Baltic J. Modern Computing, Vol. 10 (2022), No. 4, pp. 623–644
https://doi.org/10.22364/bjmc.2022.10.4.03

# An Evaluation of Machine Learning Approaches to Integrate Historical Farm Data

Philippe BELMONT GUERRÓN, Maria HALLO

Facultad de Ingeniería de Sistemas, Ladrón de Guevara E11 · 253 Edif 20, Escuela Politécnica Nacional, Quito, Ecuador

philippe.belmont@epn.edu.ec, maria.hallo@epn.edu.ec

**Abstract.** Large datasets in agriculture are increasingly available through yearly surveys. However, very few longitudinal datasets providing insights for farmer's decision are available. The main objective in this research is to match farm establishments. The purposes of this investigation is two fold: first, *to match successive yearly surveys*, producing longitudinal records into farm history; and second to use only *categorical and numerical features* to match records. We analyzed Ecuadorian national agricultural surveys from the years 2010 to 2012. In total, 125098 records were compared, using 16 different algorithms. Our results suggest that with this particular data setup, unsupervised methods using a stochastic matching approach outperform other algorithms in terms of F1 scores. Matching individuals over three consecutive years shows that ensemble techniques allowed the re-identification of 60% of individuals. In the context of Ecuador, no data are available to follow individual farms over time, longitudinal datasets could provide essential insights for local policies.

**Keywords:** Data matching, Record Linkage, Farm Matching, Machine Learning, Entity Resolution, Data Integration

## 1 Introduction

Agriculture today face a difficult challenge: the world's growing population continuously drives the need for greater food productivity while at the same time, modern agriculture threatens the environment, contributing significantly to greenhouse gas emissions and climate change (McIntyre, 2008). To this day, small-scale family farms with less than 20 hectares, play an essential role to the global food supply in middle-income countries (Woodhill et al., 2020). In this context, it is urgent to understand how efficient and sustainable agriculture can provide diverse and quality food for an ever-increasing population. Since the creation of the Food and Agriculture Organization of the United Nations (FAO) in

1945, standardized methodologies have been developed to enhance agricultural information systems worldwide.

One key component of national statistics for agricultural systems are surveys and census. In many developing countries such as Ecuador, these are often the only source of national information, yet only a few efforts for integration of yearly records have been made and mainly for health data (Kazanjian, 1998; Jarvis et al., 2017; Reppermund et al., 2019; Rowlands et al., 2021; FAO, 2015). Agricultural surveys provide complete descriptions of land ownership and farm characteristics, systematically reporting land use on a parcel level (Remans et al., 2019). Surveys usually cover small areas of geographical sampling units. Those sampling units are not modified from year to year, and with few exceptions (non-response, drop out) the same farms are surveyed in consecutive years.

These conditions should be ideal for record linkage, but in practice no identifiers and very few of the farmers' personal information are provided. On a national scale, matching datasets by hand is prohibitive, but integrating them can be done using probabilistic methods (Contiero et al., 2005). Modern Machine learning techniques offer new and efficient ways for managing large amounts of data. This is especially advantageous when the quantity of observations is important, which is the case with agricultural surveys.

In the context of Ecuador, where small-scale farming prevails, very few sources of national data exist. Agricultural statistics often exist only in isolation, and are usually poorly-shared and understood by other agencies. Even when the necessity has long been identified (Hill, 1996), public institutions fail to recognize that rural economies are intrinsically diverse across farms. Without integration between datasets and the definition of common identifiers, little effort is made to support analytical applications in developing countries. Therefore, it is essential to understand farmers' practices and drivers susceptible to affect production over time (O'Donoghue et al., 2017).

The main objective of this study is to adapt previous work related to agricultural data matching (Winkler, 1995; Aiken et al., 2019) to the context of yearly surveys. The originality of this research is dual: first, matching successive yearly datasets, with the aim to produce longitudinal records of farm history; and secondly, match records using only numeric features, an uncommon case in data matching where textual descriptors are usually employed (producer and farm names, addresses).

We applied various matching procedures to successive yearly surveys. We used public datasets from the Ecuadorian National Statistical Institute: the Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC: Agricultural land use and continuous production survey) from the years 2010 to 2012. Little to no variation in survey design occurred during those years, representing a rich source of information for agricultural policies (Guillermo Otañez, 2004). We compared 125098 records from three datasets, and the results were evaluated over three pairs of datasets, using two different sets of variables and 16 algorithms leading to 96 matching trials.

Records did not include names or address, nor consistent identifiers of farm

households. We explored using numerical features such as production characteristics, crop area, level of production and sampling features as pseudo-identifiers. These variables are subject to yearly variations as farms evolve in time, for instance by acquisition or session of land, or simply change in land use. The matching algorithm should provide robust matching results in spite of these variations of farm activities. A wide variety of record linkage methods were evaluated, including probabilistic methods and a different set of unsupervised and supervised machine learning techniques. For each algorithm, calculations were repeated over various pairs of datasets, which allowed us to evaluate "out of the sample" and generalization quality.
In the next section this article provides an overview of record linkage standard procedure; describing the necessary steps and emphasizing evaluation metrics. The following section details the "data setup": context, preprocessing and selection of pseudo-identifiers, and a short description of the matching methods that were applied to the datasets. In the last section, results are reported and discussed in regards to their implications for agricultural statistical systems.

## 2 Related works

### 2.1 Record linkage

Record linkage consists of merging datasets based on common entities. In this process, two records are compared. "Matches" are identified when two records are considered the same entity and "non-matches" in other cases, similar to a classification problem. The data setup usually involves two datasets with no unique identifiers (Winkler, 1994). Record linkage has applications in numerous domains: health records (Contiero et al., 2005; Karr et al., 2019), administrative surveys (Abowd et al., 2019; Enamorado et al., 2019) or research on historic census (Fu, Boot, Christen and Zhou, 2014; Fu, Christen and Zhou, 2014). Previous work with record linkage in the agricultural context has focused on analyzing national census, identifying duplicate entries (deduplication) (Winkler, 1995; Bellow et al., 2016) and integrating farm records to agro-industrial datasets (Aiken et al., 2019).
The procedure for record linkage involves four steps: data preparation, indexing or blocking, classification, and evaluation (see Figure 1). Preparation of data requires common attributes between datasets to be standardized. Typically string attributes are used, such as names or addresses, and numerical measures, such as date of birth. In this process various sources of errors may increase the difficulty of record linkage: the population between datasets may differ, pseudo-identifiers vary as a result of distinct data acquisition processes, and values may be missing or changing over time (Christen, 2012).
An optional step called "indexing" or "blocking" consists of dividing the data sets into smaller groups by using group identifiers, and producing pairs to compare only from these groups. This technique reduces the number of comparisons to evaluate and the computation time required to match pairs. Without this group comparison, the number of pairs for two datasets of size m increase quadratically

