An Exploratory Research on Optimization of CFS Sampling Design

GASP 2019

Mehdi Hashemipour, PhD

Julie Parker

Commodity Flow Survey (CFS)

- Component of the Economic Census
  - Partnership: Conducted by US Census Bureau with Funding from Bureau of Transportation Statistics (BTS) and
  - Every 5 years

- Data provide a comprehensive, multimodal picture of national freight flows for sampled industries

- Information collected:
  - Commodity
  - Value
  - Weight
  - Mode of Transportation
  - Destination
  - Temperature Control, HAZMAT, Exports

Goals of the Survey

- The goal of the Commodity Flow Survey (CFS) is to make reliable estimates of:
  - Annual total value of shipments,
  - Annual total tonnage of shipments,
  - Annual total ton-miles of shipments, and other quantities.

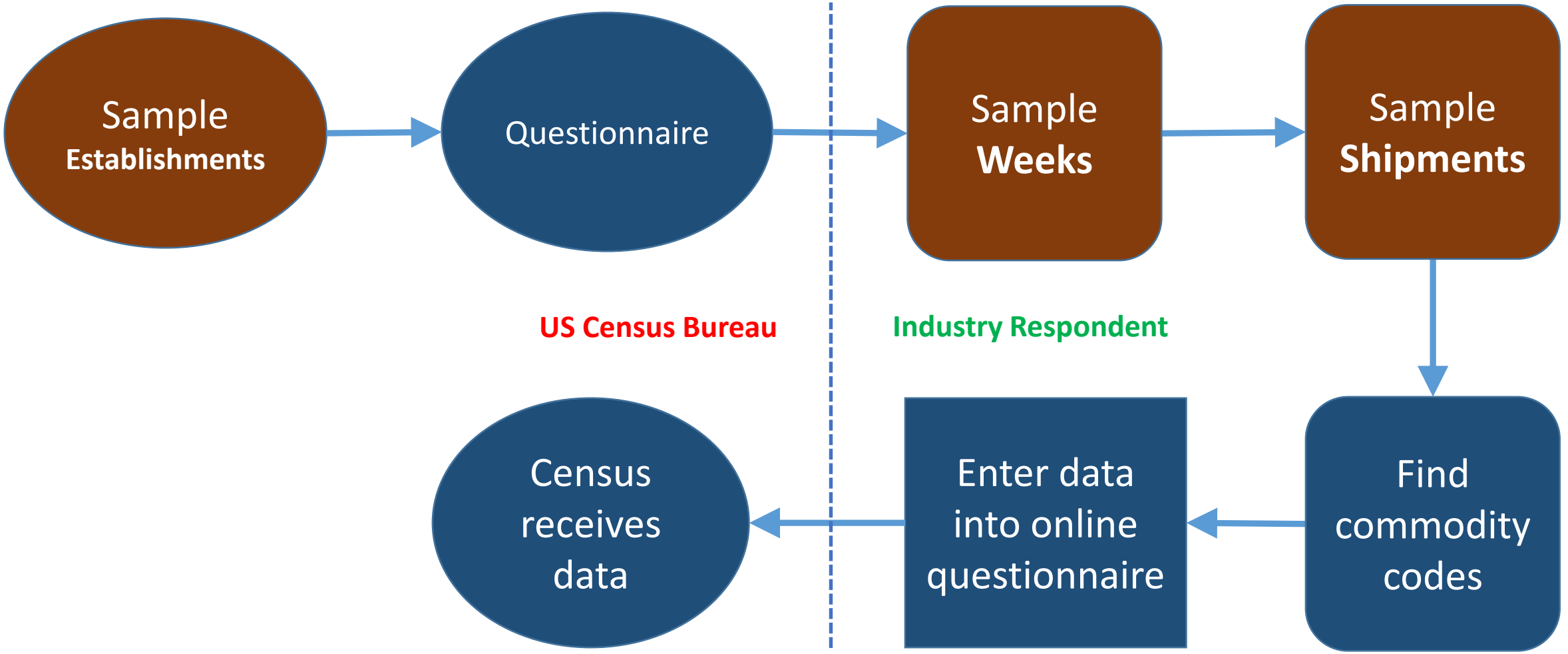
- We want to make these estimates for various areas of interest such as
  - Geographic origin of the shipment
  - Geographic destination of the shipment
  - Industry of the establishment making the shipment
  - Commodity shipped
  - Mode of transportation of the shipment
  - Other variables

Target Population

- The frame population (for the first stage of the sample) consists of establishments that are:
  - likely to ship physical products,
  - likely to be in business in 2017, and
  - for which we have data in the Business Register (the database from which we construct the frame).

