CART}(v^{(80)}\|D)$
44	$v^{(82)}$	$v^{(82)}$	$\text{CART}(v^{(82)}\|D)$
45	$v^{(83)}$	}	$\text{CART}(v^{(83)}\|D \setminus \{v^{(84)}\})\text{CART}(v^{(84)}\|D \setminus \{v^{(83)}\}\|_{\leq v^{(83)}})$
46	$v^{(84)}$	$v^{(83)} \rightarrow v^{(84)}$	
47	$v^{(86)}$	quant. / float	$\mathcal{I}(v_i^{(86)} \neq 1)[\mathcal{I}(v_i^{(87)} = v_i^{(86)} = 0) + \mathcal{I}(v_i^{(87)} \geq 1, v_i^{(86)} > 1)]$
48	$v^{(87)}$	quant. / float	$\mathcal{I}(v_i^{(87)} \leq v_i^{(22)})[\mathcal{I}(v_i^{(87)} = v_i^{(86)} = 0) + \mathcal{I}(v_i^{(87)} \geq 1, v_i^{(86)} > 1)]$

Notes: D denotes all data except the considered variable; val. refers to a validated variable value in case an equality restriction involving the variable is valid.

TABLE 8. Sampling from Distributions and Parameter Estimation

Distribution	Label	Parameters	Estimation Approach	Procedure
truncated normal	$\mathcal{TN}$	$\mu, \sigma^2, \ell, \upsilon$	moment based estimation	10 iterations of first order Taylor approximation of the system of moment conditions for expected value and variance
censored truncated normal	$\mathcal{CTN}$	$p, \mu, \sigma^2, \ell, \upsilon$		
censored normal	$\mathcal{CN}$	p, μ, σ^2		
censored truncated Poisson	$\mathcal{CTP}$	$p, \lambda, \ell, \upsilon$		
censored Poisson	$\mathcal{CP}$	p, λ		
truncated Poisson	$\mathcal{TP}$	$\mu, \sigma^2, \ell, \upsilon$		

TABLE 9. Sampling from Distributions and Parameter Estimation

Variable	Distribution	μ	σ^2	ℓ	ν	p	S
$v^{(80)}$	$\mathcal{CTN}(v^{(84)})\|S)$					$\Pr(v=0)$	D
$v^{(82)}$	$\mathcal{CTN}(v^{(84)})\|S)$					$\Pr(v=0)$	D
$v^{(83)}$	$\mathcal{CTN}(v^{(83)})\|S)$					$\Pr(v=0)$	$D \setminus v^{(84)}$
$v^{(84)}$	$\mathcal{CTN}(v^{(84)})\|S)$					$\Pr(v=0)$	$D \setminus v^{(83)}$

