Balancing of Agricultural Census Data by using Discrete Optimization*

Bianchi Gianpiero[1] · **Bruni Renato**[2] · **Reale Alessandra**[3]

Abstract In the case of large-scale surveys, such as a Census, data may contain errors or missing values. An automatic error correction procedure is therefore needed. We focus on the problem of restoring the consistency of agricultural data concerning cultivation areas and number of livestock, and we propose here an approach to this balancing problem based on Optimization. Possible alternative models, either linear, quadratic or mixed integer, are presented. The mixed integer linear one has been preferred and used for the treatment of possibly unbalanced data records. Results on real-world Agricultural Census data show the effectiveness of the proposed approach.

Keywords Data Mining, Balancing Problems, Information Reconstruction, Mixed Integer Linear Models

1 Introduction

A Census of Agriculture is a very complex, important and expensive operation for a National Statistic Office. It is an essential activity, periodically performed for monitoring the agricultural sector (see also [14]). Data collected in such a process have therefore a great intrinsic economic value, and moreover, in the case of EU countries, constitute a basis for assigning financial resources, planning production, and for several other economical European policies. As in any other large-scale survey, however, those data may contain errors or missing values, due to a variety of reasons. Nonetheless, correct information must be published and provided to the EU level, also considering that large financial resources are allocated to the sector. Therefore, error *detection* and *correction* become crucial tasks. This kind of activity is generally called *Information Reconstruction*, or also *Data Cleaning*, within the field of Data Mining (see also [18,23]), or *Data Editing and Imputation* within the field of Statistics (see also [12,31]). Note that, in contexts different from the Census, the possibility of reconstructing exact values could be useful also for counteracting possible opportunistic behaviors (e.g. willingly erroneous declarations), and knowing that exact values can be reconstructed could indeed prevent such opportunistic behaviors.

Data are generally organized into conceptual units called records (see also [28]). 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. Agriculture is a rich source of large data mining problems, and a recent overview on the use of data mining techniques in this field is in [25]. The above Information Reconstruction tasks, in particular, can be performed by following different approaches, each of which having its own features. A main approach is based on the use of rules, called *edits*, that each data record must respect in order to be declared exact (see e.g. [3,24]). Records not respecting such rules are declared erroneous. A seminal paper on the subject is [15]. However, satisfactory rules accuracy and computational efficiency often appear to be at odds. For this reason, rules are often converted into mathematical expressions, e.g. inequalities (see also [11]), and finding within a record the most probably erroneous fields or the most suitable values correcting those fields become nontrivial optimization problems (see e.g. [17] for an introduction to computational complexity). This allows to overcome the computational limits of other techniques (see e.g. [4,24,31]). Such a

* Work developed during the biennial 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] Bianchi Gianpiero Istat, Dip. per i Censimenti e gli Archivi Amm. e Statistici (DICA)
Viale Oceano Pacifico 171, 00144 Roma, Italy

[2] Bruni Renato Università di Roma "Sapienza" Dip. di Ingegneria Informatica, Automatica e Gestionale (DIAG) Via Ariosto 25, 00185 Roma, Italy
e-mail: bruni@dis.uniroma1.it

[3] Reale Alessandra Istat, Dip. per i Censimenti e gli Archivi Amm. e Statistici (DICA)
Viale Oceano Pacifico 171, 00144 Roma, Italy

methodology has been adopted within the data Editing and Imputation software system DIESIS [9,10] and in other works such as [12,27].

In the described Census, each farm specifies the *cultivation area* used for each cultivation and *number of livestock* for each type of animal, divided in some cases also by year. Moreover, they specify total areas and total numbers of livestock. However, those totals may be inconsistent with the mentioned detailed information, and a classical problem is restoring data consistency by correcting errors. These errors should be corrected by mathematically "guessing" the correct values, since it is clearly impossible to contact again the farm or inspect it somehow. The main issue is doing this on large data sets both efficiently and in order to obtain corrected data as similar as possible to the exact (but unknown) data. This work presents an innovative procedure for solving this problem based on optimization. In particular, Section 2 describes in detail the specific problem structure, analyzing also its connections to similar problems, and explains the development of the proposed integer linear programming model. Section 3 reports computational results in the case of the Italian Census of Agriculture 2010 ("Censimento Generale dell'Agricoltura 2010"), both for plants cultivations and for livestock. Note that, to the best of our knowledge, no previous attempt to treat this large-size Census problem with a discrete optimization approach was made, and only *ad hoc* procedures, designed by experts after an analysis of the specific available data, were used.

