**Transforming Freight Flow Data Collection**

Milestone #3

Optimization Model Design and Preliminary Designs

Saeed Ghanbartehrani, Sara Akbar Ghanadian

Ohio University

Athens, OH

Rev. 0, Apr 2019

Contents

[1. Introduction 3](#_Toc6986070)

[2. Background 3](#_Toc6986071)

[3. Case study 4](#_Toc6986072)

[3.1. Data sources 4](#_Toc6986073)

[3.2. Pre-processing of the data 5](#_Toc6986074)

[3.3. Frame data generation 5](#_Toc6986075)

[3.4. Optimal stratification and allocation 6](#_Toc6986076)

[4. Results 10](#_Toc6986077)

[5. Appendix 10](#_Toc6986078)

[6. References 10](#_Toc6986079)

# Introduction

In this milestone, a case study is presented to demonstrate the optimal stratification and allocation method based on Genetic Algorithm (GA) in a CFS like scenario. A high level discussion of the optimal stratification and allocation method is discussed in the background section. Data sources, pre-processing, frame data, and the optimal solution are discussed in the case study section. Results are discussed in the results section.

# Background

The optimal stratification and allocation method proposed by Ballin and Barcaroli (2013) 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. Since the set of all possible stratifications based on the atomic strata is quite large even for cases with a few auxiliary variables (e.g. 4 auxiliary variables each with 3 levels, result in an atomic strata of size 12 with 4,213,597 possible stratifications), full enumeration of the solution space is not possible in reasonable time. To address that, Ballin and Barcaroli (2013) used Genetic Algorithm (GA) which is a heuristic search technique inspired by evolutionary biology. Therefore, this method starts with with an initial set of potential solutions and evolve them using inheritance, mutation, selection, and crossover operators at each iteration to improve the solution in future iterations and finally reach a good solution while there is no guarantee to find the optimal solution. This means that only a fraction of possible stratifications is explored in the process. For each stratification, the optimal allocation is determined by Bethel's (1989) multivariate method. The auxiliary variables need to be categorical. Continuous variables are therefore converted to categorical ones using the k-means clustering method proposed by Hartigan and Wong (1979).

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. For simplicity, relative sampling costs (i.e. cost of 2 for units requiring twice as much effort compared to the regular units with cost of 1) can be used in the model. If all sampling costs are set to 1, the model minimizes the total sample size.

Barcaroli (2014) implemented their proposed method in and R package titled “SamplingStrata”. The R package “SamplingStrata” (Barcaroli, 2014b) is available on the Comprehensive R Archive Network (CRAN) (“The Comprehensive R Archive Network,” 2019).

# Case study

The optimal stratification and allocation method based on Genetic algorithm is evaluated on a case study involving a sampling frame with 100,000 units. The units in the case study are designed to replicate the establishments in CFS.

## Data sources

The state level freight transportation value and weight were used from FAF 2016 (Bureau of Transportation Statistics, 2019b) estimates Access database. The selected FAF dataset with over 1.5 million records is about 50 MBs in size in CSV format. The complete county file for 2016 County Business Patterns (CBP) (US Census Bureau, 2018) was used to estimate average freight value and weights for each industry at the county level. more details on the steps involved in the data processing are presented in the next section. The CBP dataset with over 2 million records is about 12 MBs in CSV format. A mapping between NAICS and SCTG codes was created based on “NAICS Industries In-scope to the 2017 CFS” list from 2017 CFS methodology (Bureau of Transportation Statistics, 2019a). The raw data sources are available in CSV format in “Raw\_Data” folder on the GitHub repository (Ghanbartehrani, 2019).

## Pre-processing of the data

All data files were imported in a PostgreSQL (The PostgreSQL Global Development Group, 2019b) relational database to facilitate the pre-processing stage. Total value and weight for each SCTG code were aggregated at the state level in the FAF table. In CBP table, total number of establishments was calculated for each state, county, and NAICS category based on the list of industries in-scope to the 2017 CFS. Then, CBP and FAF tables were joined based on the NAICS/SCTG mapping mentioned in the previous section to add total number of establishments in each industry-state combination in the FAF table. Finally, county level value and weights for each industry were estimated by multiplying state level numbers by the ratio of the number of establishments in each county (and industry) over the total number of establishments in each state (and industry). All the SQL scripts used to perform the steps involved in the pre-processing stage are available in “SQL\_Scripts.sql” file available in “SQL” folder on the GitHub repository (Ghanbartehrani, 2019).

