Applied Mathematical Sciences, Vol. 6, 2012, no. 125, 6241 - 6251

Information Reconstruction via Discrete Optimization for Agricultural Census Data*

Gianpiero Bianchi

Istat, Dip. per i Censimenti e gli Archivi Amm. e Statistici (DICA)
Viale Oceano Pacifico 171, 00144 Roma, Italy
gianbia@istat.it

Renato Bruni

Università di Roma "Sapienza"
Dip. di Ingegneria Informatica, Automatica e Gestionale (DIAG)
Via Ariosto 25, 00185 Roma, Italy
bruni@dis.uniroma1.it

Alessandra Reale

Istat, Dip. per i Censimenti e gli Archivi Amm. e Statistici (DICA)
Viale Oceano Pacifico 171, 00144 Roma, Italy
reale@istat.it

Abstract

In the case of large-scale surveys, such as a Census, data may contain errors or missing values. An automatic error detection and correction procedure is therefore needed. We propose here an approach to this problem based on Discrete Optimization. The treatment of each data record is converted into a mixed integer linear programming model and solved by means of state-of-the-art branch and cut procedures. Results on real-world Agricultural Census data show the effectiveness of the proposed procedure.

Keywords: Data Mining, Information Reconstruction, Integer Linear Programming

---

* Work developed during the research collaboration between the Italian Statistic Office (Istat) and the University of Roma "Sapienza" on the data processing of the 2010 Census of Italian Agriculture.

1. Introduction

A Census of Agriculture is a very complex, important and expensive activity for a National Statistic Office. Data collected in such a process have therefore a great economic value. As in any other large-scale survey, however, those data may contain errors or missing values, and their automatic *detection* and *correction* are crucial tasks. This kind of activity is generally called *Information Reconstruction*, or also *Data Cleaning*, within the field of Data Mining [9], or *Data Editing and Imputation* within the field of Statistics [6,15]. Data are generally organized into conceptual units called *records*. A record [14] has the formal structure of a set of n fields $R = (f_1, \ldots, f_n)$, and by giving each field f_i a value v_i we obtain a record instance, or, simply, a record $r = (v_1, \ldots, v_n)$. In the case of a Census of Agriculture data are typically constituted by farm codes, cultivation codes, size of cultivation areas and other amounts, years, etc., so we restrict our attention to numerical data.

The above Information Reconstruction tasks can be performed by following several different approaches, each of which having its own features. However, satisfactory data quality and computational efficiency often appear to be at odds. A main approach is based on the use of rules, called *edits*, that each data record must respect in order to be declared exact [1,4]. Records not respecting such rules are declared erroneous. A seminal paper on the subject is due to Fellegi and Holt [7]. Rules can often be converted into mathematical expressions, e.g. inequalities, and finding within a record the most probably erroneous fields or the most suitable values correcting those fields can be modeled as nontrivial optimization problems (see e.g. [8] for an introduction to the field of computational complexity). This allows to overcome the computational limits of other methodologies (see e.g. [2,15]) based on the Fellegi Holt approach. This has been done within the data Editing and Imputation software system DIESIS [4,5], and, subsequently, in other works such as [6,13].

In the described Census, each *farm* specifies the *cultivation area* used for each cultivation. A classical problem is verifying the accuracy of this information. Errors should be detected and corrected, by mathematically "guessing" the correct values, since it is clearly impossible to contact again the farm or inspect somehow the cultivations. Farms can extend on one or more districts, and the area owned by each farm in each district is known. The compatibility of each cultivation with each district can also be evaluated (some cultivation can grow only on specific types of soils, or need specific climatic conditions, latitude, altitude, etc.). Therefore, in principle, the elements for checking the above information can be gathered, but the problem is doing this on large data sets both efficiently and in an unbiased manner.

This work presents an innovative automatic procedure for solving this problem based on a discrete mathematical optimization model. In particular, Section 2 describes problem details and the proposed mixed integer linear programming model [3,12], explaining its features. Section 3 reports

computational results in the case of the Italian Census of Agriculture 2010, with specific respect to the case of vineyards. This is probably the most important case for the considered problem, since dozens of vine varieties exist, and they determine the type and the quality of wines produced. Relevant economic aspects are therefore involved, and vine cultivation and use are also regulated by legislation. Clearly, the proposed model is not limited to the case of a Census of Agriculture, but can be used for any other problem sharing the same characteristics.

