**Data Science Project**

**Nick Marchio**

**Fall 2014**

**Estimating Metro Exports with Constraint Satisfaction Algorithms**

**Background Problem**

My project is an offshoot of an existing research project pertaining to metropolitan exports. The research aims to generate export estimates for over 130 goods and services industries over time for each of the 3,113 counties in the U.S. The resulting data is then provided to local economic development departments interested in devising plans to foster export-oriented growth.

Due to limitations that exist in federal data there are no available sources of subnational exports that reflect the actual location of production outflows.

Given this data deficit, traditional data science methods that rely on predictions based on samples or training data are not as helpful and we are forced to explore other techniques outside the traditional taxonomy of models.

|  |  |  |
| --- | --- | --- |
|  | **Continuous** | **Categorical** |
| **Supervised:** | Regression | classification |
| **Unsupervised:** | dimension reduction | clustering |

One available solution is called allocation, a method that involves distributing a known control total available at higher levels of aggregation down to more detailed sub-levels using shares calculated from some other data source that strongly correlates with the data we’re trying to estimate. In the context of exports, in the past I have allocated national exports to counties based on each counties share of national GDP for each industry. Mathematically, this process is as simple as multiplying a national export total against each county’s share of the total GDP for a given industry.

While this method clearly solves the problem of finding ‘second-best’ estimates when no data is available, its main flaw is that it assumes that the distribution of some outside measure is perfectly correlated with the measure we are trying to estimate. Clearly, exports do not have a one-to-one relationship with output since some metros might produce products that better align with international demand than other places, i.e. Witchita and Denver may produce the same amount of aircraft parts, but differ in their actual export propensity (Denver may primarily produce military aircraft sold domestically, Witchita might more produce aircraft parts sold internationally).

**My Project**

The innovation I performed for my project is that I introduced several new data sets that better approximate actual export performance than GDP.

These data sets include (see spatial allocation section):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Function** | **Data Series** | **Year** | **Source** | | |
| Goods Exports | NAICS 4 Domestic Goods Exports by U.S. | 2003-2013 | *Census USA Trade Online Goods Exports* | | |
| Service Exports | Services Exports by U.S. | 2003-2013 | *BEA Table 3.1 Service Exports* | | |
| Detailed Services Exports by U.S. | 2003-2012\* | *BEA Table 1a Detailed Other Business Professional Technical* | | |
| Detailed Travel & Tourism Spending by U.S. | 2003-2013 | *BEA Travel Tourism Satellite Industry Accounts* | | |
| Detailed Royalties Industry Receipts by U.S. | 2003-2011\* | *IRS Royalties Table 7 Corporate Returns* | | |
| Inflation Adjustments | GDP Price Index by U.S. | 2003-2013 | *BEA Table 1.1.4 Price Indexes for Gross Domestic Product* | | |
| Goods Export Price Index by U.S. | 2003-2013 | *BLS Producer Price Index* | | |
| Agriculture & Services Export Price Index by U.S. | 2003-2013 | *BEA NIPA Table 4.2.4 Price Indexes for Exports and Imports of Goods and Services by Type of Product* | | |
| Spatial Allocation | NAICS 4 GDP by County | 2003-2013 | *Moody's Analytics* | | |
| NAICS 3241 Exports by Petrol. Admin. Defense Districts (PADDs) | 2003-2013 | *Energy Information Administration's Total Crude Oil and Petroleum Products Exports* | | |
| Intl. Origin, Destination, Passthrough Flights by Airport | 2003, 2010\* | *Sabre analysis of U.S. Department of Transportation’s Bureau of Transportation Statistics International Flight Data* | | |
| International Student Spending by Institution | 2003-2013 | *NAFSA analysis of IIE's Open Doors* | | |
| SCTG 2 Exports by MSA and non-MSA | 2010\* | *EDR analysis of U.S. Department of Transportation's Freight Analysis Framwork, IMPLAN, WISER Trade, Oak Ridge Intercounty Impedances* | | |
| Jobs Estimation | NAICS 3-4 Direct & Total Multipliers by U.S. | 2003-2012\* | *BLS Employment Requirements Matrix* | | |
| NAICS 3-4 Direct & Total Tourism-related Emp. by U.S. | 2003-2013 | *BEA Travel Tourism Satellite Industry Accounts* | | |
| *\*Some values were imputed or interpolated* | | |  |  |  |

The main challenge in introducing several different data sets is that we’ll need to utilize a constraint satisfaction algorithm to ensure the solutions satisfy the multiple control totals. Moreover, this is a fairly basic combinatorial optimization problem somewhat in the vein of a scaled-up version of Sudoku.