## Frame data generation

A function developed in PostgreSQL procedural language (The PostgreSQL Global Development Group, 2019a) was developed for generating sampling frames with user defined size based on the pre-processed data described in the previous section. The function signature is as follows.

generate\_est(frame\_size, source\_table, value\_CV, wgt\_CV, mile\_CV)

“Frame\_size” is the desired number of units in the frame, “source\_table” is the name of the table in which the pre\_processed data is stored, while “value\_CV”, “wgt\_CV”, and “mile\_CV” parameters are the desired Coefficient of Variations for generated values, weights, and mileages for each establishment. The following is an example call to the function to generate a frame with 100,000 establishments based on the data stored in “fafcbp” table with 0.1 CVs for value, weight, and mileage.

SELECT \* FROM generate\_est(100000, 'fafcbp', 0.1, 0.1, 0.1);

The function distributes the number of units proportional to number of establishments in each county, state, and industry combination. Value, weight, and mileage values for each establishment are generated from the normal distribution using the estimated average values stored in the input tables and standard deviations calculated based on the user provided CVs. The actual size of the generated frame is typically less than the user provided number due to rounding errors. The function does not generate an establishment for a state-county-industry combination if the number of allocated units is less than one (after rounding to the closest integer). The example code provided above resulted in a frame with 98,388 establishments which is available in “100K\_Frame.csv” file available in “R\_Scripts” folder on the GitHub repository (Ghanbartehrani, 2019).

The source code for the function is available in “Generate\_est.sql” file available in “SQL” folder on the GitHub repository (Ghanbartehrani, 2019).

## Optimal stratification and allocation

“SamplingStrata” package needs to be installed prior to running which requires R (The R Foundation, 2019) version 2.15 or newer. The package can be installed by clicking on “Install Packages” from “Tools” menu in R Studio (R Consortium, 2014) and Typing “SamplingStrata” in the search box. “Repository (CRAN)” needs to be chosen as source. The source code, documentation, and samples are available on the package’s GitHub repository (Barcaroli, 2019).

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")

df is the matrix where the input data is stored, id is the column used to uniquely identify the units, auxiliary variables are listed in X, and Y is the list of target variables. The column corresponding with the domain variable is specified in “domainvalue “. State, county, and naics are categorical used as auxiliary variables. Value is converted to 15 categories and used as the forth auxiliary variable in the frame in the next step.

CFSFrame$X4 <- var.bin(CFSFrameData$value, bins=15)

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)

The size of the atomic strata in this example is 26,436 and the first few rows are displayed below.