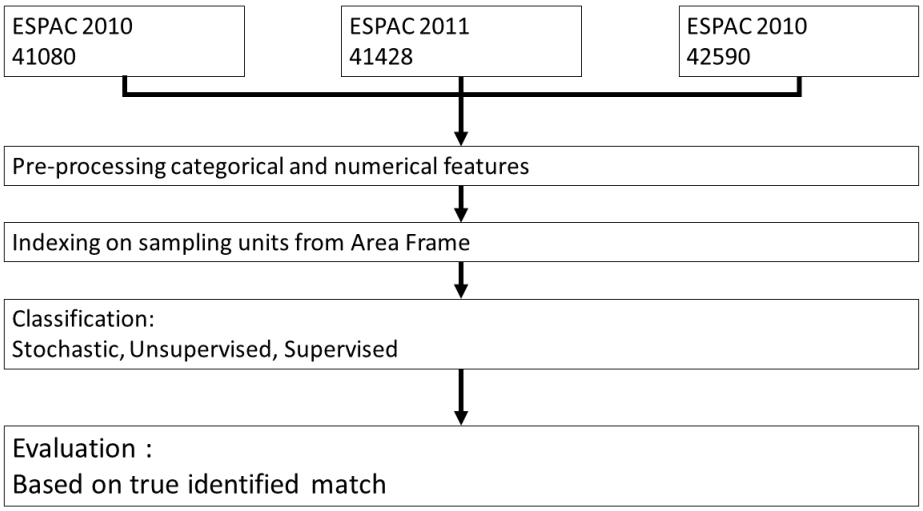


**Fig. 1.** Data process for Agricultural Survey record linkage.

(m squared). This step of fractionating pairs using a common key is especially relevant in our case where blocking can be related to sampling structure (Fellegi and Sunter, 1969). The surveys are conducted through yearly visits to geographical sampling units. In each unit, systematic sampling of all farms was carried out each year (Guillermo Otañez, 2008b); further details on blocking are provided in the data setup section.

The choice of classification algorithms may produce widely different results depending on the comparison function and selected variables to compare. When comparing two records, the classification algorithm will receive a similarity vector based on the considered attributes, and label it as match or non-match. As this process is realized using blocking, and optimizing execution of methods lead to run times not exceeding a few hours, computational efficiency will not be assessed. The evaluation step is equivalent to the evaluation of a binary classifier. Results can be summarized in a contingency table, comparing true match status to predicted outcome. Here, the first entry indicates reference true match or non-match for any given pair and classifiers output as shown in Table 1.

Predicted status is based on a similarity threshold, below which a pair of records are considered to be a non-match. When comparing methods, evaluating algorithms in terms of quality of classification is not trivial. Indeed, methods can produce different sets of pairs, with different distance metrics produced when comparing results. The similarity value from one method may not be related to another and could produce misleading comparisons. Comparing different algorithms with the same threshold should be avoided (Hand and Christen, 2018). Another problem arises from the fact that record linkage produces a strong

**Table 1.** Contingency table used in record linkage

| Predicted class: | | Match | Non-match |
|---|---|---|---|
| True status: | Match | True Positives (true match: TM) | False Negatives (false non-match: FNM) |
| | Non-match | False Positive (false match: FM) | True Negatives (true non-match: TNM) |

imbalance between the number of non-match and matches. Consequently, high accuracy may arise from a classifier that only predicts non-match without match identified (Belin and Rubin, 1995).
As imbalance between match and non-match is usually very high, F-score statistics (harmonic mean of precision and sensitivity) is preferred in record linkage (Hand and Christen, 2018). We propose here to employ an evaluation method proposed by Hand and Christen (2018) that overcomes the limitation of comparing thresholds, providing that for given a value of F-score, the same number of predicted matches are compared. The main idea is to rewrite F-score as a weighted mean of precision and recall:

$$F = \frac{2}{P^{-1} + R^{-1}} = \frac{2P * R}{P + R} = \frac{2TM}{FNM + FM + 2TM} \tag{1}$$

and rewrite F-score as:

$$F = p * R + (1 - p) * P \tag{2}$$

where:

$$p = \frac{(FNM + TM)}{(FNM + FM + 2TM)} \tag{3}$$

Comparing F and p, the weights p could inform on the *relative importance* given to precision and recall. Using p in relation to F to evaluate algorithms produce a fair comparison as the same number of predicted matches are compared. We can use this metric p to graphically compare various algorithms without the use of thresholds.
Finally, to evaluate matched databases retaining only predicted pairs, a final step called deduplication eliminates multiple occurrences of records (Murray, 2016). In fact, farms are uniquely defined in each dataset: only one match per record should occur between two given datasets. This additional step was added to produce a single match per observation, using linear assignment, as proposed by Jaro Jaro (1989); Enamorado et al. (2019). The final result is a longitudinal dataset for multiple year linkage.

## 3 Data setup

Each record linkage exercise is adapted to the nature and availability of information. In our case, true match status is available, allowing us to compare results on a common reference. This true status has been obtained in previous work, obtaining personal data from each dataset. In this part, we will first describe the context of the agricultural survey and the comparison that has been performed. Then the preparation and selection of variables is presented. Finally, a description of the algorithms is provided and, how evaluation was performed.

### 3.1 Context

The continuous production and area surveys (ESPAC) are yearly national surveys, which describe land use, crops, forestry, labor, and breeding activities in roughly 41,700 observations each year. Three consecutive years were selected: 2010-2012 (see Figure 1). To account for variation between years, we evaluated all three combinations between 2010, 2011, and 2012 datasets.
The goal was to produce a longitudinal dataset for the three selected years. True match status was established in a previous work (Belmont 2019, unpublished) using farmer and farm name and address.

Overlapping observations between years is estimated over 80%, enhancing the potential performance of the linkage algorithm, as noted by (Enamorado et al., 2019). The remaining observations are missing, possibly due to non-response or when no agricultural activities were registered on farms. Surveys are usually based on a multiple frame (Davies, 2009). The sampling design is consituted by an area frame and a list frame. The area frame is divided into standardized strata, based upon the predominant land-use in the sampling units: pasture, temporary and annual crops, forest and natural vegetation, and an additional stratum for Amazonian region. The list frame is a list of the main farms in extention or production in a specific sector. The two sampling frames may overlap, 1 or 2% of units, and usually require deduplication (Winkler, 1994); this aspect of the process is not taken into consideration in this research, as duplicates can be identified.
As mentioned above, record linkage was performed between the three pairs of annual datasets for each algorithm: 2010-2011, 2010-2012, and 2011-2012. Processing steps are described in Figure 1: identification of features, indexing, clasification and evaluation of linkage quality. Corresponding weights were normalized using minmax (0,1). The analysis was performed using R programming language (version 4.0.2) and data sets are available online (here). Original datasets have thousands of variables; in the following section we describe the data preprocessing and selection methods that we applied.

