Dubrovnik, Croatia
June 16-19 2009

Creation of a database

Sampling unit, sample size, stratification criteria and imputation methods

Creating Databases

- Practical approach on sources of micro-information
  1. Limited information on ***ALL*** corporations
  2. Extended and validated information on sampled corporations
- First file can be used for auditing, administration and as a base for the sampling process
- Second file (sample) can be used for tax modeling

Collecting data on the population

- Main file purpose
  - Auditing
  - Evaluation of Policy
  - Revenue Statistics
  - Base for sampling
- This file can be used by auditors and is dynamic (updated with results of audits and carry-back)

Collecting data on the population

- Important considerations about the type of data to be collected:
  - Tax revenue, tax base and major deductions/credits
  - Unique Corporation identifier to allow for longitudinal analysis
  - Homogeneity of the data

Collecting data on the population

- Breakdown variables should:

1. Include many measures that target specific types of corporations (e.g. financial institutions, small business)
    - Industrial breakdown (NAICS – NACE)
    - Size variables (Assets, gross revenues, salaries, etc)
    - Geographic sub sections
    - National status (domestic company?)
2. Allow for creation of homogeneous strata for sampling

Why a sample?

- A three line lesson in the economics of creating estimates
  - Experts (intuitive) knowledge
  - Theory (ITA)
  - Empirical (Economic and Tax Data)
- Empirical studies ask for detailed information, thus the need for a sample

Why a sample?

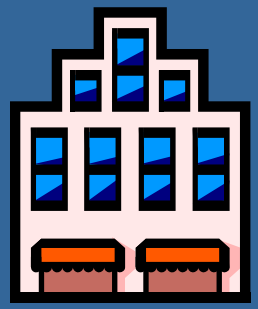
- The “economics” of a sample
  - Optimize quality information under fiscal constraints
- Provides detailed validated information
- Focus on important players and variables
- Allow for micro simulation / analysis

A Sample Example

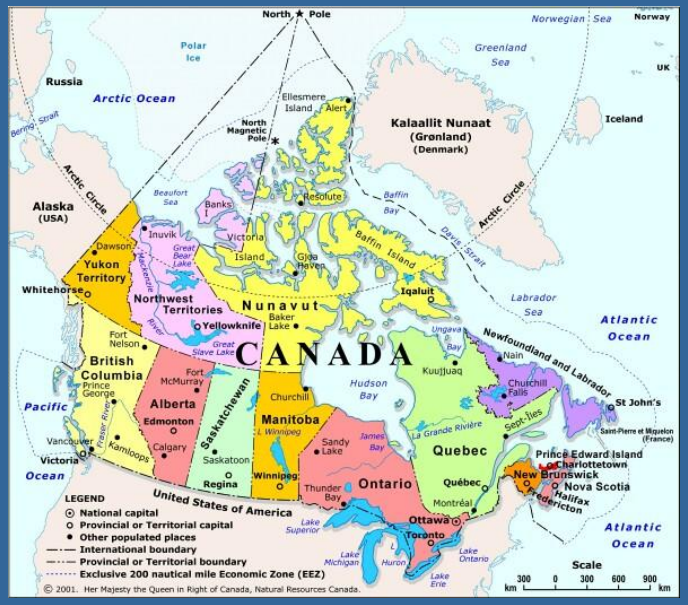
- Canadian Corporate Sample File (CSF)
  - Contains around 1,900 fields on 45,000 corporations (out of a population of around 1M)
  - All major fields from the main form (called T2)
  - Over 600 fields from accompanying schedules
  - Selected information from balance sheet and income statement

Creating a sample – How it is built

- The best sample is the population, and even then, it wont be perfect
- If we had a sample of one:



X 1,200,000 =



Representative ???

Creating a sample - Stratification

- That is why we split the population in sub groups that exhibit the same profile and have a similar economic cycle.
- For example, we could group together corporations of similar size (using asset value)
- A random selection would then be done in each of these sub groups and would provide a better representation of the overall population

Live exercise on stratification

- Using population in the Corp_Size.xls file
- Population of 100 corporations
  - Sampling first 10 corporations (weight of 10)
  - 73% over shooting on tax variables
- Creation of size stratification
  - Sampling of 5 small, 3 medium, 2 large
  - What is the impact?

Creating the Target population

- Population file has information on all corporations
- Create the target population (referred to as the Frame) by removing:
  1. Inactive Corporations
  2. Corporations with less than 60 days of activity in the fiscal year
  3. Very small corporations

From the Frame to the Stratum

- The Frame is then stratified to reflect the economic activity
- For example, stratification criteria could include:
  - National status (x2)
  - Taxation status (x2)
  - Level of asset (size) (x4)
  - Industry sub-grouping (x25)
  - Geographic region (x7)
  - Outliers (3 stratum)