2 Problem Structure and Optimization Model

Data obtained from each *farm* during the described Census contain information (called *microdata*, information about details) about the *cultivation area* used by that farm for each cultivation and the *number of livestock* for each type of animal. 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. Moreover, each farm also declares other information (called *macrodata*, information about totals): the total cultivation area and the total number of livestock, and in some cases those totals are also divided into *subtotals* by year of planting. Clearly, balancing conditions must hold between all the above microdata and the corresponding macrodata: each total (or year subtotal) must be equal to the sum of those details concerning its parts. When such conditions do not hold, data are inconsistent.

Records incurring in this problem are detected by checking the balancing conditions, which are called *balance edits*. However, when a balance edit is violated, the error could be either on the detail side or on the total side of the equation. The less reliable information should be changed in order to restore consistency. It is generally assumed, in these cases, that details constitute the less reliable information, since totals have already been confirmed from other sources. This mathematical problem of adjusting the entries (here the microdata) of a large matrix to satisfy prior consistency requirements (here given by the macrodata) is called *matrix balancing* [29] and occurs in several fields, such as economics, urban planning, statistics, demography, etc. The problem is also related to the matrix rounding problem [1], consisting in rounding off the elements of a matrix consistently with its row and column sums, often arising in economic statistics, and belongs to the broad category of *matrix scaling* problems [2].

In some cases of matrix balancing problems the only aim is restoring balancing without further objectives, and iterative scaling algorithms can be used, e.g. the RAS algorithm [22]. In other cases, on the contrary, the variations introduced for balancing the matrix should pursue an objective that typically depends on the specific application. In the case of Census data, the choice of the objective is a delicate issue for avoiding data distortions, and makes this problem different from other types of balancing problems. Errors in microdata could broadly be divided into systematic errors and random errors [16]. Systematic errors are those caused by specific (and often traceable) mechanisms, e.g. usage of a wrong unit of measurement, OCR error, etc., and are generally treated during a preliminary correction phase [13]. Our central problem is therefore correcting microdata values affected by random errors. In this case, changes from the available microdata values should be minimized, according to specific distance criteria, since it is generally deemed that this should produce data as similar as possible to the unknown exact data (Fellegi-Holt paradigm [15,24]). An optimization approach is therefore required.

The models proposed for the above problem will be hereinafter explained by referring to the specific case of *vineyards*. This is one of the most important cases: dozens of grapes varieties exist, and they determine type and quality of wines produced. The case has great economic relevance and, due to its large dimension, is also computationally demanding. Moreover, those

data are used when allocating European financial resources and when reorganizing wine production. However, the proposed models are clearly not limited to that case, but can be used for any other similar problem.

Each farm could have several vine types, and each of them could have been planted in a different time period (e.g. a specific year). Denote by

$I = \{1,\ldots, n\}$ the set of indices of all possible vine types; with $n = 442$

$K = \{1,\ldots, m\}$ the set of indices of all possible time periods; with $m = 6$.

For each farm, denote by

a_{ik} (real valued ≥ 0) the area of vine type i planted in period k declared by the farm, with $i \in I$ and $k \in K$;

a_{i0} (real valued ≥ 0) the total area of vine type i (planted during any of the periods) declared by the farm, with $i \in I$;

T_k (real valued ≥ 0) the total vine area planted in period k declared by the farm, with $k \in K$;

T (real valued ≥ 0) the total vine area owned by the farm.

In order to reconstruct the erroneous information, we need the following set of decision variables:

x_{ik} (real valued $\geq 0, \leq S$) = the area of vine type i that, according to our reconstruction, has been planted in period k by the farm, with $i \in I$ and $k \in K$.

x_{i0} (real valued $\geq 0, \leq mS$) = the total area of vine type i that, according to our reconstruction, has been planted (during any of the periods) by the farm, with $i \in I$.