2. A Discrete Mathematical Model

Data obtained from each farm during the described Census contain information about the area used by that farm for each cultivation. Those data may sometimes be erroneous or missing, due to a variety of reasons. In such cases, errors should be automatically detected and corrected, i.e. the information that was corrupted and lost should be "reconstructed" in order to be as similar as possible to the (unknown) exact value.

Farms are located over the state territory. This territory is subdivided into many districts. Each farm can extend on one or more district. Denote by

$I = \{1,\ldots, n\}$ the set of all possible cultivations;
$J = \{1,\ldots, m\}$ the set of all possible districts.

We focus on a single farm, and denote by f its total area. We check all the cultivations declared by that farm. Some of them verify a set of rules and conditions prepared for this aim and are therefore considered reliable, while some other do not. This may happen either because some of the declarations appear erroneous, or because there is a discordance between the total area declared and the sum of the areas declared for each cultivation. Denote by a the total farm area reliably assigned, i.e. the area for which the farm declaration are considered reliable. On the contrary, by grouping all the unreliable declarations, a nonempty area often remains for which the cultivation is not known. That area will be called *unassigned area* and denoted by u. Clearly, $f = a+u$. The central problem of our Information Reconstruction process consist now in assigning the cultivations to the mentioned unassigned area. We propose in this Section a discrete mathematical model for this problem.

For each farm, denote by

s_i (real value ≥ 0) the total area that the farm uses for cultivation i, with $i \in I$. Note that this area may span on one or more districts, and the farm does not declare, nor generally even consider, such subdivision. These values are only the ones, among all the cultivation data declared by farms, that can be considered reliable, so $\sum_i s_i = a$.

d_j (real value ≥ 0) the total area owned by the farm in district j, with $j \in J$. These values are not surveyed during the considered Census but are already available and are reliable.

p_{ij} (real value $\in [0,1]$) the likelihood of having cultivation i in district j, with $i \in I$ and $j \in J$. Values near to 1 means high likelihood, near to 0 means very low likelihood. This values are estimated on the basis of agricultural registrations and studies, not surveyed during the considered Census.

Moreover, there are areas where specific cultivations may be used to produce foods having "controlled origin" (In Italian DOC). In particular, for the unassigned area u, it is possible to partition it into a portion that is suitable for "controlled origin" and a portion that is not suitable for that. Denote by

C (real value ≥ 0) the total unassigned area owned by the farm in cultivations suitable for "controlled origin";
N (real value ≥ 0) the total unassigned area owned by the farm in cultivations not suitable for "controlled origin", so that $C+N = u$.

Those areas C and N should be assigned in order to maximize the likelihood of the assignment. Note that it is not known which district the unassigned area u is located into. On the other hand, the likelihood values depend on the districts. As a consequence, we need to locate the unassigned area u on the districts. This is apparently hard to obtain. A way of doing so is locating on the districts each of the reliable cultivation areas s_i, , and then obtaining the location of u as the portion of farm area f not covered by a. In order to model the described problem, we need to introduce the following sets of decision variables:

x_{ij} (real value ≥ 0) the area of cultivation i that, according to our reconstruction, is localized in district j, with $i \in I$ and $j \in J$;
v_{ij} (real value ≥ 0) the portion of C that, according to our reconstruction, is used for cultivation i and localized in district j, with $i \in I$ and $j \in J$;
w_{ij} (real value ≥ 0) the portion of N that, according to our reconstruction, is used for cultivation i and localized in district j, with $i \in I$ and $j \in J$.

Moreover, each of the farm unassigned areas C and N generally contains only a specific cultivation, and not a mixture of different cultivations. We therefore want to assign all C to one single type of cultivation, and not to fragment it among all the cultivations compatible with that area. A similar requirement holds for N. This requires the use of additional binary decision variables

$$y_i = \begin{cases} 1 & \text{if } C \text{ is assigned in our reconstruction to cultivation } i, \text{ with } i \in I \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if } N \text{ is assigned in our reconstruction to cultivation } i \text{, with } i \in I \\ 0 & \text{otherwise} \end{cases}$$

We can now formulate a mixed integer linear programming model for each farm. Cultivation assignment to areas should be done in order to maximize the likelihood. Our objective function is therefore

$$\max \sum_{\forall i,j} p_{ij} x_{ij} + \sum_{\forall i,j} p_{ij} v_{ij} + \sum_{\forall i,j} p_{ij} w_{ij}$$

