Data Mining, Data Integrity -- Florida Experience

By **Yen Chen**
Florida Department of Revenue

History

- **Florida Department of Revenue has been migrating from the legacy system of tax returns and data to an integrated tax system – SAP since 2000.**
- **Florida Corporate Income tax was first migrated (2000)**
- **Sales and Use tax was migrated in 2003**
- **Fuel taxes and Documentary stamps tax were migrated in 2005**

New System -- SAP

- SAP is a very powerful and complex system.
- Widely used in private sectors, the majority of the top 500 companies in the states.
- About 10% of the public sector has implemented SAP.

New System - SAP

- Currently the online transaction (operational) system is SAP R/3.
- SAP/BW is the SAP business information warehouse which stores the historical data from SAP R/3 through ETL (extraction, transformation and loading).

New System -- SAP

- Myth of any new system or technology

  "The problem will be solved once we move to SAP".
- The system will work the way you want it to work – business rules are very crucial during the implementation period.

New System -- SAP

- Myth of the 'cleanness' of data in SAP/BW since the process of moving from SAP R/3 to SAP/BW involves normalization, cleansing & etc.
- Normalization and cleansing are technical terms in the ETL process. They do not clean the data. Business rules should catch and correct the errors in data.

Things to learn from Florida Experience

- New system, new experience, lots to learn and adapt.

  SAP is our new integrated tax system. With the new system, we need to adapt to new terminology

  Taxpayer – Business Partner

  Tax Returns – Sales Orders

  Business Partner number (integrated ID)

  Contract Object number (unique tax source number)

  Sales Order number

  etc.

Things to learn from Florida Experience

- Revenue collecting vs. Research

  Research unit needs to be involved in the migration process and have its voice heard

- The integrated system tends to focus on the revenue collected; this is especially true when the economy is good, and ignore other fields that are crucial for research units

Things to learn from Florida Experience

- Corporate Income tax: NAICS codes are crucial for research units to distinguish the type of businesses, yet were not incorporated in the migration from the legacy system to the integrated system. Nor were many other details on the CIT returns.
- Sales and use tax: many of our forecasts and estimates are based on taxable sales, yet there are errors. Some taxable sales were left blank and a computerized math audit failed to build a routine to capture them.

Things to learn from Florida Experience

- Florida has 67 counties and our integrated system has two sets of county codes co-existing in the system, (01-67) and (11-77).
- For consolidated filers: how to split into each county's shares, e.g. when more than one minor tax is on a return, county-by county distributions aren't always easily attainable.

Things to learn from Florida Experience

- Questioning the data and checking on the integrity of the data – always
- Training on the new system. It is very important to know the system and how the data are retrieved from the system for you.
- Access to the new system is also challenging.

Things to learn from Florida Experience

- The Pros and Cons of running the old and new systems in parallel.
  - checking on the differences, e.g. why revenue from certain businesses tends to be higher or lower on one system than the other.
  - drain on the IT resource – having to maintain two systems at the same time.

Matching data – one example

- Quite often, researchers need to use data from different sources in their analyses.
- Data may be on different platforms, in different formats and time frames, e.g. annual, quarterly, monthly or weekly; annual data could be calendar year, state fiscal year or local government fiscal year;
- Your data may be pulled by applied date or validation date; etc

Matching data – one example

- In 2006 legislature session, Florida Department of Revenue was asked to assist in a property tax study.
- In 1992, a constitutional amendment (Save Our Home Amendment) was passed that allows Florida homestead owners to have the assessed value of their properties capped at the lesser of CPI growth or 3% annually. When homestead owners sell their homestead property and buy another homestead property in Florida, their assessed value of the new homestead property will be adjusted back to the full value of the property.

Matching Data – one example

- Study addresses demographic and economic characteristics of the Florida homestead households:
  - the length of stay in the homestead house before move;
  - Full property and assessed value by age group of the homestead owners;
  - Full property and assessed value by income levels;
  - for movers, do they move within the county where they stay or another county in Florida or out of state;
  - do they move up (bigger house) or move down;
  - Difference between taxable and full value of the property; etc.

Matching data – one example

- Florida allows a $25,000 homestead exemption on residential property that is lived in by the property owner. The data contains one or two social security numbers depending household status.
- IRS data are matched to obtain age and income information for the study.