- Shipper survey of ~100,000 establishments
  - Mining
  - Manufacturing
  - Wholesale Trade
  - Select Retail and Services

Current CFS Process and Sample Design



Current CFS Process and Sample Design cont.

- The sample design of the first stage is **stratified simple random sampling** without replacement of establishments in the in-scope industries.
  - The strata are (establishment) **Origin x Industry x Measure of Size (MOS) size** classes

- **Sample Size:**
- To determine the sample size in each stratum, we assign a target coefficient of variation (**CV**) to each **origin** x **industry** cell.
  - Within each **origin** x **industry** cell, we determine the boundaries of the MOS size class cells that minimize the sample size needed to achieve the target CV, where the sample is allocated according to the **Neyman Allocation** among the noncertainty strata.

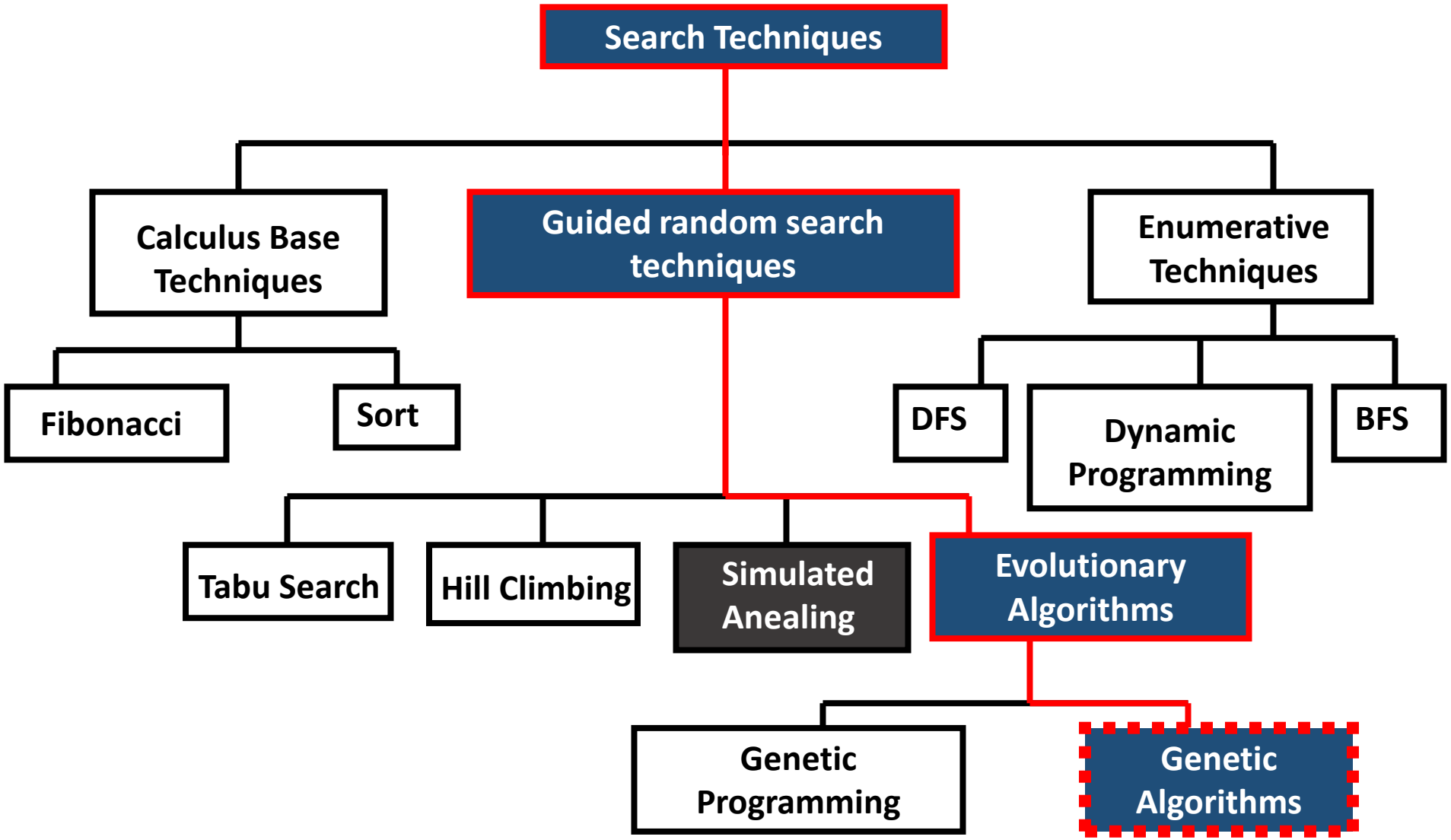
Joint Stratification and Allocation Method