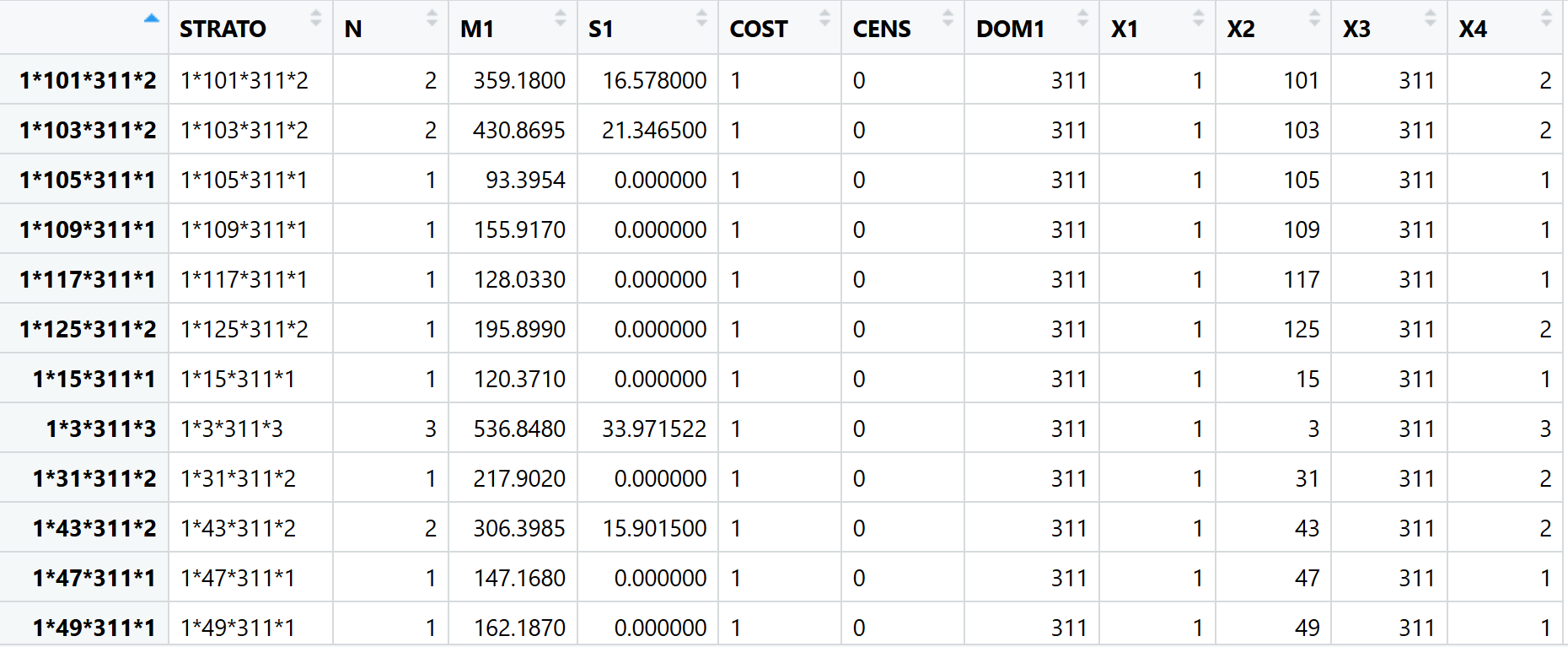


Figure 1. Atomic strata for the case study

The first column shows the combination of the values for the auxiliary variables (X1, X2, X3, and X4) identifying each stratum, N is the number of units in each stratum, M1 and S1 are mean and standard deviation, cost is the assigned sampling costs (all one), CENS column allows defining take-all (certainty) strata (strata from which all units must be included in the sample) when set to one. In this case study, no take-all strata are defined. X1, …,X4 columns are the values of the four auxiliary variables state, county, NAICS, and value class.

Next, “CV.csv” which contains the CV constraints for each domain (industry identified by NAICS in this case) is imported.

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

Following shows selected rows from “CV.csv”. Each row in the file corresponds with a CV constraint that corresponds with each value of the domain variable. In this case, a CV constraint of 0.05 is defined for each NAICS code (presented in “domainvalue” column). Note that the values in “domainvalue” column in “CV.csv” file and the variable assigned to “domanvalue” in “buildFrameDF” function discussed earlier need to be consistent.

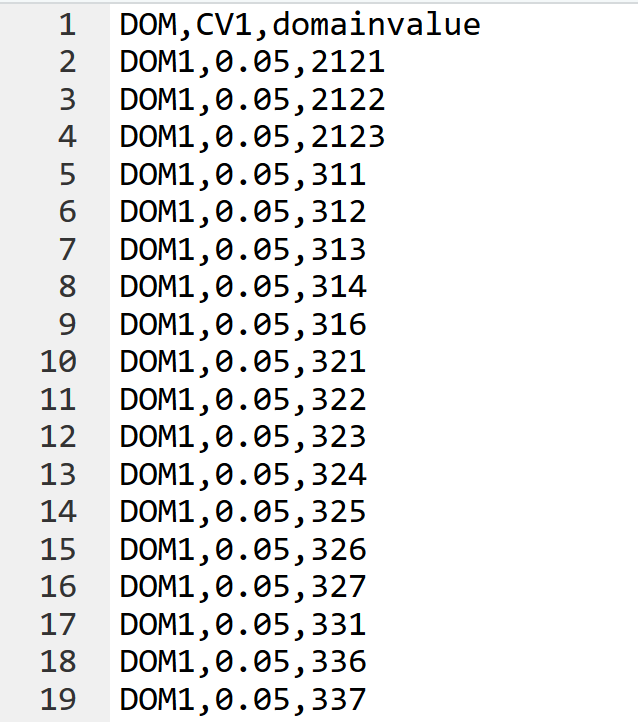


Figure 2. Selected rows from CV.csv

Then, all input data needs to be checked to ensure consistency using “checkInput” function.

checkInput(errors = CVConst,

strata = AtomicStrata,

sampframe = CFSFrame)

The last step is to call “optimizeStrata” function which performs the optimal stratification and allocation based on Genetic Algorithm. In this case, a few parameters such as parallel processing, number of iterations (i.e. generations in the genetic algorithm), output files and plots are specified along with the two required input matrices which are errors (CV constraints) and atomic strata.

solution <- optimizeStrata(errors = CVConst,

strata = AtomicStrata,

parallel = TRUE,

iter = 100,

writeFiles = FALSE,

showPlot = FALSE)

Although the entire results are stored in “solution”, selected elements can be stored in separate csv files for convenience and further analysis. The two major outputs are “aggr\_strata” and “indices” columns.

write.table(solution$aggr\_strata,file="./aggr\_strata.csv", sep=",")

write.table(solution$indices,file="./indices.csv", sep=",")

The stratification is presented in “indices” while “aggr\_strata” shows the number of samples allocated to each stratum along with some other details.