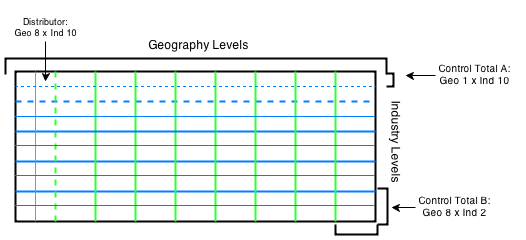
**Implementation**

Fortunately, there is an algorithm that can generate a solution that satisfies all constraints imposed by the control totals.

Its key assumption is that all controls are mutually consistent and are structured in an orderly hierarchical manner so that it’s possible to decompose the control totals into common subparts. In my data, I forced all the control totals to tie out with national statistics and built non-overlapping schemas so that it would be possible to cleanly move between levels of aggregation.

In the case of this method, the two control totals were the national detailed industry exports and metropolitan exports reported at a broader industry level and the distributors were detailed industry county GDP shares of each total. In this process, the control totals are iteratively allocated, first to the shares of one total, and then to the shares of the other. This process is repeated, recalculating shares along the way, until the estimates converge on a set of mutually consistent solutions, thus satisfying the constraints.

The following graphic demonstrates the general structure of the problem. For the sake of illustration, I chose are arbitrarily small matrix to summarize this set up. Control Total A imposes a hard constraint at a more detailed industry level, whereas Control Total B imposes a hard constraint at a more detailed geographic level and at broad industry groupings that align with the Control Total A detailed industries. Once the Control Totals are in place, a distributor is chosen that correlates with what we are trying to estimate and that is available at the lowest common multiple of the most detailed geographic and industry levels. Finally, we use the algorithmic process to fit an estimate within each ‘constraint box’ created by Control Total A and B.



Algebraically, the algorithm may be expressed as follows:

In terms of pseudocode, the algorithm may be implemented as follows:

For in***X****-***Data**iterate until converge:

*Sum-A* = = Aggregate by j groupings in

*Share-A* = = Calculate share of

= Allocate to *Share-A*

**=**

*Sum-B* = = Aggregate by j groupings in

*Share-B* = = Calculate share of

= Allocate to *Share-B*

**=**

It’s also possible to implement the method using the Python Constraint module.

**Applications**

The main takeaway is that this technique is a handy way to derive more detailed estimates from aggregate data as long as you have other measures that you think correlate with the data you’re trying to estimate. It is attractive because it easily allows you to fit your data within the framework of other more reliable control data sets.

Believe it or not, this algorithm is used by a wide array of data organizations. BEA uses it to estimate state level estimates of GDP using compensation data obtained from the IRS. Moody’s Analytics also uses the technique to allocate state level GDP to the county level using the QCEW compensation data from the BLS.

**Results Discussion:**

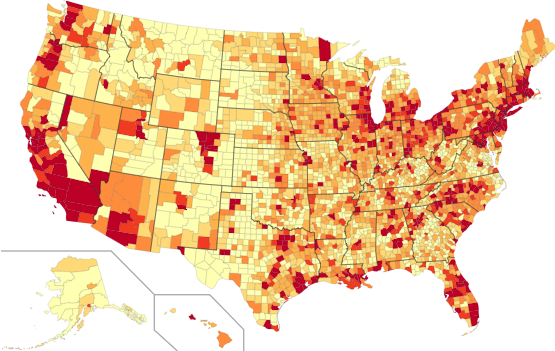
There were a number of new findings that resulted from adding the additional constraints:

-Oil exports migrated south: massive oil refineries in San Francisco and Minneapolis that supplied gas to their local market were downgraded as exporters after we added the constraint from the EIA’s petroleum export regions.

-NYC and Miami became tourism juggernauts after adjusting for international flights: my estimate roughly tied out with NYC’s own estimate of international tourism spending which exceeds $30 billion annually.

-Portland lost exports to San Jose after adjusting using the goods trade flows data: I found that while Portland is an important player in export supply chains, it isn’t exporting to the degree that we originally thought since Intel which locates production in Portland supplies most of its semiconductors to San Jose which is the last stop in the production process before the products are sold abroad).

**Real Exports, 2013 (color scale: dark more exports, light less exports)**



**Export Share of GDP, 2013 (color scale: dark more intensive, light less intensive)**