- In this exploratory research, we applied the optimal stratification and allocation method based on **Genetic Algorithm (GA)** and **Simulated Annealing (SA)** in a CFS like scenario.

- **Objective:**
  - With this method, we aim at minimizing the total sample cost while satisfying the target Coefficient of Variation (CV) constraints.

- Experimenting these methods by using domain variables such as different CFS areas and Industries may result in different CFS sampling design.

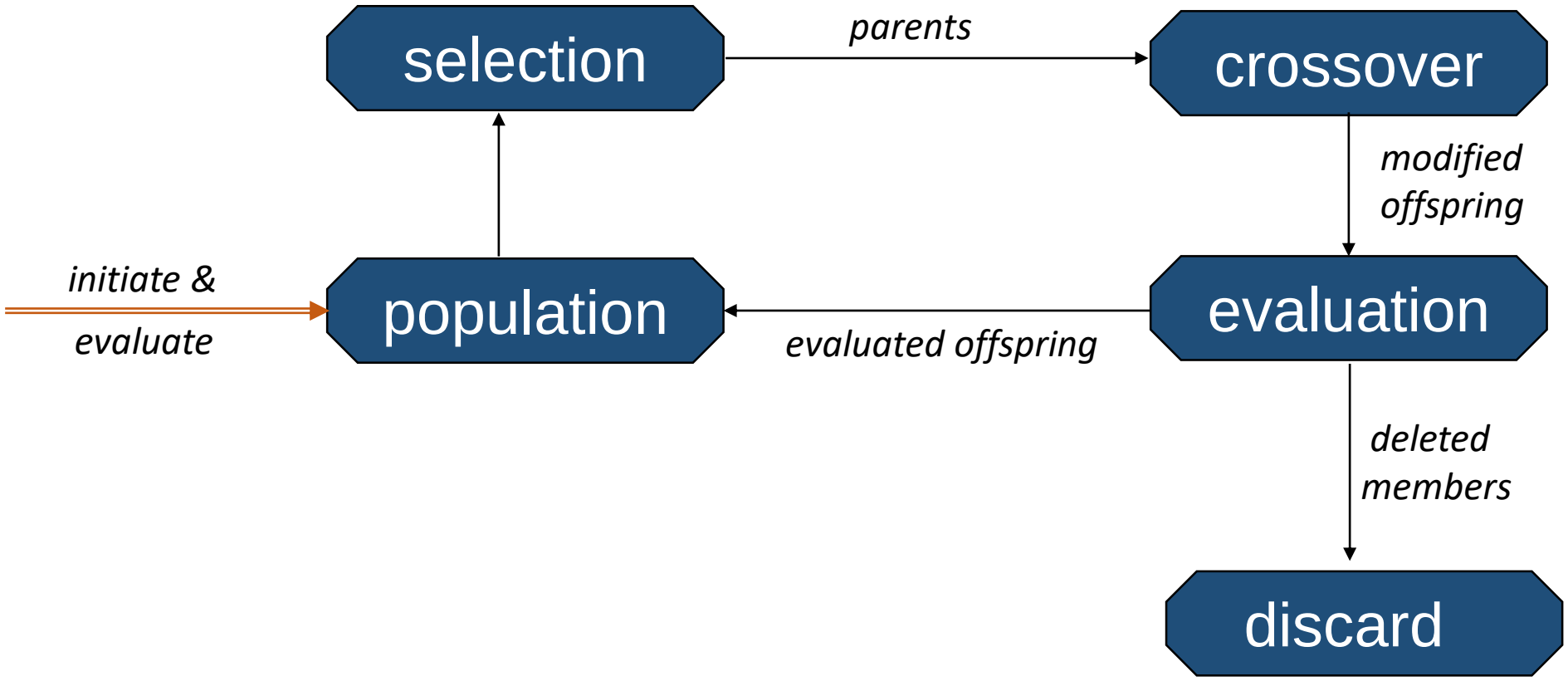
Classes of Search Techniques



Genetic Algorithms (GA) OVERVIEW

- A class of probabilistic optimization algorithms
- Inspired by the biological evolution process
- Uses concepts of “Natural Selection” and “Genetic Inheritance” (Darwin 1859)
- known about the underlying search space
- Widely-used in business, science and engineering
- A genetic algorithm maintains a population of candidate solutions for the problem at hand, and makes it evolve by iteratively applying a set of stochastic operators

The Evolutionary Cycle



Optimal stratification and allocation based on Genetic Algorithm for CFS

- The optimal stratification and allocation method aims at minimizing the total sample cost while satisfying the precision (CV) constraints.
- This method explores the set of all possible stratifications (referred to as the universe of stratifications) based on atomic strata which is the most detailed stratification derived from the Cartesian product of all auxiliary variables as the solution space.
- The objective function minimizes the total sampling cost. Cost of sampling per unit can be set according to the effort associated with collecting and processing each unit.
