Methods in “Synthetic Individual Income Tax Data: Promises and Challenges,” by Bowen et al.

Synthesis uses three steps:

1. ~~Preparation~~
   1. ~~Specific to the dataset~~
2. ~~Generation~~
   1. ~~Categorical variables~~
      1. ~~Simultaneously synthesize sets of related variables from their joint distribution (draw sets at random from the joint distribution)~~
   2. Continuous variables
      1. Apply a sequence of regression trees
      2. For a given synthesized value, draw an observation at random from the final node
      3. Variables synthesized later in the sequence tend to be noisier, so order matters
         1. Ordering based on the weighted sum of absolute values yields the highest quality data
   3. Dealing with Outliers
      1. When synthesizing a variable
         1. Since a tree is designed to minimize the sum of squared errors, large values carry more weight in estimation
         2. Yeo-Johnson transformation is used prior to synthesis, and after predicting a value with the regression tree
         3. Yeo-Johnson is similar to log or box-cox transformation but is applicable to variables with negative and zero values
   4. Smoothing
      1. Noise is added to sampled values to further protect privacy
      2. Noise is drawn from a zero-mean normal distribution
         1. Variance for that distribution is separately determined for each observation
         2. More noise is added when observations come from a unique part of the distribution, and less is added when they come from a common part of the distribution
3. Post-processing
   1. Adding constraints
      1. Imposing hard boundaries, z-bounding, etc.
      2. Whether these are needed depends on the data