Figures

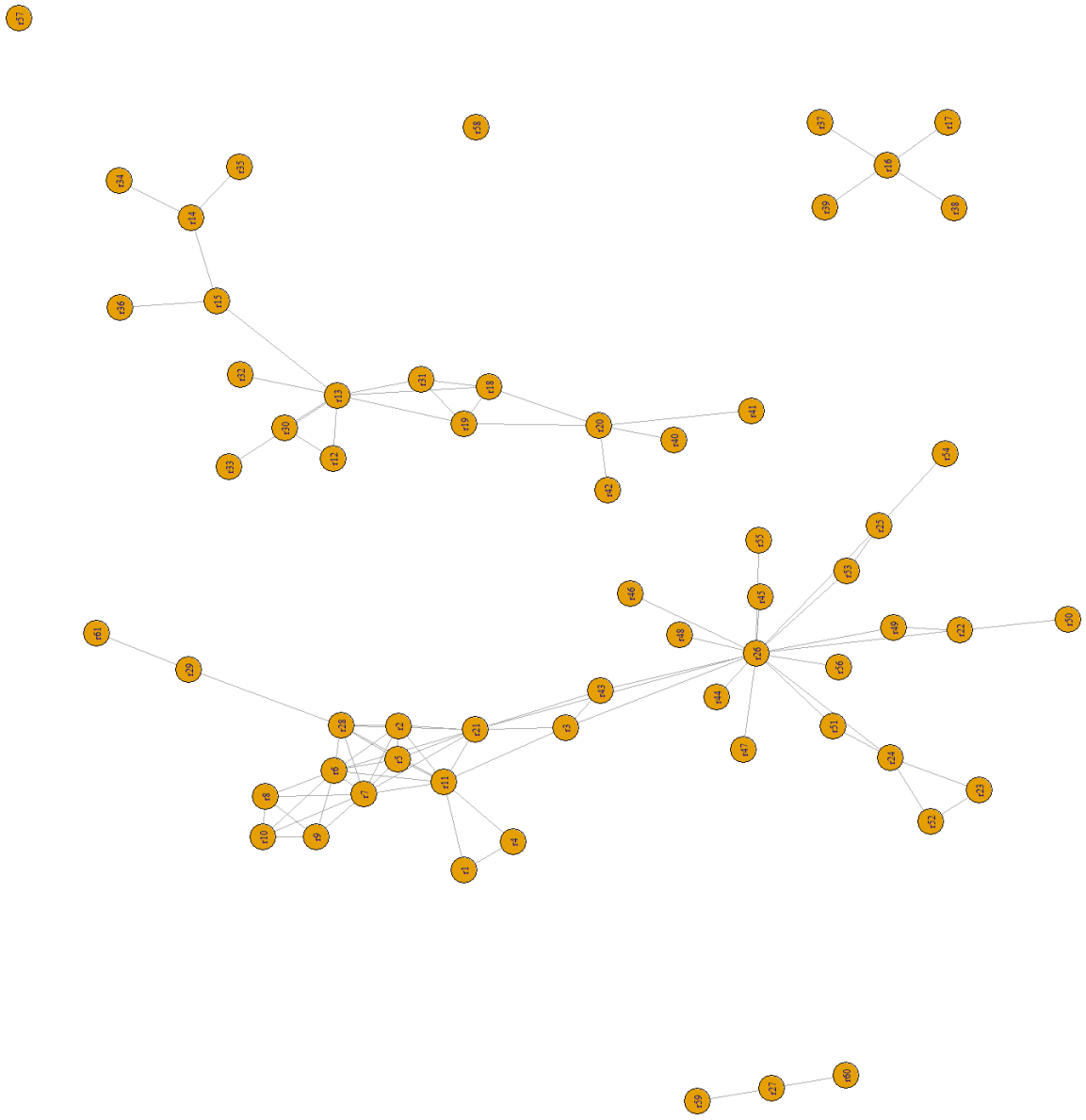

FIGURE 1. Rules related via variables.