In other words, x_{ik} is the correct value for a_{ik}. When reconstructing information for a Census, as in the case of other large-scaled surveys, it is generally assumed that the changes introduced in the data should be somehow minimized. This because, in absence of further information, being as similar as possible to the exact (unknown) data corresponds to being as similar as possible to the available (even if possibly erroneous) data. By following this minimum change paradigm, two basic alternatives exist: one is minimizing the number of changes, the other minimizing the amount of those changes.

If we need to distinguish when our reconstruction provides a result which is different form the available declaration (i.e. a change), we need the following set of binary variables:

$$y_{ik} = \begin{cases} 1 & \text{if } x_{ik} \text{ is different from } a_{ik} \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m$$

The presence of binary variables clearly has its impact on the complexity of the model: by adding the other constraints needed for this problem, which are linear, we obtain an Integer Linear Program. Minimizing the total number of changes corresponds to the following objective function

$$\min \sum_{i=1}^{n} \sum_{k=0}^{m} y_{ik} \tag{1}$$

When variables y are used, they should be linked to the x variables by constraints imposing that y_{ik} takes value 1 when $x_{ik} <> a_{ik}$ (using a certain numerical precision), otherwise those variables could be inconsistent. There is no need for constraints imposing $y_{ik} = 0$ when $x_{ik} = a_{ik}$ because the objective (1) itself does that. Value M is a real number greater than all possible values of the left-hand side of the following inequalities.

$$a_{ik} - x_{ik} \leq M y_{ik} \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m \tag{2}$$

$$x_{ik} - a_{ik} \leq M y_{ik} \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m \tag{3}$$

When, on the other hand, we are interested in measuring the difference between our reconstruction x_{ik} and the available declaration a_{ik}, we should consider a generic norm of this difference:

$$\| a_{ik} - x_{ik} \|_p$$

Several norm types and norm-induced functions exist [5]. We consider more suitable to our reconstruction problems the following three:

- The squared Euclidean norm, defined as $(\|u - v\|_2)^2 = \sum_{h=1}^{q} (u_h - v_h)^2$
- the so-called Manhattan norm, defined as $\|u - v\|_1 = \sum_{h=1}^{q} |u_h - v_h|$
- the so-called Chebyshev norm, defined as $\|u - v\|_\infty = \max_h \{ |u_h - v_h| \}$.

Clearly, the structure of the optimization model that we must solve depends now on this choice. In the first case (squared Euclidean norm), minimizing the total amount of the changes corresponds to the following objective function, containing quadratic terms.

$$\min \sum_{i=1}^{n} \sum_{k=0}^{m} (a_{ik} - x_{ik})^2 = \min \sum_{i=1}^{n} \sum_{k=0}^{m} (a_{ik}^2 - 2 a_{ik} x_{ik} + x_{ik}^2) \qquad (4)$$

However, all of them are simply squared variables $(x_{ik})^2$, so they are strictly convex, and a conic combination of those strictly convex terms produces a separable strictly convex function [8]. By adding to that the linear terms of (4) and the constraints needed for this problem (described later), which are linear, the problem remains efficiently solvable (see e.g. [6,19]).

In the second case (Manhattan norm), there are absolute values in the objective. However, they can be easily linearized by introducing additional variables:

$$s_{ik} \text{ (real valued } \geq 0) = \text{the value of } |a_{ik} - x_{ik}|, \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m$$

and linear constraints enforcing their meaning

$$s_{ik} \geq a_{ik} - x_{ik}, \quad s_{ik} \geq x_{ik} - a_{ik} \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m \qquad (5)$$

We can now minimize the linear function $\sum_{i=1}^{n} \sum_{k=0}^{m} s_{ik}$. When adding the other constraints needed for this problem, which are linear, the problem becomes an easily solvable Linear Program.