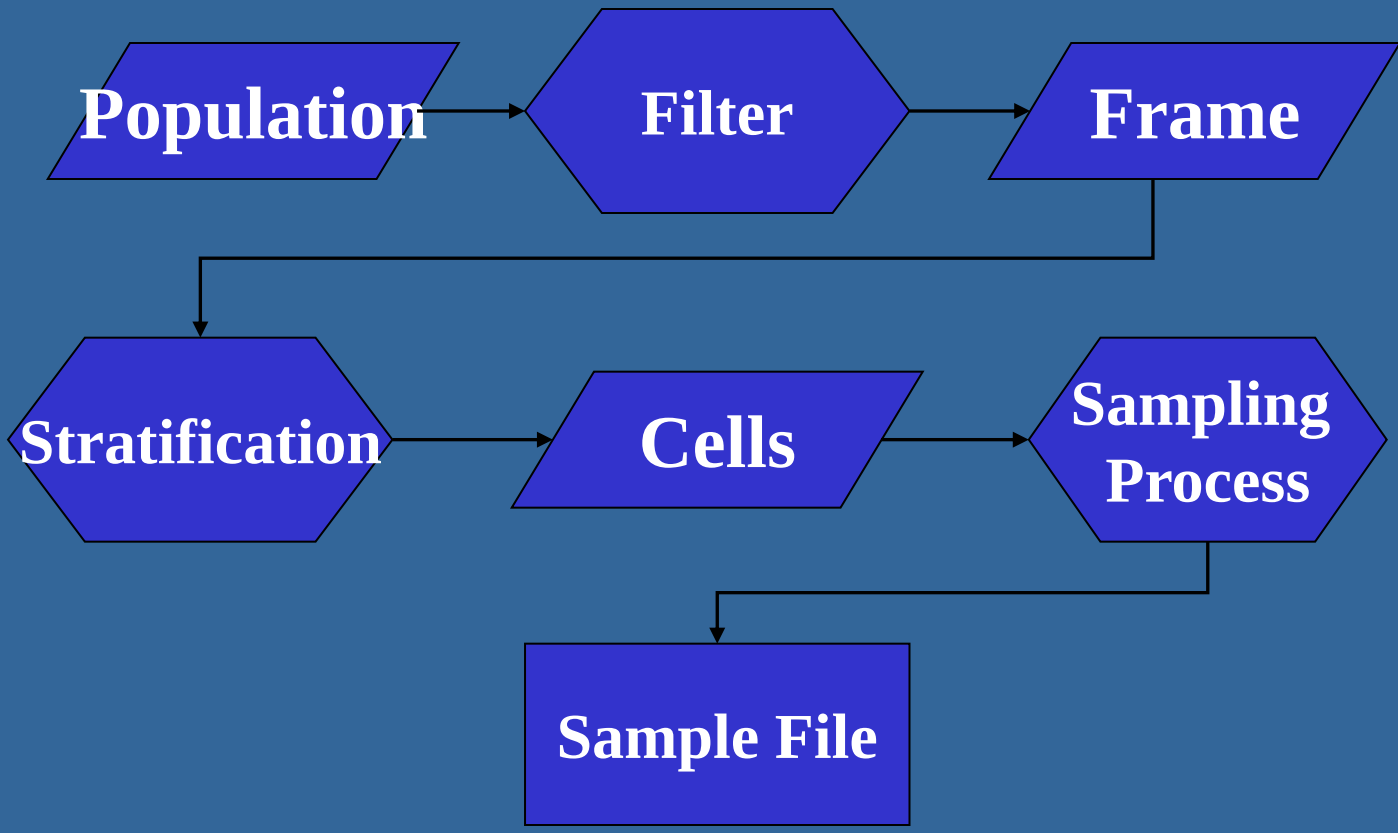
From the strata to the sample

- This results in 2,803 strata
- Once those stratification criteria are defined, a random sample is done in each resulting stratum (note that some stratum may not contain any corporations)
- Special rules can apply
  - e.g. compulsory selection of every corporations with a Asset of level 4 (Large)
  - maintaining a Hit-list of corporations that are added automatically to the sample.

From the Strata to the Sample

- The Revenue Agency keys manually the required information from all sampled corporations.
  - For 2003, the Canadian Corporate Sample File was composed of approximately 45,000 corporations representing a total population of around 1 million.
- Evolution of the population and the sample over time:
  - Put mechanisms in place so that the size of the sample increases with the population
  - Periodical review of methodology and data collected
  - Should allow for time series analysis

The path


Population
Filter
Frame
Stratification
Cells
Sampling Process
Sample File

Weighting / Estimation

- The sample will be weighted by stratum so that it reflects the overall economy
- Based on the fact that stratum should be homogeneous

- Weight = $\frac{\text{\# observations in the “Frame”}}{\text{\# observations in the sample}}$

- The weighted data creates a sample that represent the overall national economy

Size of a sample and descriptive statistics

- What size is appropriate ?
- Size depend on the homogeneity of the population.
- Will look for a maximum variation coefficient of 5% or 10% for pre specified key variables in each stratum
- A good stratification with homogeneous population in each stratum will allow for smaller sample

Creating a Sample – Short falls

- It is a sample file and, as such, does not have micro information on all the population
- It cannot be used for robust analysis in a sub section of the economy that is not a stratification criteria (e.g. analysing Capital Taxes paid by corporations doing R&D)
- Longitudinal analysis can be difficult, especially when dealing with pools.

Creating a Sample – Strong points

- Gives detailed information
- Format is manageable for micro simulation (i.e. I/O and CPU wise)
- Unlike the universe file, there is a thorough validation of all fields.
- For corporations, the Hit List ensures access to micro information on major corporations

Creating a Sample

- Population Database: How to use it to create a sample (Post vs Dynamic sampling)
- Post sampling
  - Provide exact size for sample
  - Allow validation tests before gathering the sample information
  - Delay the capture of information and completion of the sample
  - Increase Administration costs

Creating a Sample

- Dynamic Sampling
  - Allow to capture information as it comes in
    - (Reduce time and costs – keying information once)
  - Stratification tests has to be done on prior years (wont be as accurate)
  - Size of sample will vary (within certain limits)
  - Need post sampling for missing strata

After thoughts

- Need for a robust unique ID over time (CIT)
  - Time series and carry over (back) of pools
- Management of variable names over time
  - Why, especially for micro simulation models, alphanumeric is better than numeric?
- New measures, new variables
- Formal review of sampling methodology