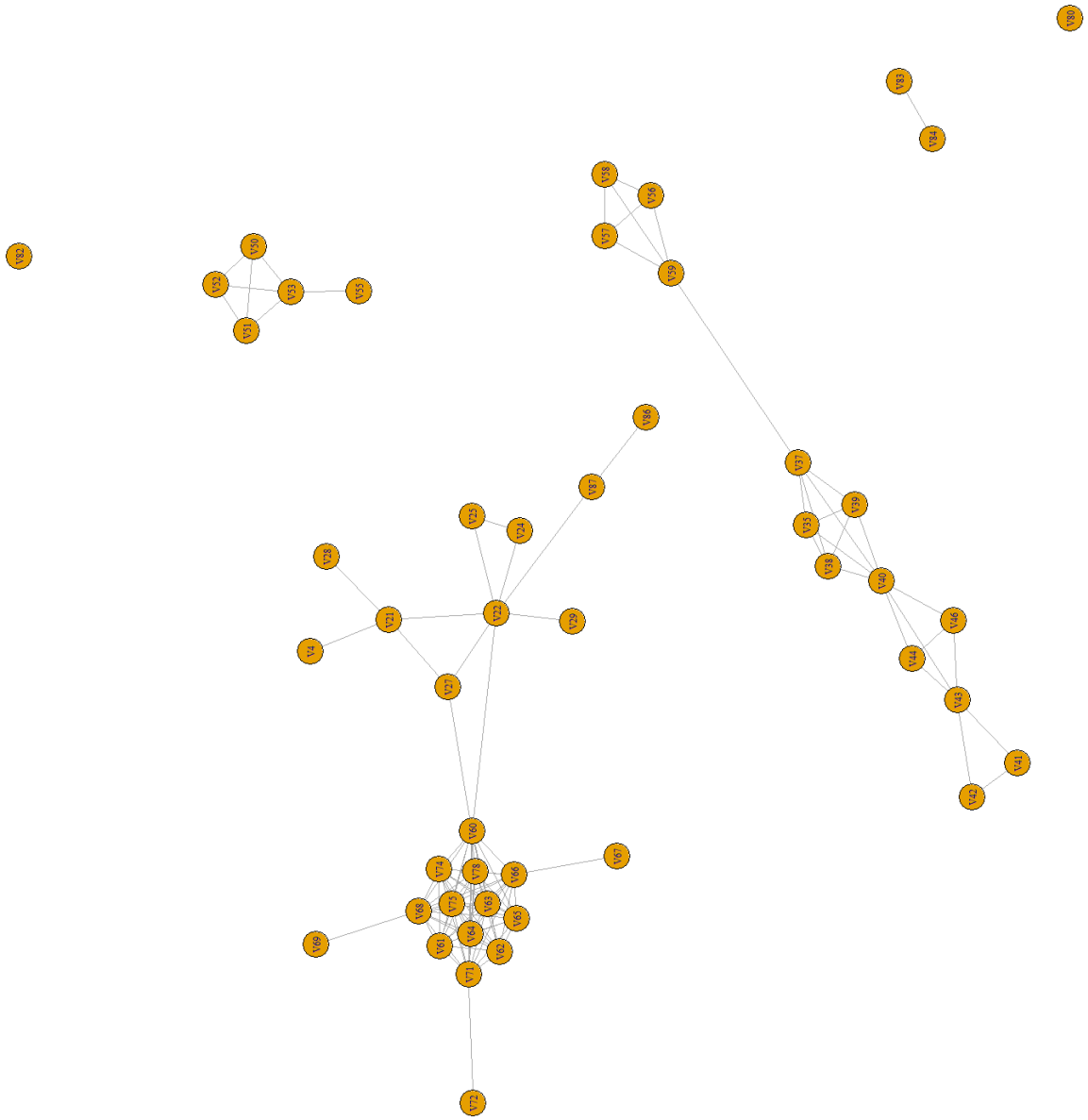

FIGURE 2. Variables related via rules.

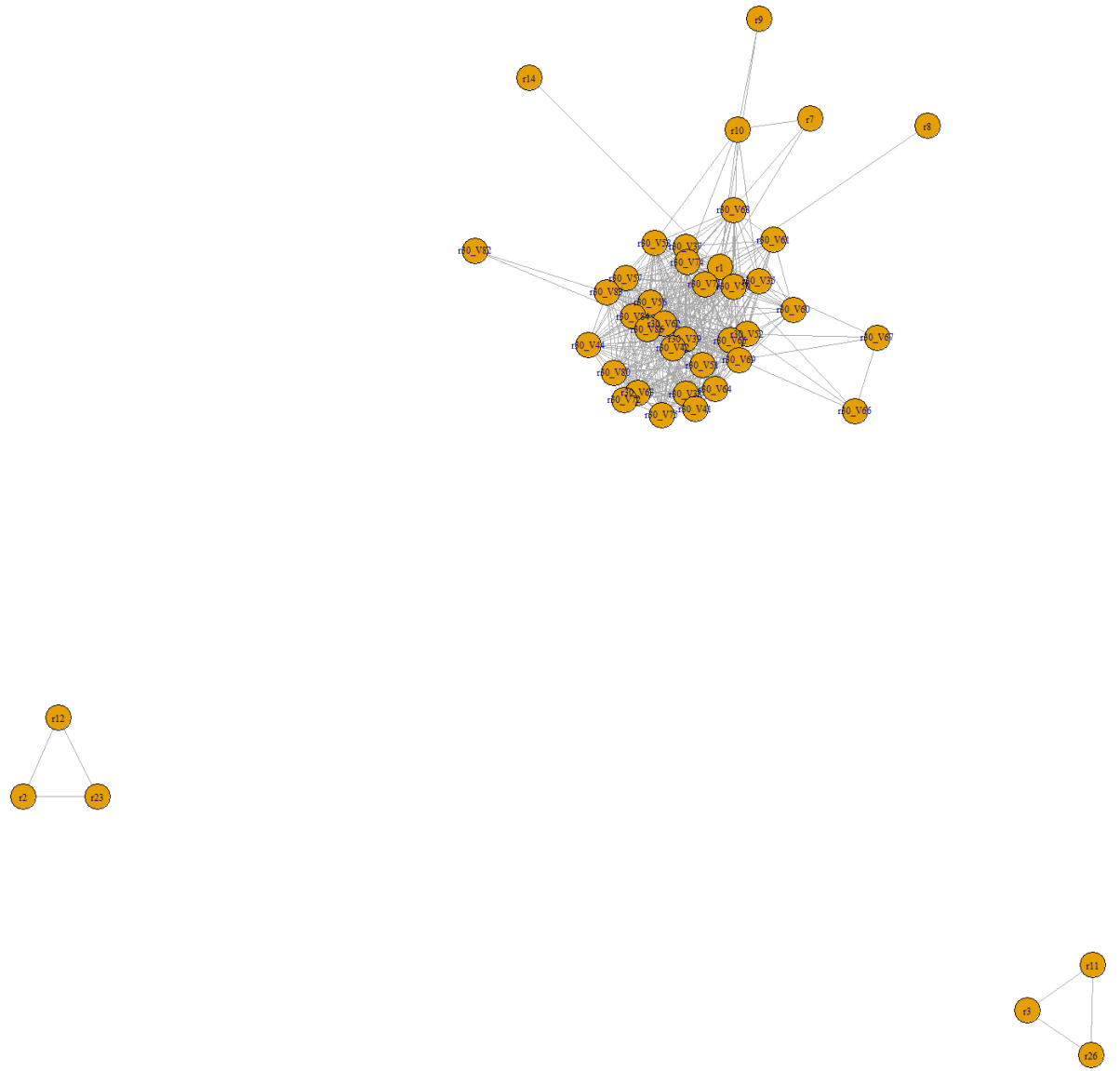

FIGURE 3. Rules that are jointly subject to violation.

