Cutting Planes for Surveying Italian Agriculture and Industry

R. Bruni, Univ. of Perugia (speaker)

bruni@diei.unipg.it or bruni@dis.uniroma1.it

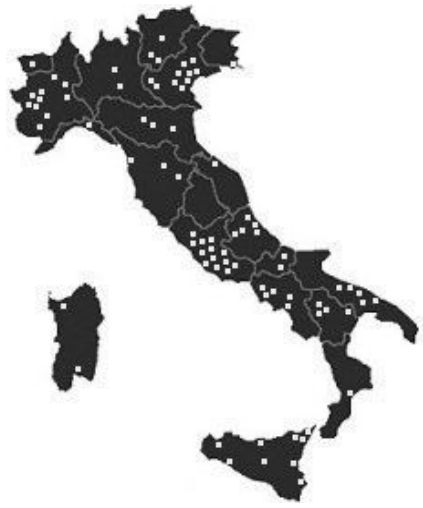
F. Bianchi, G. Bianchi, A. Reale, Istat

XXXIX AIRO Conference - September 8-11, 2008 - Ischia, Italy

Outline

- What is a Censuses of Agriculture or Industry?
- The Universe Selection Problem
- A (Large) Knapsack Model
- How to Solve many of this Large Problems?
- Test on Data from 5° General Census of Agriculture
- Results

Censuses of Agriculture or Industry



- Periodically held in every nation
- Send officials to collect many data from a large number of farms or companies
- Very expensive, but necessary for analyzing many economical and social aspects and therefore plan government policies
- Differently from a population Census, we do not survey everything! The more we survey, the more money we spend
- We have to decide what to survey and what to exclude (called Universe Selection problem): want a fair snapshot of the situation without wasting too much money

Available Data

- Consider, w.l.o.g., the example of agriculture
- We have a list U of n units (the farms) and a set of m cultivations (e.g. oranges, apples, etc.)
- Each unit is $u_i = \{\mathrm{ID}, d_{i1}, d_{i2}, \ldots, d_{im}, d_{i\,TOT}\}$

  Total cultivation area of farm i

  …

  Area used for 2nd cultivation in farm i

  Area used for 1st cultivation in farm i

- Unfortunately, these data are just the data of last census (10 years ago) so things may have changed!
- The single cultivation areas may easily have changed. Total cultivation area of the farm is more stable

Goals and Requirements

- We have some shares $q_1, q_2, \ldots, q_m$ to respect, one for each cultivation (e.g survey at least 0.8 of the total cultivation of oranges, at least 0.5 of the total cultivation of apples, etc.)
- We need to find the maximum cardinality set of units that we can exclude from the survey still respecting the quotes
- Logical conditions on excluded units may sometimes exist
- Since data are uncertain, we want a compromise between saving and sureness. We do the above selection for different ranges of total cultivation area. After this, we survey all of the units belonging to the range that appeared to be the best choice, according to statistical indicators computed using the above set. This was set up by Istat

The Binary Knapsack Model

- Binary variables $x_i = \begin{cases} 1 & \text{if we exclude unit } i \\ 0 & \text{if we survey unit } i \end{cases}$

$$\begin{cases} \max \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} d_{ij}\, x_i \le (1-q_j) \sum_{i=1}^{n} d_{ij} \quad \forall\, j=1,2,\ldots,m \quad \text{(knapsack constraints)} \\ x_i \in \{0, 1\} \qquad i=1,2,\ldots,n \end{cases}$$

- In practice we want to maximize the number of units we do not survey but the excluded area for each cultivation cannot be more than the maximum area we are allowed to exclude
- Many variables, and we need to solve many of this problems
- Just solving by B&B a sequence of problems like this, with >10.000 variables, is quite slow. We can do better applying known results

How to Solve It ?

- We solve the linear relaxation of the problem, finding $\bar{x}$
- After this, we have the separation problem: find a valid inequality for cutting away $\bar{x}$, or conclude that such inequality does not exist
- We look for a cut in the form of a cover of one of the knapsack constraints, i.e. a set of units that, if excluded, exclude from the survey a cultivation area that is too big.
- Call u the incidence vector of a cover C. All d_{ij} are for sure ≥ 0

having a cover means $\sum_{i=1}^{n} d_{ij}\, u_i > (1- q_j)\sum_{i=1}^{n} d_{ij}$ $\quad j=1,2,\ldots, m$

that becomes $\sum_{i=1}^{n} d_{ij}\, u_i \geq [(1- q_j)\sum_{i=1}^{n} d_{ij}\,]+1$

Finding the Cover

- And among all those covers, we want a cover C such that the components of $\bar{x}$ corresponding to elements of C sum to a value $>|C|-1$ ($\bar{x}$ can be **cut away** by the cutting plane generated by C)
- This means a set of units that, if excluded as soon as their value in $\bar{x}$ is > 0, exclude from the survey a too big cultivation area

$$\sum_{i \in C} \bar{x}_i > |C| - 1 \quad \Rightarrow \quad \sum_{i=1}^{n} \bar{x}_i u_i > \sum_{i=1}^{n} u_i - 1$$

$$\Rightarrow \quad \sum_{i=1}^{n} (\bar{x}_i - 1) u_i > -1 \quad \Rightarrow \quad \sum_{i=1}^{n} (1 - \bar{x}_i) u_i < 1$$