Match IRS and PTA data

- Before matching, try to get each dataset as clean as possible.
- IRS data: some duplicate SSNs due to

  1. filing for multiple years
  2. some amended returns (some new information, e.g. date of birth, name change, etc)
  3. marriage or divorce or death during the year
  4. errors in Social Security numbers

Matching IRS and PTA data

- Both IRS and PTA data have one or two social security numbers depending on the household status.

  Sometimes both social security numbers are identical or sometimes as in PTA data, SSN1 is blank while SSN2 contains a valid social security number.

Matching IRS data and PTA data

- IRS data preparation: checking duplicates on fields like tax year; the first SSN and the second SSN, both name lines if filed jointly or married yet file separately; adjusted gross income; filing status; addresses; date of birth.
- Many married yet filing separately have crossed SSNs in the data, we need to match SSN1 with SSN2 to find those records and add the adjusted gross income together to derive the household income.

Matching IRS data and PTA data

- IRS data preparation: some tricks on addresses and names, addresses may be spelled differently ('street' vs. 'st'; '21st' vs. '21'), names may be spelled differently, English spelling vs. Spanish spelling.
- Due to discrepancies in SSNs, we may have to keep the duplicate records.

Matching IRS and PTA data

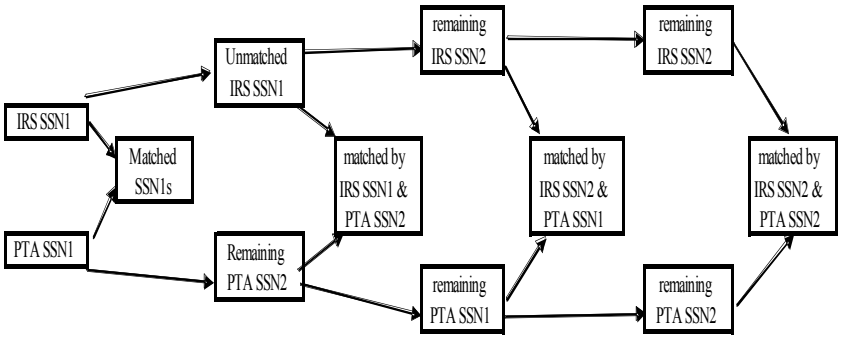
- PTA data: property tax data has a unique parcel ID, yet there are duplicate social security numbers too due to:
  1. marriage or divorce during the year
  2. one parent and college student
  3. discrepancies in social security numbers
- Property tax data are from 67 counties in Florida, their parcel IDs may vary from year to year. The way of reporting also varies from county to county.

Match IRS data with PTA data

- There are different ways of matching

  1. matching by elimination: first match SSN1s of the two datasets, those matched, put aside; then matching SSN1 in one remaining file against the SSN2 in the other remaining file and put the matched on the side; then matching SSN2 in the second remaining file against the SSN1 in the other second remaining file and put the matched on the side; last matching SSN2s in the third round remaining files. The drawback of this method is you might lose some records if SSN2 in the first round matched are unique and have a match in SSN1 in the second file.

Matching IRS and PTA data



Matching IRS data and PTA data

- There are different ways of matching

  2. matching all possible ways and then eliminate duplication: create a larger file for IRS data, create SSN=SSN1 and SSN=SSN2 and appending the two files; do the same thing for PTA file; sort the data by SSN and matching the two, then delete the duplicate using the unique PTA parcel number and SSNs.

Match IRS data and PTA data

- There are different ways of matching

  2. – continued

  since we kept names from both IRS and PTA files, we may now delete those records with error SSNs in the matched file.

  We also linked returns of couples who filed separately with IRS yet we failed to put them together when preparing IRS data for matching.

Matching IRS and PTA Data

- To make the process easy, two social security numbers in each dataset were rearranged, the larger one was assigned TSSN1, the smaller one (could be blank) was assigned TSSN2.

Matching IRS data and PTA data

- From 4.1 million PTA homestead records (2005) with SSNs, 3.3 million records were matched with IRS data, about 82%.
- Explanation for non-match:

  1. filed income tax return in other states
  2. low income households and are not required to file income tax return.
  3. errors in SSNs.

Matching PTA data

- One question that the study asks is how many homestead houses moved from one year to next? How many moved within the state and where?
- We used the similar methodology above to match two years' PTA data to find the answer. A unique master parcel ID field is created for all parcels cross all years involved in the study.

Matching PTA Data

- The project is still in progress and the preliminary result is due in November 2006. We have to do lots of data mining and maintain the integrity of the data.
- The result? – To be continued at the next FTA Conference**.**