- Barcaroli (2014) implemented this method in and R package titled "SamplingStrata".

Barcaroli, G. (2014a). SamplingStrata: An R Package for the Optimization of Stratified Sampling. Journal of Statistical Software, 61(1), 1–24. https://doi.org/10.18637/jss.v061.i04

Genetic Algorithm Procedure

- In the first step, the input data is read and loaded in “CFSFrameData” matrix.

```
CFSFrameData <- read.csv(file="./100K_Frame.csv", header=TRUE, sep=",")
```

- Then, the frame based on the loaded data is created as follows.

```
CFSFrame <- buildFrameDF(df = CFSFrameData,
                         id = "estno",
                         X = c("state","county","naics"),
                         Y = c("value"),
                         domainvalue = "naics")
```

- Atomic strata which is the most detailed strata resulting from the cartesian product of all auxiliary variables is then constructed and stored in “AtomicStrata” matrix.

```
AtomicStrata <- buildStrataDF(CFSFrame, progress = TRUE)
```

Atomic strata for the case study

STRATO	N	M1	S1	COST	CENS	DOM1	X1	X2	X3	X4
1*101*311*2	2	359.1800	16.578000	1	0	311	1	101	311	2
1*103*311*2	2	430.8695	21.346500	1	0	311	1	103	311	2
1*105*311*1	1	93.3954	0.000000	1	0	311	1	105	311	1
1*109*311*1	1	155.9170	0.000000	1	0	311	1	109	311	1
1*117*311*1	1	128.0330	0.000000	1	0	311	1	117	311	1
1*125*311*2	1	195.8990	0.000000	1	0	311	1	125	311	2
1*15*311*1	1	120.3710	0.000000	1	0	311	1	15	311	1
1*3*311*3	3	536.8480	33.971522	1	0	311	1	3	311	3

- (X1, X2, X3, and X4): auxiliary variables
- N: the number of units (establishments) in each stratum,
- M1 and S1: mean and standard deviation of the value for each stratum,
- Cost: the assigned sampling costs
- CENS: certainty
- DOM1: domain variable (industry)