### 3.2 Pre-processing and attribute selection

Identifying a set of farm features for record linkage is especially challenging in Ecuador. Small scale farming activities vary significantly over time, as farmers

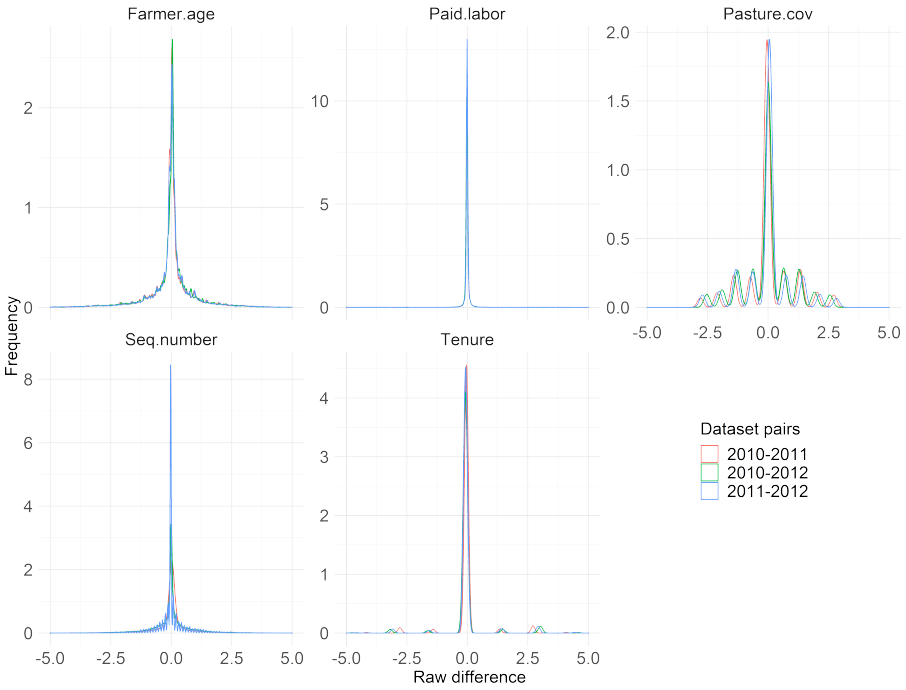


**Fig. 2.** True matches raw difference in value for selected variables: farmers age (Farmer.age), number of paid workers (Paid.labor), pasture coverage (Pasture.cov), land tenure (Tenure), sequential survey number (Seq.number).

employ different adaptive strategies (Chen et al., 2018).

In the absence of string identifiers, farms are described numerically. The inclusion of continuous variables provoked a considerable increase in computation time when testing different algorithms. Converting numerical variables to categorical variables and using quantiles and normalization within blocks helped to reduce variance among identified pairs and reduce computational time. We evaluated consistency of variables based on the true match subset (see Figure 2) in order to confirm that these remained constant or very similar over time. For instance, the number of parcels between years is centered on zero but annual change appears when farmers merge or divide parcels, or as land is transfered, acquired or leased. A subset of 13 descriptors was defined, selecting variables with lower contribution to variance using principal component analysis.

Common characteristics representing consistency over time include: (i) category of farm ownership: (privately owned, rented); (ii) farm category: subsistence farms which generate no income by selling products, family farms, "capitalist" farms, business farms, and haciendas, where no family member works on the farm, (iii) farmer's age and (iv) sex, (v) labor on farm, (vi) cattle density citep(see:Alkemade 2013), (vii) average milk production per cow, (viii) presence

of horses, donkeys or both; (ix) farm size (as quintile of land distribution), (x) pasture and (xi) irrigation cover (as percentage of total land), and (xii) forest cover (as quintile of the percentage of total land) and number of parcels (xiii). A second reduced set was built to test survey design variables: two "agronomic" variables: land tenure and farm size as quintile of land distribution, survey ponderation factors assigned to each farm with very few adjustments from year to year; finally, a sequential survey number allows for partial identification. Sequential survey numbers correspond to a determined sequence of farms that each survey taker has to follow. Each year, the sequence begins with the same first farm and proceeds in the same order as previous years (Guillermo Otañez, 2008a).

In table 2, we report descriptive statistics of retained variables. Valid and missing data in percentage is described, and number of categories in the left column and mean value right column are reported. Quantity of missing value did not exceed 11.2 %, which is within the range of effective record linkage as tested by Enamorado et al. (2019). Those attributes are common descriptors in standard agricultural surveys (Winkler, 1995).

### 3.3 Indexing

As mentioned above, by using group identifiers, we can reduce the number of comparisons to evaluate. For instance, individuals are paired together using a region code as a common identifier. One individual from a group in the first dataset is compared to the others from the same group in the second dataset, and not from the whole dataset, thus reducing computation of pairs.
Agricultural surveys utilize a rotation scheme of sampling units to avoid repeated interviews and response burden. Here, consistency of the sampling design over years, with no sampling rotation, allowed us to define consistent blocking indexes. In this dataset, only small modifications were made in order to adapt to new political divisions which occurred in 2009, but the majority of sampling units remained constant. This sampling design would assure a high overlapping between years. Sampling units contain 16 farms on average and up to 92 in most populated areas. The total of pairs computed reached 1,036,906 pairs for 2010-2011, 1,021,664 pairs for 2011-2012, and 1,003,527 pairs for 2010-2012. For each pair, a distance metric is computed using a classification algorithm. The next section describes the algorithms employed.

### 3.4 Classification Algorithms

In this section we summarize the types of algorithms employed for classification. For a given set of pairs, a distance metric, or weight is assigned to a pair. The value of the weights indicates if the pair is a match or a non-match, based upon a defined threshold. We selected a wide range of classification algorithms to provide an overview of the best-performing algorithms for this task using existing methods (Aiken et al., 2019). Different methods can produce very different results.

**Table 2.** Variable characteristics

| Categorical attributes: | | Valid | Missing (%) | Categories |
|---|---|---|---|---|
| | 2010 | 41081 | - | 5 |
| Farm type | 2011 | 41428 | - | 5 |
| | 2012 | 42590 | - | 5 |
| | 2010 | 41081 | - | 2 |
| Farm ownership: (proper, rent...) | 2011 | 41428 | - | 2 |
| | 2012 | 42590 | - | 2 |
| | 2010 | 41081 | - | 2 |
| Farmer sex | 2011 | 41428 | - | 2 |
| | 2012 | 42590 | - | 2 |
| | 2010 | 41081 | - | 5 |
| Land extension | 2011 | 41428 | - | 5 |
| | 2012 | 42590 | - | 5 |
| | 2010 | 41081 | - | 5 |
| Pasture extension quantile | 2011 | 41428 | - | 5 |
| | 2012 | 42590 | - | 5 |
| | 2010 | 41081 | - | 6 |
| Forest extension quantile | 2011 | 41428 | - | 6 |
| | 2012 | 42590 | - | 6 |
| | 2010 | 41081 | - | 3 |
| Presence/Absence of horses and/or donkeys | 2011 | 41428 | - | 3 |
| | 2012 | 42590 | - | 3 |
| Numerical attributes: | Year: | Valid | Missing (%) | Mean |
| Age (birth year) | 2010 | 40218 | 863(2.1%) | 1967 |
| | 2011 | 40887 | 541(1.3%) | 1968 |
| | 2012 | 42050 | 540(1.3%) | 1969 |
| Paid Labor (persons) | 2010 | 41081 | - | 2.06 |
| | 2011 | 41428 | - | 1.9 |
| | 2012 | 42590 | - | 1.85 |
| Area under irrigation (%) | 2010 | 41081 | - | 0.09 |
| | 2011 | 41428 | - | 0.1 |
| | 2012 | 42590 | - | 0.08 |
| Number of parcels | 2010 | 41081 | - | 1.49 |
| | 2011 | 41427 | 1(0%) | 1.5 |
| | 2012 | 42590 | - | 1.53 |
| Animal density | 2010 | 41081 | - | 0.36 |
| (cattle unit / | 2011 | 41428 | - | 0.37 |
| hectare) | 2012 | 42590 | - | 0.34 |
| Average milk | 2010 | 38703 | 2378(5.8%) | 1.28 |
| production | 2011 | 39033 | 2395(5.8%) | 1.34 |
| (liters/animal/day) | 2012 | 37803 | 4787(11.2%) | 1.21 |
| Ponderation factor | 2010 | 41081 | - | 0.81 |
| | 2011 | 41428 | - | 0.81 |
| | 2012 | 42590 | - | 0.8 |