This assignment should obviously verify a set of constraints. First of all, the sum of the areas assigned to the different cultivations in each district j must be equal to the area owned by the farm in district j :

$$\sum_{\forall i} x_{ij} + \sum_{\forall i} v_{ij} + \sum_{\forall i} w_{ij} = d_j \qquad \forall j = 1,\ldots, m$$

The sum of the areas used by the farm for cultivation i over all the districts must be equal to the total area used by the farm for cultivation i :

$$\sum_{\forall j} x_{ij} = s_i \qquad \forall i = 1,\ldots, n$$

The sum of the portions of C assigned to all cultivations in all districts must be equal to C. A similar condition must hold for N.

$$\sum_{\forall i,j} v_{ij} = C \qquad \sum_{\forall i,j} w_{ij} = N$$

In order to connect the y and z variables to v and w, we need to impose that it is not possible assigning a portion of C [resp. of N] to cultivation i (regardless to the district) when the corresponding variable y_i [resp. z_i] is 0. Note that M is a constant value greater than all possible left-hand-side values.

$$v_{ij} \le M y_i \qquad \forall i = 1,\ldots, n \;\; \forall j = 1,\ldots, m$$
$$w_{ij} \le M z_i \qquad \forall i = 1,\ldots, n \;\; \forall j = 1,\ldots, m$$

The whole C must be assigned to only one cultivation. A similar condition must hold for N.

$$\sum_{\forall i} y_i = 1 \qquad \sum_{\forall i} z_i = 1$$

The above constraints have the effect of letting only one y [resp. only one z] be 1, and so the previous constraints can only assigning C [resp. N] to a unique cultivation.

Finally, an assignment for C or for N cannot be accepted when the likelihood of that assignment, although being the greatest possible for the current problem, is too low. In such a case, indeed, that assignment cannot be considered reliable. For this reason we introduce two thresholds, denoted by SC and SN, respectively for the assignments made on C and N.

$$\sum_{\forall j} v_{ij} - \sum_{\forall j} p_{ij} v_{ij} - SC \leq M(1 - y_i) \qquad \forall i = 1,\ldots, n$$

$$\sum_{\forall j} w_{ij} - \sum_{\forall j} p_{ij} w_{ij} - SN \leq M(1 - z_i) \qquad \forall i = 1,\ldots, n$$

When the likelihood of assigning C to cultivation i is good (= near to 1) for the different districts where C have been located, that assignment can hold, that means y_i can assume value 1. On the other hand, when that likelihood is not good (= near to 0), that assignment cannot hold, that means y_i must be forced to value 0. Note that, if no assignment has a sufficient likelihood, those constraints cannot be satisfied and the model correctly becomes infeasible.
The above is obtained because, for assignments having good likelihood, the value of $\sum_{\forall j} p_{ij} w_{ij}$ is only a bit smaller than the value of $\sum_{\forall j} v_{ij}$, and by subtracting SC the left-hand-side of the inequality becomes smaller than or equal to 0, leaving y_i free.

When on the contrary the likelihood is not good, the value of $\sum_{\forall j} p_{ij} w_{ij}$ is much smaller than the value of $\sum_{\forall j} v_{ij}$, and even subtracting SC (whose reasonable value is therefore just a fraction of $\sum_{\forall j} v_{ij}$, for instance one half) the left-hand-side of the inequality becomes positive. As a consequence, $M(1 - y_i)$ must have a strictly positive value, and so y_i must have value 0.

On the whole, the complete mixed integer linear programming model for assigning the unassigned area of a single farm is the following:

$$
\begin{cases}
\max \sum_{\forall i,j} p_{ij} x_{ij} + \sum_{\forall i,j} p_{ij} v_{ij} + \sum_{\forall i,j} p_{ij} w_{ij} \\
\sum_{\forall i} x_{ij} + \sum_{\forall i} v_{ij} + \sum_{\forall i} w_{ij} = d_j \qquad \forall j = 1,\ldots,m \\
\sum_{\forall j} x_{ij} = s_i \qquad \forall i = 1,\ldots,n \\
\sum_{\forall i,j} v_{ij} = C \qquad \sum_{\forall i,j} w_{ij} = N \\
v_{ij} \le M y_i \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
w_{ij} \le M z_i \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
\sum_{\forall i} y_i = 1 \qquad \sum_{\forall i} z_i = 1 \\
\sum_{\forall j} v_{ij} - \sum_{\forall j} p_{ij} v_{ij} - SC \le M(1 - y_i) \qquad \forall i = 1,\ldots,n \\
\sum_{\forall j} w_{ij} - \sum_{\forall j} p_{ij} w_{ij} - SN \le M(1 - z_i) \qquad \forall i = 1,\ldots,n \\
x_{ij} \ge 0 \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
v_{ij} \ge 0 \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
w_{ij} \ge 0 \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
x_{ij} \in \Re \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
v_{ij} \in \Re \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
w_{ij} \in \Re \qquad \forall i = 1,\ldots,n \;\; \forall j = 1,\ldots,m \\
y_i \in \{0,1\} \qquad \forall i = 1,\ldots,n \\
z_i \in \{0,1\} \qquad \forall i = 1,\ldots,n
\end{cases}
$$

3. Computational Results

By sequentially solving the above model for each farm, we perform the requested Information Reconstruction process. This procedure was implemented in C++,

using ILOG Concert Technology [10] in order to express the described optimization models. The models themselves are solved by means of the state-of-the-art branch-and-cut [3,12] procedure implemented by the solver ILOG Cplex [11], running on a 16 cores server having 128Gb of RAM and Linux Operating System. The resulting software system has been tested for the treatment of data from the Italian Census of Agriculture 2010, with specific respect to the case of vineyards. This is probably the most important case for the considered problem, since dozens of vine varieties exist, and they determine the type and the quality of wines produced. The case has therefore great economic relevance and, due to its large dimension, is also computationally demanding. Moreover, the total area that a region uses for each vine type determines the European Community funding obtained for that region. Note that, to the best of our knowledge, no previous attempt to treat this problem with a discrete optimization approach was practically successful. Clearly, the proposed model is not limited to the case of vineyards, but can be used for any other problem sharing the same characteristics.

The practical behavior of the proposed procedure has been evaluated both from the computational and from the data quality points of view, as follows. One large dataset including all farms producing vine from all Italian regions has been assembled. This dataset included 388,487 farms, and the total number of vineyard areas declared was 804,930, corresponding to a total area of 625,700 ha (hectares, 1 hectare being 10,000 m^2). Each vineyard declaration constitutes a record, so the dataset is considerably large. After this, the cultivation declarations where checked by means of rules, and an unassigned area resulted for 18,263 farms. The total unassigned area was 34,783 ha, with 30,226 ha suitable for "controlled origin" and 4,557 ha not suitable for "controlled origin". As remarked in Section 1, thresholds were assigned in order to reject low likelihood cultivation assignments, in particular $SC = \frac{1}{2}\sum_{\forall j} v_{ij}$ and $SN = \frac{1}{2}\sum_{\forall j} w_{ij}$.

In order to evaluate the computational behavior, Table 1 reports regional detail, that are: the total unassigned area for each region, and the corresponding computational times for processing the whole region. As observable, the procedure is very fast, and the processing of all the Italian farms having an unassigned area requires only about 17 minutes.

The quality of the reconstructed information has been evaluated by considering: (i) the ability to assign the unassigned area; and (ii) the variation produced in the data by the reconstruction process. As for the first aspect, the procedure was able to assign the unassigned area with likelihood values high enough to satisfy the thresholds *SC* and *SN* in the totality of the cases (100%). As for the second aspect, Table 2 reports an analysis of the percentages of the two big groups of vine cultivations (red and white) on the initial data, i.e. before applying the described reconstruction process, and on the final data, i.e. after the reconstruction process, followed by the computation of the variation introduced by this process. Note that, in the general case of the correction of a survey, a small

variation of the frequency distributions of the data means that the reconstruction procedure was able to reconstruct information without distorting the data, so it is a positive feature. In this case, as observable from the Table, the variation value has been very small, so the quality of the reconstructed information is extremely satisfactory. Moreover, we report in the same Table the described percentages computed only over the farms not having unassigned areas ("Exact only"). Those percentages are again very similar to the percentages after reconstruction, confirming the high quality of the reconstructed information.