In the third case (Chebyshev norm), we have a min-max objective in the problem that again can be easily linearized by introducing one additional variable

$$t \text{ (real valued } \geq 0) = \text{the value of } \max_{ik} \{ |a_{ik} - x_{ik}| \}, \quad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m$$

and linear constraints enforcing the above meaning

$$t \geq a_{ik} - x_{ik}, \quad t \geq x_{ik} - a_{ik} \qquad \forall\, i = 1,\ldots, n \qquad \forall\, k = 0,\ldots, m \qquad (6)$$

We now simply minimize t. When adding the other constraints needed for this problem, which are linear, the problem becomes again an easily solvable Linear Program.

Clearly, also a combination of the above alternatives can be considered. The characteristics of the specific real problem will determine, from case to case, the choice of the objective among the described ones or their possible combinations. In our case, we consider more representative of the real problem's aim the minimization of the total number of changes, and, in second place, the minimization of the amount of those changes. This because a change with respect to a value that has been deliberately declared has intrinsically a very high cost. Therefore, we prefer maintaining the maximum number of those declared values, even if this may result in a greater amount of the changes that we are forced to introduce. The objective function becomes:

$$\min \left(M' \sum_{i=1}^{n} \sum_{k=0}^{m} y_{ik} + \sum_{i=1}^{n} \sum_{k=0}^{m} s_{ik} \right) \qquad (7)$$

where the first sums are multiplied by a numerical value M' weighting the relative importance of the first part with respect to the second one. We chose $M'=S$, so that a single change weights as much as the maximum amount of a change.

We now describe the balancing conditions that should be respected in our case. The sum of vine areas of any type planted in period k must be equal to the total vine area planted in period k (called balancing over vine types)

$$\sum_{i=1}^{n} x_{ik} = T_k \qquad \forall\ k \in K \tag{8}$$

The sum of the areas of vine type i planted in periods from 1 to m must be equal to the area of the same vine type planted along all the periods (called balancing over time periods)

$$x_{i0} = \sum_{k=1}^{m} x_{ik} \qquad \forall\ i \in I \tag{9}$$

The sum of vine areas of any type planted in any period must be equal to the total vine area owned by the farm (called overall balancing)

$$\sum_{i=1}^{n} \sum_{k=1}^{m} x_{ik} = T \tag{10}$$

Clearly, any other type of balancing condition could be expressed as other linear constraints. Note that the structure of balancing constraints (8) and (9) could be considered as defining a transportation problem (see e.g. [5,30]) with a set of origins I and a set of destinations K, values a_{i0} being the supply at origin i, values T_k being the demand at destination k, variables x_{ik} being the amount to be shipped from source i to destination k. However, the values a_{i0} are in our case declared values that we may change (using the x_{i0} variables), and moreover there is no guarantee that the following condition, essential for the feasibility of a transportation problem, is respected:

$$\sum_{i=1}^{n} a_{i0} = \sum_{k=1}^{m} T_k$$

The complete mixed integer linear programming model is therefore the following:

$$\begin{cases}
\min\ (M' \sum_{i=1}^{n} \sum_{k=0}^{m} y_{ik} + \sum_{i=1}^{n} \sum_{k=0}^{m} s_{ik}) & \\
\sum_{i=1}^{n} x_{ik} = T_k & \forall\ k \in K \\
x_{i0} = \sum_{k=2}^{m} x_{ik} & \forall\ i \in I \\
\sum_{i=1}^{n} \sum_{k=1}^{m} x_{ik} = T & \\
a_{ik} - x_{ik} \le M y_{ik} & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
x_{ik} - a_{ik} \le M y_{ik} & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
s_{ik} \ge a_{ik} - x_{ik} & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
s_{ik} \ge x_{ik} - a_{ik} & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
0 \le x_{ik} \le S & \forall\ i \in I \quad \forall\ k \in K \\
0 \le x_{i0} \le mS & \forall\ i \in I \\
s_{ik} \ge 0 & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
x_{ik},\ s_{ik} \in \Re & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m \\
y_{ik} \in \{0,1\} & \forall\ i = 1,\ldots,n \quad \forall\ k = 0,\ldots,m
\end{cases} \tag{11}$$

3 Computational Analysis

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 [20] in order to express the described optimization models. The models themselves are solved by means of the state-of-the-art branch-and-cut (see e.g. [5,26]) procedure implemented by the solver ILOG Cplex [21], 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 ("Censimento Generale dell'Agricoltura 2010"), with specific respect to the cases of:

(I) Vineyards suitable for "controlled origin" wine, considered in Table 1;
(II) Vineyards not suitable for "controlled origin" wine, considered in Table 2;
(III) Generic cultivations, considered in Table 3;
(IV) Livestock, considered in Table 4.

