Title of technical report

European Food Safety Authority (EFSA),   
Individual authors (format Name Surname)

Abstract

(Max. 300 words, no paragraph breaks; no tables, footnotes, graphs or figures. Note that the abstract should end with the copyright)

© European Food Safety Authority, 20YY

**Key words:** (max. seven key words)

**Requestor:** add requesting party

**Question number:** EFSA-Q-YYYY-NNNNN

**Correspondence:** xxx@efsa.europa.eu

**Acknowledgements:** [Scientific Committee OR EFSA] wishes to thank the following for the support provided to this scientific output: [staff members or others who made a contribution but are not eligible as authors]. The Panel [Scientific Committee OR EFSA] wishes to acknowledge all European competent institutions, Member State bodies and other organisations that provided data for this scientific output.

**Suggested citation:** EFSA (European Food Safety Authority), Individual authors [add names in the format Surname followed by Initial(s), Surname followed by Initial(s) and Surname followed by Initial(s)], 20YY. Title of the report. EFSA supporting publication 20YY:EN-NNNN. 3 pp. doi:10.2903/sp.efsa.20YY.EN-NNNN

**ISSN:** 2397-8325

© European Food Safety Authority, 20YY

Reproduction is authorised provided the source is acknowledged.

Reproduction of the images listed below is prohibited and permission must be sought directly from the copyright holder:

Figure 1: © Stockphoto; Figure 5: © WHO

Summary

The summary should not include tables, footnotes, graphs or pictures or references.

**Table of contents**

[Abstract 1](#_Toc487211940)

[Summary 3](#_Toc487211941)

[1. Introduction 3](#_Toc487211942)

[1.1. Background and Terms of Reference as provided by the requestor 3](#_Toc487211943)

[1.2. Interpretation of the Terms of Reference (if appropriate) 3](#_Toc487211944)

[1.3. Additional information (if appropriate) 3](#_Toc487211945)

[2. Data and Methodologies 3](#_Toc487211946)

[2.1. Data 3](#_Toc487211947)

[2.2. Methodologies 3](#_Toc487211948)

[2.2.1. Title level 3 3](#_Toc487211949)

[2.2.1.1. Title level 4 3](#_Toc487211950)

[3. Assessment 3](#_Toc487211951)

[4. Conclusions 3](#_Toc487211952)

[5. Recommendations (if appropriate) 3](#_Toc487211953)

[Documentation provided to EFSA (if appropriate) 3](#_Toc487211954)

[References 3](#_Toc487211955)

[Glossary [and/or] Abbreviations 3](#_Toc487211956)

[Appendix A – Title of the appendix 3](#_Toc487211957)

1. Introduction
   1. Background and Terms of Reference as provided by the requestor
   2. Interpretation of the Terms of Reference (if appropriate)
   3. Additional information (if appropriate)
2. Algorithm Description
   1. Overview

I dati vengono prima partizionati poi pairs vengono fatti solo nella stessa partizione (evito esplosione dei dati)

* 1. Data cleaning

The first action needed to develop a robust duplicates detection model is to clean the input data. In fact, a standardized and cleaned dataset allows reaching much more accurate results, since the overall data complexity decreases. The data cleaning steps used in this project are the following:

* Records with unknown tested parameter were removed;
* Records which were not validated by the Data Collection Framework (DCF) were removed;
* Records which were reported without result unit were dropped;
* Records which were reported with too general terms (i.e. root terms) for foodEx1 and parameter were dropped;
* Unicode accented letters were converted into standard letters in all the strings fields, for example ‘Á’ and ‘Ç’ become ‘A’ and ‘C’;
* Special characters, such as ‘#’, ‘$’, ‘%’ were removed from some strings fields.
  1. Partitioning phase

As mentioned in Section 2.1, it is not possible to simply create pairs comparing one record with all the other ones, since this would generate a massive dataset. For this reason, the data should be clustered into several partitions to avoid comparisons which do not make sense (e.g. two records which were collected in two different years are for sure non-duplicates). In fact, each record will be compared only with the records contained in its partition, resulting in a much smaller and manageable dataset. The partitions are created by splitting the data using some specific columns (named key variables), which immediately identify records as non-duplicates. For example, if a record is sampled in France and another one is sampled in Germany instead, then they can be safely labelled as non-duplicates since they are clearly two different records.

In this project, the data partitioning was achieved using several key variables of different type, such as categorical and numerical variables. In particular, each variable type required a different partitioning methodology for splitting the data, as explained in the following Sections.