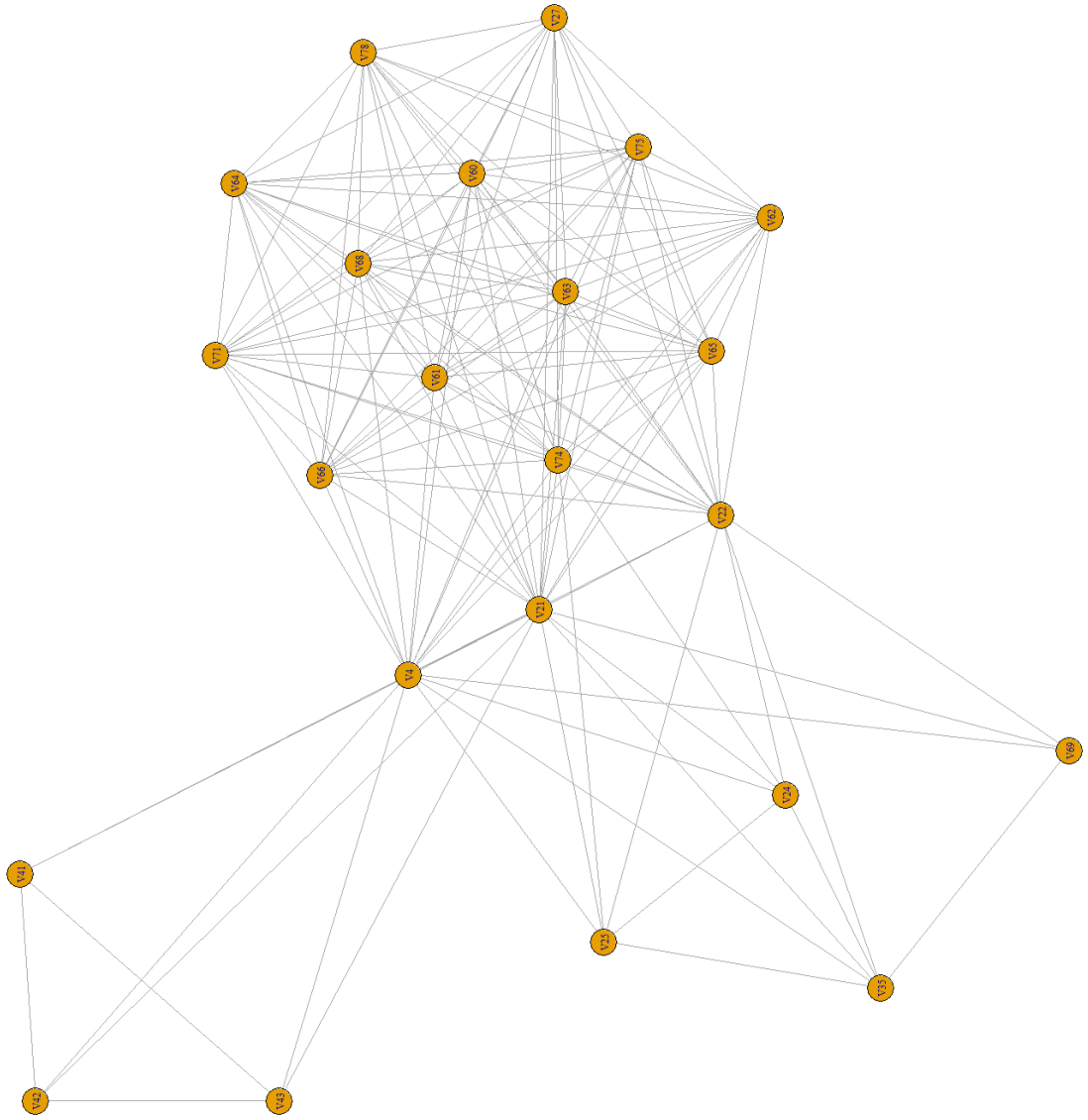

FIGURE 4. Variables that are jointly subject to violation.