Note that, in the last two cases, microdata are not subdivided by year of planting but by geographical area. In the above four cases, we report results for each Italian region and for all Italy (1st column): the total number of farms not respecting the balancing conditions (2nd column); the total number of records involved in those unsatisfied balancing conditions (3rd column); the total number of changes operated by the reconstruction process (4th column). Moreover, we analyze in greater detail those changes: we report the percentages of area (or heads) modified by the procedure, computed with respect to the total area involved in that case (or to the total number of animals). Such modifications can be done by adding (5th column) and/or by subtracting (6th column), and note that those quantities are not bounded to be equal, since errors are not so. Finally, we report the total processing time in seconds (7th column).

The practical behavior of the proposed procedure should now be evaluated both from the computational and from the data quality points of view. As observable, the procedure is very fast: each single model is solved to optimality in extremely short times (generally about 0.02 sec.) so that the processing of all the Italian farms requires, for the 4 cases together, only about 50 minutes. The quality of the obtained data has been evaluated by considering: (i) the ability to restore balancing; and (ii) the variation produced in the data by the reconstruction process.

Region	Farms	# Records	Changes	Added Area	Subtracted Area	Time (sec.)
Piemonte	696	2055	1866	0.15%	-1.70%	23.5
Valle d'Aosta	22	55	46	0.00%	0.00%	0.6
Lombardia	468	1488	1423	0.16%	-0.75%	17.0
Veneto	3817	6916	5602	4.23%	-0.96%	79.2
Friuli-Venezia Giulia	286	1528	945	0.13%	-0.38%	17.5
Liguria	124	257	493	0.01%	-0.04%	2.9
Emilia-Romagna	336	940	640	0.07%	-0.74%	10.8
Toscana	3392	6099	4332	3.20%	-1.01%	69.8
Umbria	73	248	173	0.05%	-0.13%	2.8
Marche	1359	2243	1541	0.79%	-0.21%	25.7
Lazio	432	925	1455	0.13%	-0.55%	10.6
Abruzzo	197	324	300	0.04%	-0.36%	3.7
Molise	407	479	428	0.12%	-0.01%	5.5
Campania	420	953	1212	0.13%	-0.41%	10.9
Puglia	6854	7895	8016	3.04%	-0.72%	90.4
Basilicata	61	78	104	0.02%	-0.06%	0.9
Calabria	168	223	369	0.04%	-0.12%	2.6
Sicilia	620	973	1764	0.43%	-0.94%	11.1
Sardegna	221	394	668	0.05%	-0.45%	4.5
Bolzano	125	446	241	0.01%	-0.35%	5.1
Trento	268	899	482	0.04%	-0.34%	10.3
Italy total	20346	35418	32100	12.85%	-10.21%	405.5

Table 1: Results on vineyards suitable for controlled origin wine

Region	Farms	# Records	Changes	Added Area	Subtracted Area	Time (sec.)
Piemonte	528	866	1961	0.05%	-0.28%	9.9
Valle d'Aosta	168	254	173	0.01%	-0.01%	2.9
Lombardia	2312	3761	3172	0.38%	-0.25%	43.1
Veneto	1346	3811	2593	0.33%	-1.05%	43.7
Friuli-Venezia Giulia	1489	2931	2223	0.19%	-0.19%	33.6
Liguria	868	1465	1321	0.06%	-0.04%	16.8
Emilia-Romagna	436	918	814	0.10%	-0.62%	10.5
Toscana	4558	11481	6747	0.83%	-4.88%	131.6
Umbria	2200	4924	2466	0.18%	-0.17%	56.4
Marche	3005	5850	3472	0.33%	-0.09%	67.0
Lazio	4333	7363	7768	0.43%	-0.80%	84.4
Abruzzo	5392	10664	5673	0.71%	-0.94%	122.2
Molise	1732	3392	1841	0.24%	-0.05%	38.9
Campania	10904	18092	13728	0.92%	-0.67%	207.3
Puglia	10897	15251	13383	2.86%	-1.48%	174.8
Basilicata	370	486	659	0.04%	-0.15%	5.6
Calabria	2455	3391	4513	0.41%	-0.89%	38.9
Sicilia	4224	7497	9630	2.00%	-3.54%	85.9
Sardegna	642	1385	2528	0.07%	-0.72%	15.9
Bolzano	11	21	19	0.00%	-0.01%	0.2
Trento	1936	2400	1977	0.13%	-0.05%	27.5
Italy total	59806	106203	86661	10.25%	-16.87%	1217.0