- So, we want C such that the above sum is < 1

Separation Problems

- Putting all together, we have the separation problems (one for each j)

$$\begin{cases} \min \sum_{i=1}^{n} (1 - \bar{x}_i)\, u_i \\ \sum_{i=1}^{n} d_{ij}\, u_i \geq [(1-q_j) \sum_{i=1}^{n} d_{ij}\,] + 1 \qquad j=1,2,\ldots,m \end{cases}$$

- If this minimum is < 1 we have the cover C, otherwise we try next j. If we don't find it, it does not exist
- Easily solvable: every u_i has a cost $(1- \bar{x}_i)$ and a value d_{ij} we just order by increasing cost/value and, following this order, put

$u_i = 1$ until $\text{LHS}^i \leq \text{RHS}$, $u_i = (\text{RHS} - \text{LHS}^{i-1})/d_{ij}$, $u_i = 0$ all the rest

Using the Generated Inequalities

- Obtained cover C, we compute its **extension** E(C) by adding all elements whose d_{ij} are ≥ of those of all elements in C
- Obtained the incidence vectors $\bar{u}$ and $\bar{e}$ of the cover and of its extension, we generate the valid **inequality cutting away** $\bar{x}$

$$\sum_{i=1}^{n} \bar{e}_i x_i \leq \sum_{i=1}^{n} \bar{u}_i - 1$$

- We add the inequality to the formulation and solve again the linear relaxation obtaining a **different** solution $\bar{\bar{x}}$
- We repeat the procedure until we have the optimal integer solution

Solving the Sequence of Problems

- When solving the sequence of problems, we keep all generated inequalities, since in every problem we have a different range of total cultivation area, so we have a different subset of the same variables (some inequalities may be useless but it doesn't matter)
- Conditions for detecting covers which are minimal (better) or that produce facets (even better) are known. We are exploring how this could be useful for us (future work)

Test Methodology

- Test on the 5° General Census of Agriculture (the last one)
- 2.594.825 records, one for each farm (many are very small! how expensive would be to survey all of them??)
- Fields : identifier; region; 6 principal cultures; total cultivation area
- C++ procedure calling Cplex 8 (by now, but will move to open source) for solving the LP on a PC Pentium IV 3GHz 1Gb RAM
- Test were performed entirely in Istat, according to laws about data privacy and security

Details on a Single Region

- A sort of preliminary test on a small region (Valle d'Aosta)

Range	Total units	Excluded units	Time (sec.)
$d_{\text{i TOT}} >= 0$	6595	4139	55
$d_{\text{i TOT}} >= 0.1$	6360	3904	43
$d_{\text{i TOT}} >= 0.2$	6098	3642	35
$d_{\text{i TOT}} >= 0.3$	5732	3275	22
$d_{\text{i TOT}} >= 0.4$	5435	2964	18

- According to the statistical indicators set up by Istat (following UE direct. etc.) the most convenient and safe range was the last ($d_{\text{i TOT}} >= 0.4$)
- This would correspond to surveying 5435 farms with 20% savings (a lot !)
- This for being cautious. Otherwise, we could save up to 60% !

Results on All Regions

Dataset	Units	Minimum set	Time (min.)
PIEMONTE	97.982	56.594	15
VALLE D'AOSTA	5.435	2.471	3
LOMBARDIA	67.560	35.886	7
TRENTINO-ALTO ADIGE	44.415	25.675	8
VENETO	166.924	97.737	18
FRIULI-VENEZIA GIULIA	31.939	18.104	5
LIGURIA	28.906	19.281	5
EMILIA-ROMAGNA	101.717	50.838	10
TOSCANA	110.119	58.760	16
UMBRIA	48.445	28.141	8
MARCHE	60.282	36.594	8
LAZIO	170.871	105.284	18
ABBRUZZO	73.810	49.763	8
MOLISE	30.581	20.227	5
CAMPANIA	196.351	130.209	18
PUGLIA	304.481	209.832	30
BASILICATA	72.146	45.132	9
CALABRIA	158.637	103.818	18
SICILIA	297.638	192.473	30
SARDEGNA	92.507	61.594	15

Conclusions

- In a Census of Agriculture or of Industry, given a list of units, we want to find the minimum cardinality set of units ensuring the quotes
- Since data are generally those of the last survey, they may have changed. In order be safe, we do something more: we want to find a range of some reliable filed such that we survey all units in that range.
- So, we need to solve a sequence of large knapsack problems. We use a dynamic simplex approach, with separation routine based on covers
- Results on data from last Census of Agriculture (2000) are very good (good savings being cautious, and more savings possible)
- This approach will be used for the next Census of Agriculture (2010)!