# Results

A detailed discussion of the results along with tables and visualizations will be added in this section.

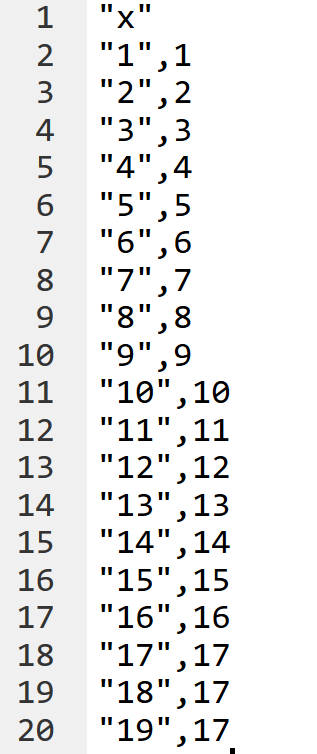


Figure 3. Selected rows from “indices.csv” file

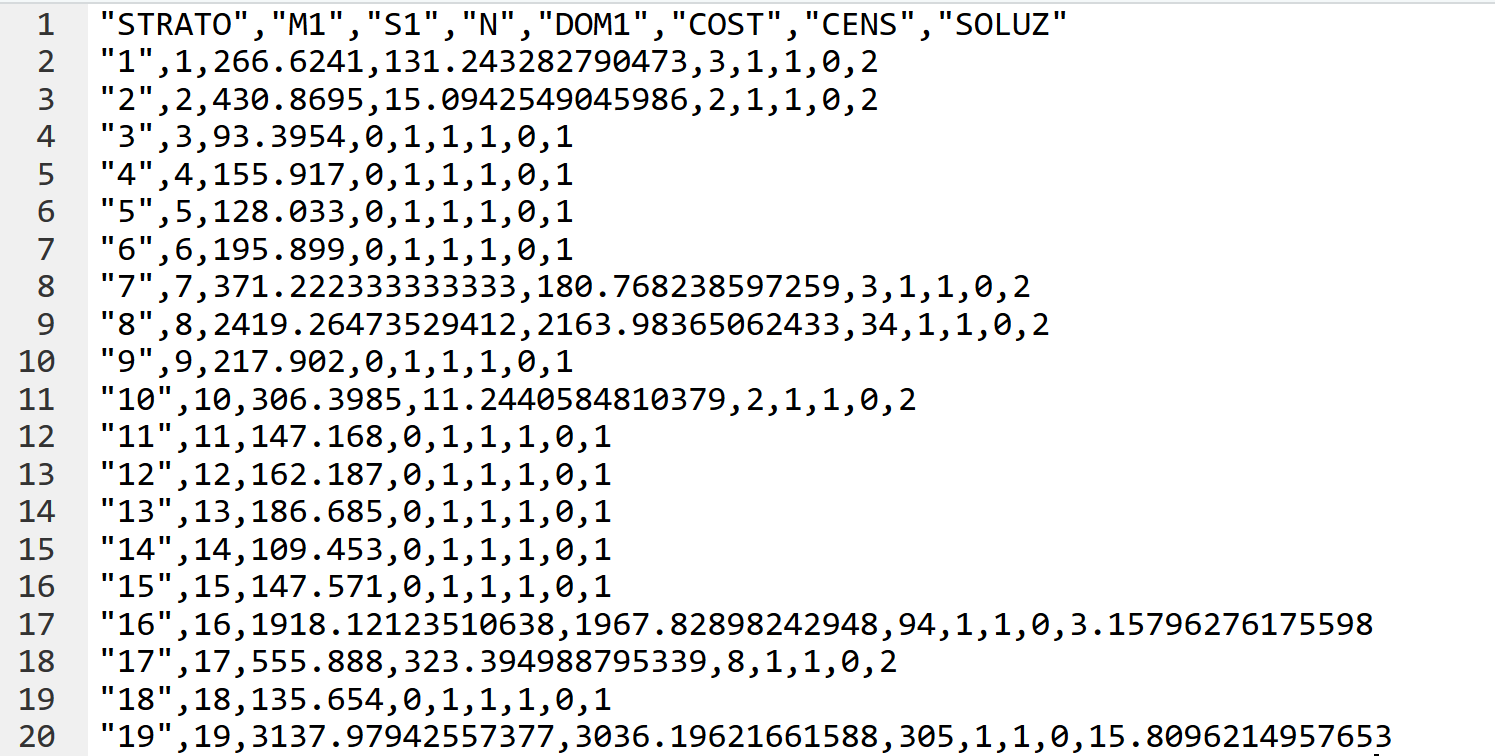


Figure 4. Selected rows from “aggr\_strata.csv” file

# Appendix

The complete R script for the case study is presented below. The latest code is also available in “cfs1.R” file available in “R\_Scripts” folder on the GitHub repository (Ghanbartehrani, 2019).

library(SamplingStrata)

# Read the input frame data from 100K\_Frame.csv file

CFSFrameData <- read.csv(file="./100K\_Frame.csv", header=TRUE, sep=",")

# Build the frame using estno column as identifier, state, county, and naics

# as auxiliary variables and value as target variable

CFSFrame <- buildFrameDF(df = CFSFrameData,

id = "estno",

X = c("state","county","naics"),

Y = c("value"),

domainvalue = "naics")

# Converting value to 15 categories and using it as the fourth auxiliary variable in the frame

CFSFrame$X4 <- var.bin(CFSFrameData$value, bins=15)

# Building the atomic strat based on the frame

AtomicStrata <- buildStrataDF(CFSFrame, progress = TRUE)

# Uncomment and run the following line to view the atomic strata

#str(AtomicStrata)

# Read the CV constraints from CV.csv file

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

# Check the input data for errors

checkInput(errors = CVConst,

strata = AtomicStrata,

sampframe = CFSFrame)

# Optimization of stratification

solution <- optimizeStrata(errors = CVConst,

strata = AtomicStrata,

parallel = TRUE,

iter = 100,

writeFiles = FALSE,

showPlot = FALSE)

# Writing the stratification and allocation results to csv files

write.table(solution$aggr\_strata,file="./aggr\_strata.csv", sep=",")

write.table(solution$indices,file="./indices.csv", sep=",")

# References

Ballin, M., & Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 27.