Table 2: Results on vineyards not suitable for controlled origin wine

As for the first aspect, data obtained by the procedure were able to satisfy the balancing conditions in the totality of the cases (100%). As for the second aspect, a positive feature for a general information reconstruction procedure is satisfying the requirements while not changing the data exceedingly. In the analyzed cases, in addition to the theoretical guarantee that the number of changes is minimal, we observe that the amount of the variations is always a small percentage. This means that the procedure was able to reconstruct information without distorting the data.

Region	Farms	# Records	Changes	Added Area	Subtracted Area	Time (sec.)
Piemonte	624	1682	1191	0.037%	-0.016%	19.2
Valle d'Aosta	84	221	157	0.000%	0.000%	2.5
Lombardia	1436	3784	2162	0.030%	-0.069%	43.3
Veneto	5715	13604	11033	0.383%	-0.670%	155.8
Friuli-Venezia Giulia	734	1955	1123	0.010%	-0.002%	22.4
Liguria	264	645	458	0.001%	0.000%	7.3
Emilia-Romagna	252	742	976	0.168%	-0.042%	8.5
Toscana	3006	6571	6274	0.486%	-0.319%	75.2
Umbria	513	1160	931	0.007%	-0.003%	13.2
Marche	2210	5347	4615	0.229%	-0.088%	61.2
Lazio	912	2023	1553	0.009%	-0.002%	23.1
Abruzzo	1960	4540	2650	0.038%	-0.103%	52.0
Molise	2006	4649	3817	0.047%	-0.035%	53.2
Campania	2854	6538	4403	0.015%	-0.003%	74.9
Puglia	18205	41924	34881	0.561%	-0.527%	480.3
Basilicata	433	996	710	0.011%	-0.005%	11.4
Calabria	506	1166	851	0.007%	-0.008%	13.3
Sicilia	1117	2546	1772	0.035%	-0.002%	29.1
Sardegna	223	512	407	0.006%	-0.004%	5.8
Bolzano	23	96	83	0.012%	-0.037%	1.1
Trento	889	2280	1344	0.005%	-0.003%	26.1
Italy total	43966	102981	81391	2.098%	-1.938%	1180.0

Table 3: Results on other cultivations

Region	Farms	# Records	Changes	Added Heads	Subtracted Heads	Time (sec.)
Piemonte	704	947	719	0.317%	-0.107%	9.7
Valle d'Aosta	38	51	38	0.000%	0.000%	0.5
Lombardia	528	784	574	0.390%	-0.130%	8.0
Veneto	3797	4449	4489	7.348%	-2.763%	45.4
Friuli-Venezia Giulia	164	210	169	0.017%	-0.285%	2.1
Liguria	71	92	71	0.000%	0.000%	0.9
Emilia-Romagna	333	593	500	5.585%	-8.403%	6.1
Toscana	1554	1878	1973	0.265%	-0.037%	19.2
Umbria	124	160	124	0.000%	-0.554%	1.6
Marche	1171	1488	1592	0.823%	-0.428%	15.2
Lazio	302	380	305	0.000%	0.000%	3.9
Abruzzo	215	286	217	1.634%	0.000%	2.9
Molise	1003	1303	1281	0.437%	-0.477%	13.3
Campania	387	471	394	0.001%	0.000%	4.8
Puglia	3734	4144	4556	0.595%	-0.570%	42.3
Basilicata	123	173	126	0.001%	0.000%	1.8
Calabria	266	320	270	0.000%	0.000%	3.3
Sicilia	338	441	340	0.000%	0.000%	4.5
Sardegna	276	489	281	0.004%	0.000%	5.0
Bolzano	127	181	127	0.000%	0.000%	1.8
Trento	97	111	97	0.000%	0.000%	1.1
Italy total	15352	18951	18243	17.418%	-13.754%	193.5