Before implementing classification methods, we use deterministic matching as baseline. Deterministic matching evaluates a pair given the assumption that all fields are equal. Matching methods can be categorized into two general families: Stochastic or probabilistic matching and "Machine learning" methods. In total, 16 methods were applied, leading to 96 evaluations including variation in parameters, and each producing weight output for the three paired datasets. Probabilistic approaches included: propensity score matching adapted to the context of data linkage, Epilink method, Stochastic matching approaches using an expectation–maximization algorithm with the Fullegi Sunter model, and a scaling algorithm.
Machine learning algorithms consisted of unsupervised methods: clustering (Fuzzy C-Means), and supervised methods: Artificial neural network, Recursive partition tree, Bagging decision tree, and Adaboost. In the case of Machine learning, models are required to handle highly skewed predicted targets. Indeed, a paired dataset is mostly composed of non-matched pairs with a very low quantity of match (Johnson and Khoshgoftaar, 2019). This data imbalance may highly decrease the capacity of the machine learning techniques to identify matches among training datasets (Pixton and Giraud-Carrier, 2006).

**3.4.1 Probabilistic and stochastic methods** In the following section, we describe methods using a probabilistic approach to matching. The main idea is to infer the distribution of distances between pairs to compute weights. These methods are most commonly used for record linkage.

*Stochastic record linkage* Stochastic record linkage makes use of the Expectation-Maximization algorithm. The Fellegi-Sunter model is commonly employed in record linkage, and a short description of the procedure is given, for more details see (Pixton and Giraud-Carrier, 2006; Christen, 2012). This method was computed using the R package Recordlinkage, with the emWeights procedure. This procedure is based on a decision model, assigning a probabilistic weighting for pairs of records (Sariyar and Borg, 2010). For a collection of potential pairs, comparison patterns are computed, then conditional probabilities over these patterns give a probability of belonging to a set of matches or a set of non-matches. We aim to merge two sets A1 and A2 of size N1 and N2 respectively, using a set X of common variables. In a sample size of N1*N2 pairs, a comparison vector noted $\gamma_x(i, j)$ is defined with the pair of the ith observation in A1 and jth observation in A2. This vector represents the level of within-pair similarity for the xth variable between the ith and jth observations, of datasets A1 and A2 respectively. As noted in Enamorado et al. (2019), corresponding elements of the comparison vector can be set according to Lx similarity levels for the xth

variable:

$$\gamma_x(i,j) = \begin{Bmatrix} 0 \\ 1 \\ \vdots \\ L_x - 2 \\ L_x - 1 \end{Bmatrix} \begin{matrix} Different \\ Similar \\ \\ Identical \end{matrix} \tag{4}$$

Thus, the conditional probability of the match status M is denoted:

$$m(\gamma_{ij}) = P(\gamma_{ij}|M = 1) \wedge u(\gamma_{ij}) = P(\gamma_{ij}|M = 0) \tag{5}$$

Where M take the value 0 for non-match with a probability $m(\gamma_{ij})$ and 1 for match with a probability $u(\gamma_{ij})$.
Finally, under the Fellegi-Sunter model weights w are computed according to:

$$w_{\gamma_{ij}} = log(\frac{m(\gamma_{ij})}{u(\gamma_{ij})}) \tag{6}$$

and used to define linkage rules distinguishing between match and non-match. The Fellegi-Sunter model has various limitations: the assumption of independence of matching variables and the treatment of missing values (Abowd et al., 2019). An extension of the Fellegi-Sunter model (see: FastLink (Enamorado et al., 2019)) proposes a different approach, relaxing the assumption of independence of matching variables. The treatment of missing data is essential and, in absence of imputation, data is usually treated as disagreement. Here, the canonical model assumes that data is missing at random conditional to the variables M (see equation 5: on conditional probability of the match above). The Fastlink algorithm has shown an important increase in computational efficiency and overall performance.

*Other probabilistic methods* We adapted Propensity Score Matching (PSM) to record linkage. For statistical matching, the PSM algorithm consists of pairing two observations according to a score (Ho et al., 2011). The approach uses a logit model to estimate a dependent variable taking a binary value in datasets to match 0 in the first dataset, 1 in the second (Rässler, 2012). The predicted probability or the propensity score, is a conditional probability of belonging to a dataset, given a set of variables. Based on the nearest propensity score, each observation is given a "donor unit", in this case using nearest neighbor matching. An R implementation of PSM, MatchIt was employed (see: Ho et al. (2011). A similar procedure called Epilink, using another distance metric between pairs (Contiero et al., 2005) was also evaluated (see R package: Recordlinkage, epiClassify procedure).
We also evaluated the Scaling procedure, another approach providing no explicit assumption of statistical independence, based on correspondence analysis (described in Healy and Goldstein (1976). This method allows for identification of most discriminatory identifiers based upon the minimization of a loss function (Goldstein et al., 2017). The R implementation Scalelink was employed.

**3.4.2 Machine learning Methods** We make a distinction between Machine learning methods and probabilistic ones as the design of the machine learning methods employed here are not predicting a probability distribution over a set of classes. These methods produce a likelihood of an observation to belong to a certain class. As mentioned above, supervised methods can only be evaluated using true match as training data.
Supervised classifiers were implemented using a labeled pair as training, and applying the trained model to the remaining pair dataset, ensuring that no paired records were shared between training and testing datasets. For instance, a model trained over 2011-2012 pairs was tested on 2010-2012 pairs, trained model on pairs from 2010-2012 were tested on 2010-2011 pairs, and trained model on pairs from 2010-2011 were tested on 2011-2012 pairs.
Four methods are selected: (i) Recursive partitioning tree (Therneau, 1983), using rpart R-package, using anova as the splitting rule; (ii) artificial Neural Networks (Ripley and Hjort, 1996) using nnet R-package (decay = 5*10-4, maximum iteration off 300, Initial random weights = 0.1); (iii) bagging decision tree (Breiman, 1996), and (iv) adaptive boosting, using fastAdaboost, Freund and Schapire (1996) Adaboost.M1 algorithm. The last two methods are a linear combination of weak decision tree classifiers. Finally, an unsupervised machine learning method, using clustering as a classifier, was evaluated. Fuzzy C-means clustering was tested here, in particular because of computational efficiency and the flexibility of the method (Hartigan and Wong, 1979). An R implementation of this algorithm was used from package e1071 (cmeans).
For supervised methods, an additional matching exercise was implemented separately using subsets divided per sampling strata. As for the use of neural networks, a committee of networks was evaluated, using various combinations of pseudo-identifiers and using ensemble averaging.