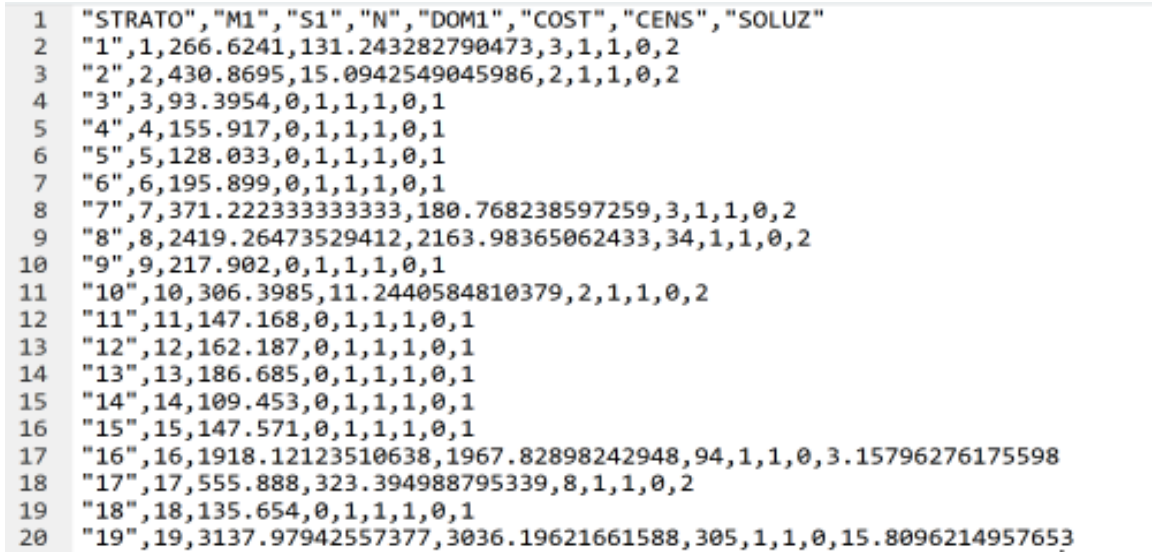
Genetic Algorithm Procedure cont.

- Next, CV constraints for each domain is imported.

```
CVConst <- read.csv("./CV.csv", header=TRUE, sep=",")
```

- The last step is to call “optimizeStrata” function which performs the optimal stratification and allocation based on Genetic Algorithm.

```
solution <- optimizeStrata(errors = CVConst,
                           strata = AtomicStrata,
                           parallel = TRUE,
                           iter = 100,
                           writeFiles = FALSE,
                           showPlot = FALSE)
```



```
"STRATO","M1","S1","N","DOM1","COST","CENS","SOLUZ"
"1",1,266.6241,131.243282790473,3,1,1,0,2
"2",2,430.8695,15.0942549045986,2,1,1,0,2
"3",3,93.3954,0,1,1,1,0,1
"4",4,155.917,0,1,1,1,0,1
"5",5,128.033,0,1,1,1,0,1
"6",6,195.899,0,1,1,1,0,1
"7",7,371.222333333333,180.768238597259,3,1,1,0,2
"8",8,2419.26473529412,2163.98365062433,34,1,1,0,2
"9",9,217.902,0,1,1,1,0,1
"10",10,306.3985,11.2440584810379,2,1,1,0,2
"11",11,147.168,0,1,1,1,0,1
"12",12,162.187,0,1,1,1,0,1
"13",13,186.685,0,1,1,1,0,1
"14",14,109.453,0,1,1,1,0,1
"15",15,147.571,0,1,1,1,0,1
"16",16,1918.12123510638,1967.82898242948,94,1,1,0,3.15796276175598
"17",17,555.888,323.394988795339,8,1,1,0,2
"18",18,135.654,0,1,1,1,0,1
"19",19,3137.97942557377,3036.19621661588,305,1,1,0,15.8096214957653
```

Future Work

- Applying Simulated Annealing (SA) Method
- Combining the GA and SA method and create a New Method
- Comparing New Methods with the Current CFS Sampling Design and Finding the Difference

- **Future Update:**
  - The Sixth International Conference on Establishment Statistics (ICES VI) will be held in New Orleans, Louisiana, USA, June 15–18, 2020.

Questions and Comments

M.Hashemipour@dot.gov

Julie.Parker@dot.gov

Special thanks to:

Saeed Ghanbartehrani

Sara Akbar Ghanadian