Table 4: Results on livestock

The accuracy of the reconstructed information has been further evaluated by setting up a specific experiment. A large dataset of 274687 records representing all vineyards obtained from about 126000 farms, all exact, were perturbed by introducing random errors with uniform distribution at 3 different intensities, so that respectively about 1%, 5% and 10% of the microdata values have been changed. This was performed 20 times, in order to obtain statistically significant results, so 60 different large erroneous datasets were obtained. After this, the reconstruction procedure was applied to all of them, and the 60 obtained (corrected) data sets were compared to the original exact one.

Statistical indicators commonly used for measuring the differences between real and predicted values, such as the Relative Root Mean Square Error (RRMSR), are practically 0 ($< 10^{-5}$) for all the corrected datasets. This means that the quality of the reconstruction is fully satisfactory. However, in order to obtain more insight, we analyzed the reconstruction at an even greater detail: we compared each single reconstructed value to its original value, and checked whether it was exactly identical or not. Note that such test is extremely strict, probably beyond the requirements of a similar reconstruction process. The results are presented in Table 5. The percentage of reconstructed values that are exactly equal to the original values has been computed by subdividing the datasets on the basis of the number of errors actually introduced in each farm. Clearly, those percentages lower when the number of errors introduced in the farm increases, but accuracy is anyway extremely high. Even when the farm data contain a considerable number of errors (from 4 to 10, that is often more than what happens in usual practice), the reconstructed values are exactly equal to the original ones in a very high percentage of the cases.

Errors per Farm	Percentage of Exactly Reconstructed Values		
	Perturbation at 1%	Perturbation at 5%	Perturbation at 10%
1	99.9%	99.9%	99.9%
2	98.1%	98.2%	98.2%
3	86.0%	86.6%	83.4%
4 ÷10	81.8%	56.3%	49.1%

Table 5: Accuracy of the reconstruction process

4 Conclusions

Information Reconstruction is a crucial task in the case of large surveys, such as a Census of Agriculture, as well as for other applications of database processing. A typical problem arising in the described Census consists in checking, and correcting when needed, the areas declared by each farm for each cultivation. This type of balancing problem is extremely important and has a great economical relevance. Moreover, in contexts different from the Census, the possibility of reconstructing exact values could be useful for counteracting opportunistic behaviors, e.g. willingly erroneous declarations for influencing resources allocation or production plans.

Similar problems could be formulated in different manners. This particular Census problem has very specific aims and requirements, and it was deemed that they were better represented by the proposed mixed integer linear model (11). The procedure has been tested in the case of the Italian Census of Agriculture 2010 with specific respect to the 4 most important cases. Clearly, the proposed class of models is not limited to the case of an Agricultural Census, but can be used for other problems sharing the same characteristics, in particular the presence of balance requirements and minimum change objective. Results are very encouraging both from the computational and from the data quality point of view. The sequence of arisen mixed integer problems can be solved to optimality by using a state-of-the-art implementation of branch-and-cut procedures. Each single model is solved in extremely short times. In the totality of the cases the reconstructed information was able to satisfy the balancing conditions without excessively distorting the data, as resulted from the analysis of the variations introduced in the whole datasets. Moreover, a specific experiment proves that the reconstructed information was exactly equal to the original uncorrupted one in an exceedingly high percentage of the cases.