### 3.5 Evaluation

In this step, two forms of evaluation are employed using F-score and p metric, as described in section 2. After identifying the best methods, we report the performance on deduplicated datasets. The first form is used to compare fairly distinct methods with the same complete set of pairs. In the second form, the deduplication step consists of removing multiple matches for the same observation, providing a matched dataset where one observation has one only match. In the latter, comparisons of methods are less accurate as they are based on different sets of pairs. We report these results in order to illustrate what can be obtained building a transversal dataset.
Once classification was completed, F-score and p (as described in section 2) were calculated for different thresholds after minmax scaling of obtained weights. Graphical observation of performance helps to fairly distinguish between methods.
The best methods were assessed as deduplicated results after threshold selection. This comparison, despite the subjectivity of threshold selection between methods, will help illustrate the application for multiple year linkage. Thresholds were

established programmatically, using extreme value theory (Sariyar et al., 2011), providing an illustration of the potential of building transversal datasets with yearly agricultural surveys.
For each pair of matched records, algorithm weights and predicted matches were evaluated. Each algorithm was evaluated over three pairs of datasets, with two sets of variables for 16 models, leading to a total of 96 evaluations.

**Table 3.** An Illustration of matched records over three consecutive years

| | Pairs of databases | | |
|---|---|---|---|
| Variable: | 2010-2011 | 2011-2012 | 2010-2012 |
| Identifier: | 996 | 42220 | 83914 |
| Year | 2010 | 2011 | 2012 |
| Name | Arcadio | Arcadio | Arcadio jose |
| Surname | Buendia | Buendia | Buendia |
| Farm Class | 3 | 3 | 3 |
| Sex | 0 | 0 | 0 |
| Year of birth | 1969 | 1969 | 1969 |
| Land tenure | 1 | 1 | 1 |
| Livestock density | 0 | 0.01 | 0 |
| Number of parcels | 1 | 2 | 1 |
| Extension area | 4 | 4 | 5 |
| Pasture | 1.5 | 1.3 | 1.5 |
| EM weight | 0.515 | 0.437 | 0.151 |
| EM predicted | 42220 | 83914 | 996 |

Table 3 shows reported results for matching over the three pairs of datasets. We obtained weights for each record to identify individuals without the use of name and surname as pseudo-identifiers. For each record, a matching weight was calculated and corresponding prediction of pairs was stored. In this example only the weights for fastLink EM algorithms are reported. Even for a single algorithm, between combinations of years, the value of weights varies significantly.

## 4 Results

In this section results are organized into three sections: (i) a graphical method using "fair metric" comparison for the whole dataset and per sampling stata; (ii) an evaluation of best performing algorithms after deduplication; (iii) results over combined datasets, to obtain transversal records between 2010 and 2012.

### 4.1 Algorithm Comparison: Graphical Methods

Here, methods are compared using the graphical method discussed in section 2: a common "similarity threshold" for a given number of matches was calculated.

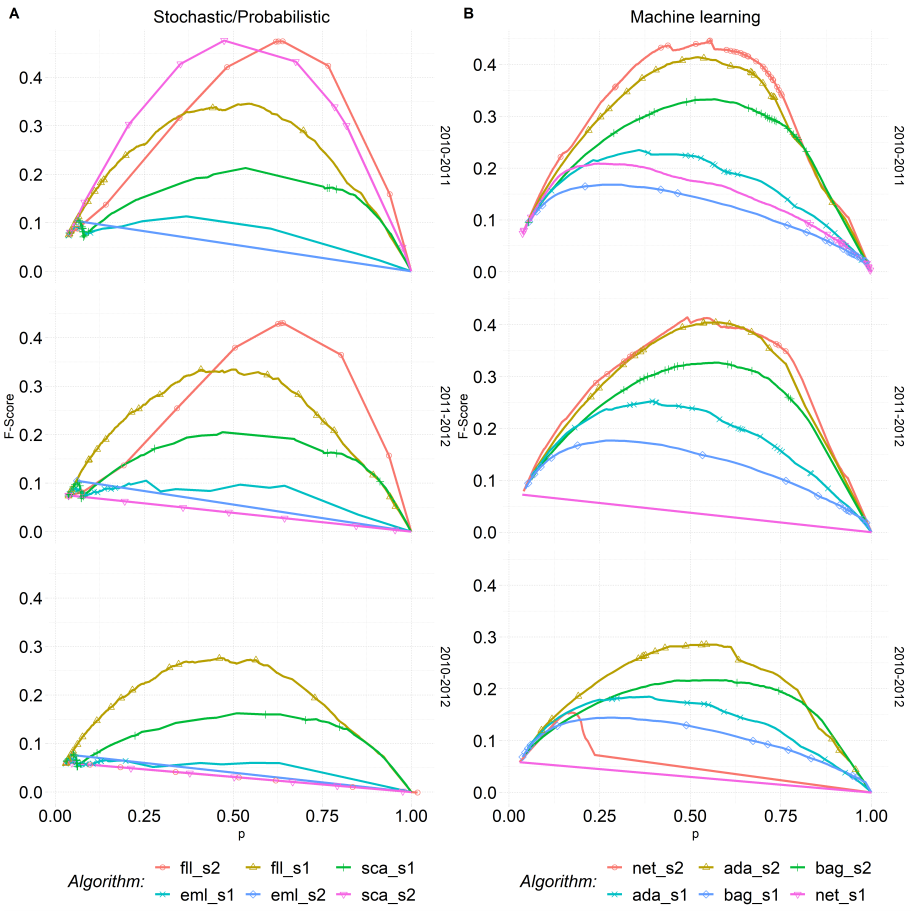


**Fig. 3.** F-score - p plots of the three paired datasets in row and by groups of algorithms in column (Probabilistic and Machine learning).

Then, using the "p" ratio of known true matches by the sum of predicted and known matches, results are shown graphically against the values of F-score (see Figure 3). In this figure, rows show evaluation of pairs formed from the three different pairs of datasets: the first row with 2010-2011, the second row showing 2011-2012, and the third row 2010-2012.

The plotted lines represent the best algorithms (higher F-scores for a given value of "p"). Each algorithm is plotted twice: with the first subset of 13 variables (suffix "_s1") and with the second with four variables (suffix "_s2"). The following probabilistic methods are plotted: sca: scaling method, fll: Fastlink : EM Fellegi-Sunter adaptation, ems: EM Fellegi-Sunter model. For machine learning methods the following methods were plotted: net: Artificial neural networks, ada: adaboost, bag: bagged clustering.