Region	Total unassigned area (ha)	Times (sec.)
Piemonte	42.89	0.82
Valle d'Aosta	10.08	1.98
Lombardia	271.19	34.61
Veneto	12510.09	165.75
Friuli-Venezia Giulia	87.94	13.95
Liguria	52.87	15.60
Emilia-Romagna	20.24	0.44
Toscana	8293.34	160.14
Umbria	100.75	14.50
Marche	2517.57	65.76
Lazio	263.03	38.02
Abruzzo	259.62	34.39
Molise	397.68	22.80
Campania	511.53	86.31
Puglia	8939.25	303.37
Basilicata	13.09	0.66
Calabria	331.33	24.45
Sicilia	49.41	1.15
Sardegna	8.35	0.49
Bolzano	0.88	0.05
Trento	101.47	22.74
Italy total	34782.60	1008.00

Table 1: Total correction times for each region and for all Italy.

	Before Reconstruction	After Reconstruction	Variation	Exact only
White	43.55%	42.84%	-0.71%	43.95%
Red	56.45%	57.16%	+0.71%	56.05%

Table 2: Analysis of frequency distributions before and after the reconstruction process.

4. Conclusions

Information Reconstruction is a crucial task for large surveys, such like a Census of Agriculture, as well as for any other application requiring database cleaning. The problem have been tackled in several different manners, but satisfactory data quality and computational efficiency often appear to be at odds. A typical problem arising in the described Census consists in checking, and correcting when needed, the areas declared by each farm for each cultivation. The problem is extremely important due to its economical and normative aspects, and is also computationally demanding due to its large dimension. This paper presented an automatic approach, based on the formulation of mixed integer linear programming models. The procedure has been tested in the case of the Italian Census of Agriculture 2010, with specific respect to the case of vineyards, which is probably the most important case for the considered problem. Clearly, the proposed model is not limited to the case of a Census of Agriculture, but can be used for any other problem sharing the same characteristics. Results are very encouraging both form the computational and from the data quality point of view. The sequence of arisen mixed integer programming problems can be solved to optimality by using state-of-the-art implementation of branch-and-cut procedures. Each single model is solved to optimality in extremely short times (generally about 0.1 sec.). In the totality of the cases the reconstructed information was able to satisfy the thresholds introduced to reject low likelihood cultivation assignments. Moreover, the reconstruction process did not distort data, as resulted from the analysis of the variations introduced in the frequency distributions of the whole dataset.

References

[1] Banff Support Team, Functional Description of the Banff System for Edit and Imputation System, Statistics Canada, Quality Assurance and Generalized Systems Section Tech. Rep, 2003.

[2] M. Bankier, Canadian Census Minimum change Donor imputation methodology, Proceedings of the Workshop on Data Editing, UN/ECE, Cardiff, United Kingdom, 2000.
[3] D. Bertsimas and J.N. Tsitsiklis, Introduction to Linear Optimization, Athena Scientific, Belmont, Massachusetts, 1997.
[4] R. Bruni, Discrete Models for Data Imputation, Discrete Applied Mathematics Vol. 144/1 (2004), 59-69.
[5] R. Bruni, Error Correction for Massive Data Sets, Optimization Methods and Software, Vol. 20/2-3 (2005), 295-314.
[6] T. De Waal, Computational Results with Various Error Localization Algorithms, UNECE Statistical Data Editing Work Session, Madrid, Spain, 2003.
[7] I.P. Fellegi and D. Holt, A systematic approach to automatic edit and imputation, Journal of the American Statistical Association 71 (1976), 17-35.
[8] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman and Company, San Francisco, CA, 1979.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, New York, NY, 2001.
[10] IBM Ilog Concert Technology 12.1 Reference Manual, International Business Machines Corporation, 2009.
[11] IBM Ilog Cplex 12.1 Reference Manual, International Business Machines Corporation, 2009.
[12] G.L. Nemhauser and L.A. Wolsey, Integer and Combinatorial Optimization, John Wiley & Sons, Inc., New York, NY, 1999.
[13] J. Riera-Ledesma and J.J. Salazar-Gonzalez, New Algorithms for the Editing and Imputation Problem, UNECE Statistical Data Editing Work Session, Madrid, Spain, 2003.
[14] R. Ramakrishnan and J. Gehrke, Database Management Systems (3rd edition), McGraw-Hill, New York, NY, 2003.
[15] W.E. Winkler, State of Statistical Data Editing and current Research Problems, Proceedings of the Workshop on Data Editing, UN/ECE, Rome, Italy, 1999.

Received: July, 2012