References

1. Bacharach, M.: Matrix rounding problems. *Management Science* **12**(9), 732-742 (1966)
2. Bacharach, M.: *Biproportional Matrices and Input-Output Change*. Cambridge University Press, Cambridge, UK (1970)
3. Banff Support Team: Functional Description of the Banff System for Edit and Imputation System. *Quality Assurance and Generalized Systems* Section Tech. Rep. Statistics Canada (2003)
4. Bankier, M.: Canadian Census Minimum change Donor imputation methodology. In *Proceedings of the Workshop on Data Editing*, UN/ECE, Cardiff, United Kingdom (2000).
5. Bertsimas, D. and Tsitsiklis, J.N.: *Introduction to Linear Optimization.* Athena Scientific, Belmont, Massachusetts (1997)
6. Bomze, I.M., Locatelli, M.: Separable standard quadratic optimization problems. *Optimization Letters* **6**(5), 857–866 (2012)
7. Bourbaki, N.: *Topological vector spaces*. Springer-Verlag, Berlin, Germany (1987)
8. Boyd, S. and Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge, UK (2004)
9. Bruni, R.: Discrete Models for Data Imputation. *Discrete Applied Mathematics* **144**(1), 59-69 (2004)
10. Bruni, R.: Error Correction for Massive Data Sets. *Optimization Methods and Software* **20**(2-3), 295-314 (2005)
11. Bruni, R. and Bianchi, G.: A Formal Procedure for Finding Contradictions into a Set of Rules. *Applied Mathematical Sciences* **6**(126), 6253-6271 (2012)
12. De Waal, T.: Computational Results with Various Error Localization Algorithms. *UNECE Statistical Data Editing Work Session*, Madrid, Spain (2003)
13. De Waal, T., Pannekoek, J., Scholtus, S.: Handbook of Statistical Data Editing and Imputation. Wiley Handbooks in Survey Methodology, John Wiley & Sons, Inc.: New York, NY (2011)
14. European Council Regulation (EEC) No 357/79 of 5 February 1979 on statistical surveys, EEC Documentation (1979)
15. Fellegi, I.P. and Holt, D.: A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* **71**, 17-35 (1976)
16. Fuller, W.A.: *Measurement Error Models*. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.: New York, NY (2006)
17. Garey, M.R. and Johnson, D.S.: *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Co, San Francisco, CA (1979)
18. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, NY (2001)
19. Hochbaum, D.S. and Shanthikumar, J.G.: Convex separable optimization is not much harder than linear optimization. *Journal of the ACM* **37**(4), 843-862 (1990)
20. IBM: *Ilog Concert Technology 12.1 Reference Manual.* International Business Machines Corporation (2009)
21. IBM: *Ilog Cplex 12.1 Reference Manual.* International Business Machines Corporation (2009)

22. Kalantari, B., Lari, I., Ricca, F., Simeone, B.: On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. Mathematical Programming, Ser. A **112**, 371–401 (2008)
23. Klösgen, W. and Żytkow, J.M. (Eds.): *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press: Oxford, UK (2002)
24. Lyberg L.E., Biemer P., Collins M., De Leeuw E.D., Dippo C., Schwarz N., Trewin D. (Eds.): *Survey Measurement and Process Quality*, *Section C, post survey processing and operations*. John Wiley & Sons, Inc.: New York, NY (1997)
25. Mucherino, A., Papajorgji, P., Pardalos, P.M.: *Data Mining in Agriculture*, Springer: New York, NY (2009)
26. Nemhauser, G.L. and Wolsey, L.A.: *Integer and Combinatorial Optimization.* John Wiley & Sons, Inc.: New York, NY (1999)
27. Riera-Ledesma, J. and Salazar-Gonzalez, J.J.: New Algorithms for the Editing and Imputation Problem. *UNECE Statistical Data Editing Work Session*, Madrid, Spain (2003)
28. Ramakrishnan, R. and Gehrke, J.: *Database Management Systems* (3rd edition). McGraw-Hill: New York, NY (2003)
29. Schneider, M.H. and Zenios, S.A.: A comparative study of algorithms for matrix balancing. *Operations Research*, **38**(3), 439-455 (1990)
30. Schrijver, A.: *Combinatorial Optimization*. Springer, Berlin; New York (2003)
31. Winkler, W.E.: State of Statistical Data Editing and current Research Problems. In *Proceedings of the Workshop on Data Editing*, UN/ECE, Rome, Italy (1999)