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

Barcaroli, G. (2014b). **SamplingStrata** : An *R* Package for the Optimization of Stratified Sampling. *Journal of Statistical Software*, *61*(4). https://doi.org/10.18637/jss.v061.i04

Barcaroli, G. (2019, April 21). R package for Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys: barcaroli/SamplingStrata. Retrieved April 23, 2019, from https://github.com/barcaroli/SamplingStrata

Bethel, J. (1989). *Sample allocation in multivariate surveys*. *15*, 47–57.

Bureau of Transportation Statistics. (2019a, February 25). 2017 Commodity Flow Survey Overview and Methodology. Retrieved April 23, 2019, from https://www.bts.gov/archive/publications/commodity\_flow\_survey/methodology\_2012

Bureau of Transportation Statistics. (2019b, March 29). Freight Analysis Framework. Retrieved April 23, 2019, from https://www.bts.gov/faf

Ghanbartehrani, S. (2019, April 10). saeedt/CFS\_Sampling. Retrieved April 23, 2019, from GitHub website: https://github.com/saeedt/CFS\_Sampling

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

R Consortium. (2014, March 27). RStudio. Retrieved April 23, 2019, from RStudio website: https://www.rstudio.com/

The Comprehensive R Archive Network. (2019, March 11). Retrieved April 18, 2019, from https://cran.r-project.org/

The PostgreSQL Global Development Group. (2019a). PostgreSQL: Documentation: 11: 43.1. Overview. Retrieved April 23, 2019, from https://www.postgresql.org/docs/current/plpgsql-overview.html

The PostgreSQL Global Development Group. (2019b). PostgreSQL: The world’s most advanced open source database. Retrieved April 23, 2019, from https://www.postgresql.org/

The R Foundation. (2019). R: The R Project for Statistical Computing. Retrieved April 23, 2019, from https://www.r-project.org/

US Census Bureau. (2018, September 12). County Business Patterns: 2016. Retrieved April 23, 2019, from https://www.census.gov/data/datasets/2016/econ/cbp/2016-cbp.html