While using only agricultural characteristics, overall, the use of sampling design variables as pseudo-identifiers performed better in terms of F-score and observing performance in F-score p plots (line above perform better). This may suggest that among small farms, variability in characteristics is too high to be considered across years. Unsupervised learning methods, supervised method with recursive partition tree, Epilink and propensity score matching performed much worse than EM and scaling algorithm and are not reported in Figure 3. Globally, algorithms performed with similar performance compared to one another between the three assessed pair datasets (see Figure 3). As for 2010-2012 pairs, as expected, globally lower performance was recorded across methods, as variability due to farm change increased as expected, hindering linkage. Between the two groups of methods, supervised methods outperform unsupervised ones.
Among unsupervised methods, sampling design subset of variables produced highly variable results among years, with a high F-score on 2010-2011 dataset and very low F-score when comparing to 2010-2012 datasets. In comparison with agronomic variables, quality of matching remained stable over the years. The EM Fellegi Sunter algorithm consistently performed below the rest of the methods (Figure 3, "eml_s1", "eml_s2"), whereas the canonical model of Fastlink procedure performed best among evaluated methods (see Figure 3, "fll_s1", "fll_s2"), when recall weight p was near 0.53 with almost equal weight to precision (0.47). Scaling, producing a similar value for F-score (in Figure 3, "sca_s1").

Among supervised methods, using sampling design variables consistently outperformed agronomical variables; all supervised algorithms remained consistent in prediction performance through tested yearly dataset pairs. Adaptive boosting and Bagged clustering attained low F-score (0.17 and 0.19 for the 2010-2011 pairs) and ANN performed better than any other method in all cases.

When comparing performance over survey strata, six principal strata of the survey results remained similar, with ANN performing better in all strata except the list frame subset of records. In Figure 4, areas with a majority of: temporary crops, perennial crops, pasture, forest; areas in the Amazon region and list frame strata are plotted. The size of blocks can vary considerably between strata: 23 farms on average in areas predominantly covered with temporary crops and 10 farms in areas that are predominantly forest. The diversity of land use and farm systems are linked to the strata, and linkage methods performed significantly better in pasture strata and significantly worse in the case of the Amazon Forest region and perennial crop strata. For temporary crops where diversity block size is superior, all methods performed at a lower level. For those strata, differences between supervised methods with sampling variables (ANN "net_s2") and unsupervised ones with agronomic variables (Fastlink "fll_s1") are almost not noticeable and ranked similarly. The list frame (Figure 4, on the last row to the right) behaves differently with an overall better F-score performance than other stratas, and the Adaptive boosting method performed better than other methods only in these strata. This sub-population presents different characteristics: in this subset of farms are only selected farms with important size (over

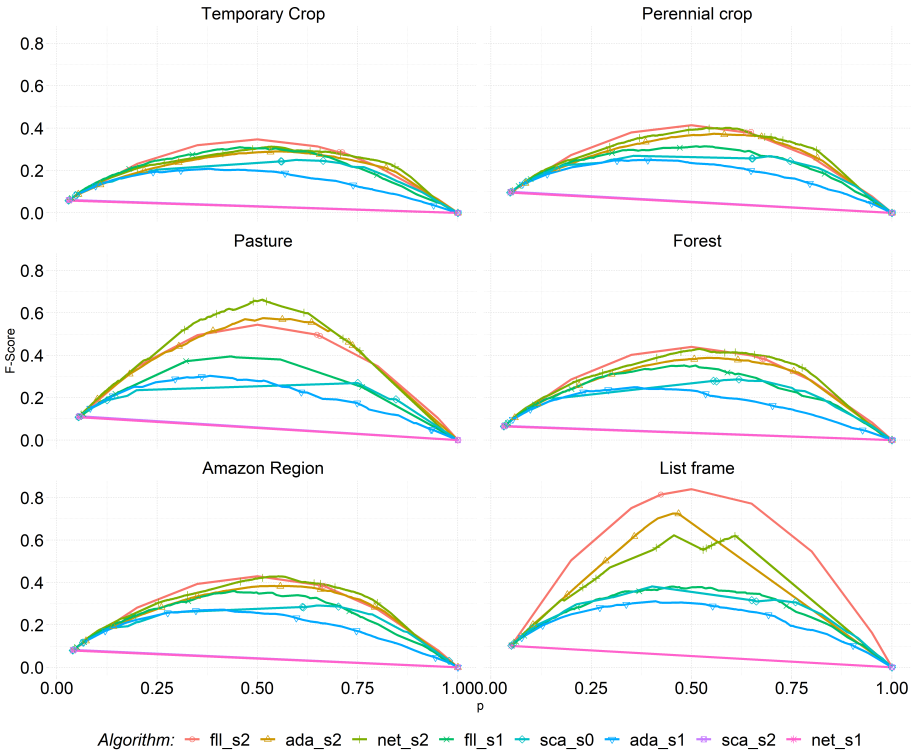


**Fig. 4.** F-score - p plots 2010-2011 pair datasets, and per sampling stratas the following methods are plotted: sca: scaling, fll: fast linkage, nxt: NN-committee, net: NN, ads: adaboost, per stratas.

a 100 hectares), specialized in one crop (over 50 hectares dedicated to only one crop) or specialized (poultry, pig production or flowers for instance). The weak learner combination of adaboost method may be more adapted to capture these variations.

When comparing performance over survey strata, six principal strata of the survey results remained similar with ANN performing better in all strata except for the list frame subset of records. In Figure 4, areas with a majority of temporary crops, perennial crops, pasture, forest, amazon region and list frame strata are plotted. The size of blocks can vary considerably between strata: 23 farms on average in areas predominantly covered with temporary crops and 10 farms in areas that are predominantly forest. Linkage methods performance was significantly better in pasture strata and worse in the Amazon Forest region and perennial crop strata. For temporary crops where block size can reach more than a hundred farms, the results are less accurate. For those stratums, differences between supervised machine learning methods (ANN "net_s2") and

unsupervised ones with agronomic variables (Fastlink "fll_s1") are similar. The list frame (see Figure 4, on the last row to the right) behaves differently with overall better F-score performance than other stratas and the method Adaptive boosting performed better than other methods only in these strata. This subpopulation presents different characteristics, with selected farms according to their degree of specialization and important size. The weak learner combination of the adaboost method may be more adapted to capture these variations.

### 4.2 Results after deduplication

Once a classification algorithm is applied, the complete dataset of pairs, including match and non-match, can be used to describe relations between datasets, using pair weights as ponderation. Nevertheless, as the entities are fixed farms, described only once in each dataset, there is an important overlap between datasets: a large proportion of the individuals are present in both datasets. The process of eliminating duplicates, or deduplication, allows one to obtain a 'clean' dataset with only one farm linked to another. Ideally, overall precision and recall should be maximized to ensure a high linkage quality during deduplication. It is especially difficult to establish a threshold that optimizes F-score, and produces a high match rate. Additionally, for this step, mean weights were averaged to produce an ensemble of learners based only on the best algorithm giving slightly better results than best algorithms (see Table 4: "ensemble").