* + 1. Partitioning based on generic categorical variables

The simplest way of partitioning data is to split them using categorical variables, which are, variables that can assume only a limited set of values in the dataset. In fact, it is sufficient to split the data based on the variable value to create a set of partitions. An example of this type of variable is the result unit column, which identifies a unit of measurement. In fact, this variable is controlled terminology (Unit catalogue) and therefore it can assume just a fixed set of values. Note that variables which represent a year are not real categorical variables since they are numbers, but actually they can be considered as such, because they can assume just a relatively small set of values in the dataset. SPIEGA PERCHè LI USIAMO

In this project, the data were partitioned with this approach using the following fields:

* Sampling year: the year when the record was sampled;
* Analysis year: the year when the record was analysed;
* Result unit: the unit of measurement for result value and other variables;
* Organisation code: who provided the record;
* Parameter code – level two: the second level parent of the reported parameter in the EFSA Param catalogue (i.e. records which have the same parent in the second level of the tree structure of the catalogue are put into the same partition);
* Efsa product code – level two: the second level parent of the reported foodEx1 code in the EFSA Foodex catalogue.
  + 1. Partitioning based on sampling country

The sampling country variable indicates where a record was sampled. In particular, it is a controlled terminology field, which means that only values contained in the related EFSA catalogue (in this case the Country catalogue) can be set. The main issue related to this field is that generic values can be reported instead of the real country, such as ‘European Union’ or ‘European Economic Area’, making the partitioning process on this variable more difficult to perform. In fact, it is very common that a duplicate has the real sampling country value in one record, and a generic entry to which the real country belongs to in the other one (e.g. ‘Italy’ and ‘European Union’). Therefore, it is not possible to use the approach presented in Section 2.3.1, because generic records and detailed records would be put in different partitions, making the detection of this type of duplicates not possible (generic and detailed records will not be compared). For this reason, it was necessary to develop a wiser partitioning procedure which takes into account generic entries, in order to be able to compare generic sampling countries with generic ones. In particular, generic entries were inserted in all the partitions generated by a detailed sampling country which is indeed related to the generic entry. For example, the ‘European Union’ records were put inside all the partitions which were created with a detailed European country, such as ‘Italy’, ‘Germany’ or ‘France’.

* + 1. Partitioning based on result value

The result value variable is a numerical field and it represents the result of the related analytical measure in the unit specified in the Result Unit variable. Note that comparing records which have a very different result value would be useless, since they are clearly non-duplicates. However, it could happen that the result value of two records which are indeed duplicates can slightly differ due to human errors. For this reason, it was necessary to create an algorithm which is capable of creating partitions based on this numerical value, that is, grouping similar result values into the same partition.

HOW?

* 1. Pairs creation phase

Si partiziona di più e poi si calcolano le distanze tra I pairs

* + 1. Further partitioning based on pairs values
    2. Distance measures

Scrivi che per tutti I casi se c’è un unknown allora è 0.5. metti annex con tutti gli unknown di tutti i campi

* + - 1. Equal distance
      2. Contains distance
      3. Country distance
      4. Numeric distance
      5. Unit distance

C’è variable che ha sia il valore che l’unità e quindi controlliamo il valore solo se ha la stessa unità altrimenti no

* + - 1. String distance

Spiega complev compged e spedis e perchè si è scelto compged (veloce e abbastanza corretta con le stringhe a parte casi rari)

* + - 1. Hierarchy distance

Spiega come funziona la hierarchy distance e Parla anche del prod treat dist che è caso particolare

* 1. Classification of duplicates
     1. Weighted distance

Spiega anche che la threshold è a 0 per evitare che o res code o lab samp code sia dist = 1, visto che org code è sempre uguale (usato per partizionare)

* + 1. Classification rule

1. Algorithm Performance Spiega come si comporta l’algoritmo al variare dei dati e quanto ci mette usando il sas server
2. Conclusions
3. Recommendations (if appropriate)

Documentation provided to EFSA (if appropriate)

1. Dossier name. Month YYYY. Submitted by [name of the company]

References

Alderman G and Stranks MH, 1967. The iodine content of bulk herd milk in summer in relation to estimated dietary iodine intake of cows. Journal of the Science of Food and Agriculture, 18, 151–153.

Glossary [and/or] Abbreviations

Glossary: an alphabetical list of words relating to a specific subject, text, or dialect, with explanations; a brief dictionary.

Abbreviation: a shortened form of a word or phrase (such as Mr., Prof.). It also includes acronyms (a group of initial letters used as an abbreviation for a name or expression, each letter being pronounced separately – such as DVD, FDA – or as a single word – such as EFSA, NATO).

|  |  |
| --- | --- |
| XXX | Dsadsadsadsa |
| YYY | Sdsdsadsad |
| ZZZ | Fdsfsafasdf |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Unknown variable values