**Table 4.** Merging results for four different methods, after deduplication

| Method | Variable subset | Year | TP | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| ANN | Sampling | 2010-2011 | 12518 | 44 | 83.6 | 57.6 |
| | Design | 2010-2012 | 6728 | 23.1 | 78.5 | 35.7 |
| | Variables | 2011-2012 | 11536 | 38.9 | 84.9 | 53.3 |
| EM fastlink | Agronomic | 2010-2011 | 12005 | 39.3 | 92.8 | 55.2 |
| | Variables | 2010-2012 | 8315 | 27.7 | 89.4 | 42.3 |
| | | 2011-2012 | 12066 | 39 | 92.8 | 54.9 |
| Ensemble | Sampling | 2010-2011 | 13052 | 48.6 | 79.1 | 60.3 |
| | Design | 2010-2012 | 7098 | 25.6 | 75 | 38.1 |
| | Variables | 2011-2012 | 12021 | 42.8 | 79.9 | 55.8 |

After deduplication: methods performed with average precision but recall remained high: for pairs of consecutive years (2010-2011 and 2011-2012) almost 50% of true match pairs were re-identified, with ANN algorithm leading to highest F1-scores and EM algorithm fastlink with only agronomic variables.

### 4.3 Results on successive years

When evaluating methods identifying pairs of records matched on the three pairs of datasets, validation dataset raised 12280 individuals.

**Table 5.** Merging results for four different methods, with identified individuals over three years

| Method | Variable subset | TP | Precision | Recall | F-score |
|---|---|---|---|---|---|
| ANN | SDv | 5638 | 27.7 | 45.9 | 34.6 |
| EM fastlink | SDv | 4895 | 39.1 | 39.9 | 39.5 |
| EM fastlink | Agv | 4624 | 43.6 | 37.7 | 40.4 |
| Ensemble: majority | SDv | 6292 | 60.4 | 51.2 | 55.4 |

Among evaluated algorithms, interestingly the unsupervised method outperforms the supervised one (see Table 5), in terms of precision and it allows for the identification of almost around 40% of individuals; whereas using ensemble "majority" over deduplicated results more than 50% of these individuals were re-identified.

## 5 Discussion

In this research, we used yearly data to test various record linkage techniques to produce longitudinal data. We showed that common linkage methods demonstrate remarkable results in allowing complex linkage with numerical descriptors between yearly databases. When no true match is available, only unsupervised methods can be used, and Fellegui Sunter algorithm showed similar accuracy as supervised methods (AdaBoost, Artificial Neural Networks), proving adequate to rebuild populations of individuals with anonymized data. In this section, we review the implications of using agriculture survey for record linkage: yearly data, numeric pseudo-identifiers and the interest of using sampling data.

Typical record linkage makes use of textual pseudo identifiers such as name, surname, or company name. This information can be used to re-identify a unique individual with high likelihood, despite differences in text fields. Here, we propose using only numeric attributes such as farmer age, number of workers, categories land tenure, land extension, or irrigation to identify the correct match between data sets. Using similar setup, recent work on the potential of re-identification in public surveys, using only demographic attributes, have shown strong evidence that the combination of pseudo-identifiers (15 socio-demographic variables) in anonymized data-sets lead to very high linkage precision (99.98%) for North-American populations (Rocher et al., 2019).

Re-identification was carried out successfully on numerous national surveys, using sampling zip codes as blocking attributes. Our hypothesis was that by using adequate blocking variables and comparison functions, similar results can be obtained. for farm surveys records. These are, by nature unmovable, but their extension can vary by acquisition or transfer of land. Despite the variable nature of the pseudo-identifiers, we could review differences between identified matches (true match), and show that little variation occurs over time for selected variables. Descriptors of farm categorical characteristics (ownership, type of farm)

and land use (size, pasture, forest, irrigated area, number of parcels) are less susceptible to change from one year to another. For production, cattle density, average milk production per cow, presence of horses or donkeys were robust characteristics to identify a farm over years. Finally, Farmer's personal information (age, sex), or the availability of permanent labor on farm were expect to have the same consistency over time but performed poorly, as a different person is selected from one year to another for the same farm.
Understanding survey sampling structure can contribute to record linkage quality. Sequential numbering within primary sampling units for instance, can be viewed as pseudo identifier. In a sampling unit, the path followed by surveyor is fixed, starting and following the same path every year. This variable, despite the fact that farms are not always surveyed consecutively, was determinant in farm re-identification, and could be used as window blocking, a proxy of the geographical sub units inside sampling units.
Sampling frames for national surveys are generally defined by census enumeration areas, or census tract. These units are the smaller administrative units, standardized in size among urban and rural areas (Aday and Cornelius, 2006).In Ecuador, geographic units were stratified according to a coarse land-use classification and then combined with a list frame. Results showed that within the diversity of rural landscapes, record linkage for the list frame produce the best results, followed by "pasture" farms, extension with prevailing forest cover and the amazon region. This could be explained considering farm density across strata, with fewer observation within sampling unit in the best performing regions. Conversely, high farm density causes an increase in comparison pairs, reducing correct matching.

## 6 Conclusions

Our results help to provide insights on how to improve data integration process for agricultural establishments: (i) carefully selected numeric farm characteristics can provide enough information for matching, (ii) for agriculture survey, geographical blocking allows to reduce calculations, and sequential identifiers and ponderation factors are survey characteristics helping re-identification. Despite the fact that this evaluation was performed on an almost constant sampling design, matching results suggest that at most 51% of individuals could be re-identified thru years, but it is necessary to provide more stable pseudo-identifiers to increase recall levels. In the context of small-scale farm, the major type of farm in developing countries, there is a wide variation in characteristics and non-response rate that affect accuracy of matching. Small scale farming is a key population for future food systems (McIntyre, 2008; Woodhill et al., 2020) and beyond the scope of agriculture production, it is necessary to build, at a national scale, a more efficient information system to understand specificities in small scale farming systems. The scope of these surveys is limited to the productive components of the farms, relevant in the context of developed countries, but incomplete for developing economies. In those countries, produc-

tion is inter-related to the social background, and the multiple incomes of a farmer's family are key drivers for production (Mundler, 2014), but this information is often lacking (Carletto et al., 2013). This evaluation of record linkage methods for agricultural survey shows that despite low false positive rates, the quality of matching experiments led to low recall in matching. For future works, complementing surveys with socio-economic background could provide enough information for better record linkage, and, at the same time, complement information of social background on each farm to provide insights for local policies and development entities.

## References

Abowd, J. M., Abramowitz, J., Levenstein, M. C., McCue, K., Patki, D., Raghunathan, T., Rodgers, A. M., Shapiro, M. D. and Wasi, N. (2019). Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data, *Technical report*.

Aday, L. A. and Cornelius, L. J. (2006). *Designing And Conducting Health Surveys: A Comprehensive Guide*, John Wiley & Sons.

Aiken, V. C. F., Dórea, J. R. R., Acedo, J. S., de Sousa, F. G., Dias, F. G. and Rosa, G. J. d. M. (2019). Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods, *Computers and Electronics in Agriculture* **163**: 104857.
**URL:** *http://www.sciencedirect.com/science/article/pii/S016816991930434X*

Belin, T. R. and Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage, *Journal of the American Statistical Association* **90**(430): 694–707. Publisher: Taylor & Francis Group.

Bellow, M. E., Daniel, K., Gorsak, M. and Erciulescu, A. L. (2016). Evaluating Record Linkage Software for Agricultural Surveys.

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Carletto, C., Jolliffe, D. and Banerjee, R. (2013). The Emperor has no data! Agricultural statistics in sub-Saharan Africa, *World Bank Working Paper* .

Chen, M., Wichmann, B., Luckert, M., Winowiecki, L., Förch, W. and Läderach, P. (2018). Diversification and intensification of agricultural adaptation from global to local scales, *PLoS ONE* **13**(5).
**URL:** *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935394/*

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*, Springer Science & Business Media.

Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., Crosignani, P. and Tessandori, R. (2005). The EpiLink record linkage software, *Methods of Information in Medicine* **44**(01): 66–71.

Davies, C. (2009). *Area frame design for agricultural surveys*, United States Department of Agriculture, National Agricultural Statistics ....

Enamorado, T., Fifield, B. and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records, *American Political Science Review* **113**(2): 353–371.

FAO (2015). World Programme for the Census of Agriculture 2020 Volume 1: Programme, concepts and definitions, *Report*, FAO.
**URL:** *http://www.fao.org/3/a-i4913e.pdf*

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association* **64**(328): 1183–1210.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *icml*, Vol. 96, Citeseer, pp. 148–156.

Fu, Z., Boot, H. M., Christen, P. and Zhou, J. (2014). Automatic record linkage of individuals and households in historical census data, *International Journal of Humanities and Arts Computing* **8**(2): 204–225. Publisher: Edinburgh University Press 22 George Square, Edinburgh EH8 9LF UK.

Fu, Z., Christen, P. and Zhou, J. (2014). A graph matching method for historical census household linkage, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, pp. 485–496.

Goldstein, H., Harron, K. and Cortina-Borja, M. (2017). A scaling approach to record linkage, *Statistics in medicine* **36**(16): 2514–2521. Publisher: Wiley Online Library.

Guillermo Otañez (2004). Diseño de muestreo de la ESPAC, *Report*, BID/INEC.

Guillermo Otañez (2008a). Encuesta de Superficie y Producción Agropecuaria Continua Manual del Encuestador, *Report*, INEC, Ecuador.

Guillermo Otañez (2008b). Plan de fortalecimiento del sistema estadistico gropecuario, *Report FAO/TCP/ECU/3102*, FAO/INEC, Ecuador.
**URL:** *https://anda.inec.gob.ec/anda/index.php/catalog/206/download/4120*

Hand, D. and Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms, *Statistics and Computing* **28**(3): 539–547.
**URL:** *https://doi.org/10.1007/s11222-017-9746-6*

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1): 100–108.

Healy, M. J. R. and Goldstein, H. (1976). An approach to the scaling of categorized attributes, *Biometrika* **63**(2): 219–229. Publisher: Oxford University Press.

Hill, B. (1996). Monitoring incomes of agricultural households within the EU's information system-new needs and new methods *, *European Review of Agricultural Economics* **23**(1): 27–48.
**URL:** *https://doi.org/10.1093/erae/23.1.27*

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference, *Journal of Statistical Software, http://gking. harvard. edu/matchit* .

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association* **84**(406): 414–420.

Jarvis, S., Parslow, R. C., Carragher, P., Beresford, B. and Fraser, L. K. (2017). How many children and young people with life-limiting conditions are clinically unstable? A national data linkage study, *Archives of disease in childhood* **102**(2): 131–138. Publisher: BMJ Publishing Group Ltd.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance, *Journal of Big Data* **6**(1): 27.
**URL:** *https://doi.org/10.1186/s40537-019-0192-5*

Karr, A. F., Taylor, M. T., West, S. L., Setoguchi, S., Kou, T. D., Gerhard, T. and Horton, D. B. (2019). Comparing record linkage software programs and algorithms using real-world data, *PLOS ONE* **14**(9): e0221459. Publisher: Public Library of Science.
**URL:** *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221459*

Kazanjian, A. (1998). Understanding women's health through data development and data linkage: implications for research and policy, *CMAJ* **159**(4): 342–345. Publisher: Can Med Assoc.

McIntyre, B. D. (2008). *International Assessment of Agricultural Knowledge, Science and Technology for Development: Global Report*, Island Press.

Mundler, P. (2014). Unité de l'agriculture et diversité des exploitations agricoles. Des représentations en évolution, *L'agriculture en famille: travailler, réinventer, transmettre* p. 65.

Murray, J. S. (2016). Probabilistic record linkage and deduplication after indexing, blocking, and filtering, *arXiv preprint arXiv:1603.07816* .

O'Donoghue, C., O'Donoghue and Pacey (2017). *Farm-Level Microsimulation Modelling*, Springer.

Pixton, B. and Giraud-Carrier, C. (2006). Using structured neural networks for record linkage, *Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research*.

Remans, R., Jones, S. K., Dulloo, E., Villani, C., Estrada-Carmona, N., Juventia, S. D. and Laporte, M. A. (2019). Agrobiodiversity Index Report 2019: risk and resilience, *Technical report*, Bioversity International.

Reppermund, S., Heintze, T., Srasuebkul, P., Reeve, R., Dean, K., Smith, M., Emerson, E., Snoyman, P., Baldry, E. and Dowse, L. (2019). Health and wellbeing of people with intellectual disability in New South Wales, Australia: a data linkage cohort, *BMJ open* **9**(9): e031624. Publisher: British Medical Journal Publishing Group.

Ripley, B. D. and Hjort, N. L. (1996). *Pattern recognition and neural networks*, Cambridge university press.

Rocher, L., Hendrickx, J. M. and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models, *Nature communications* **10**(1): 1–9.

Rowlands, I. J., Abbott, J. A., Montgomery, G. W., Hockey, R., Rogers, P. and Mishra, G. D. (2021). Prevalence and incidence of endometriosis in Australian women: a data linkage cohort study, *BJOG: An International Journal of Obstetrics & Gynaecology* **128**(4): 657–665. Publisher: Wiley Online Library.

Rässler, S. (2012). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, Vol. 168, Springer Science & Business Media.

Sariyar, M. and Borg, A. (2010). The RecordLinkage package: Detecting errors in data, *The R Journal* **2**(2): 61–67.

Sariyar, M., Borg, A. and Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory, *Journal of Biomedical Informatics* **44**(4): 648–654.
**URL:** *http://www.sciencedirect.com/science/article/pii/S1532046411000372*

Therneau, T. M. (1983). A short introduction to recursive partitioning, *Orion Technical Report* **21**.

Winkler, W. E. (1994). Advanced methods for record linkage.

Winkler, W. E. (1995). Matching and record linkage, *Business survey methods* **1**: 355–384.

Woodhill, J., Hasnain, S. and Griffith, A. (2020). What future for small-scale agriculture.

Received May 4, 2022 , revised September 27, 2022, accepted November 9, 2022