Economia agro-alimentare /
Food Economy

An International Journal on Agricultural and Food Systems

Vol. 23, Iss. 3, Art. 9, pp. 1-16 - ISSN 1126-1668 - ISSNe 1972-4802
DOI: 10.3280/ecag3-oa12760

Mapping data granularity: The case of FADN

Concetta Cardillo[*,a], **Giuliano Vitali**[b]

[a] CREA, Research Centre for Agricultural Policies and Bioeconomy, Italy
[b] Alma Mater Studiorum University of Bologna, Italy

Abstract

The present analysis looks into the issue of mapping information contained in the FADN database aimed at finding a methodology useful as a preliminary analysis to data extraction. To the purpose the concept of data granularity has been introduced. The method has been used to perform a farm-based analysis, revealing a wide heterogeneity of factors and levels that show the existence of specific data 'patches'. The work proved to be able to increase awareness regarding effective data availability as a preliminary analysis to queries performed on relational data-bases which are not designed from a systems basis, and that can be considered valid for any survey-supplied data.

Article info

Type:
Article
Submitted:
14/05/2021
Accepted:
21/09/2021
Available online:
12/01/2022

JEL codes:
C18, C81, Q12

Keywords:
FADN
Farms
Survey
Granularity
Big data

Managing Editor:
Lucia Briamonte,
Luca Cesaro,
Alfonso Scardera

* *Corresponding author*: Concetta Cardillo - Researcher - CREA, Council of Research in Agriculture and analysis of agricultural economics, Centre of Policies and Bio-economics - 00198 Rome, Italy. E-mail: concetta.cardillo@crea.gov.it.

Introduction

Understanding and monitoring the agricultural sector, exploring farm structure and dynamics is a fundamental task of every country, and one of the most powerful tools developed from the EU is represented by the Farm Accountancy Data Network (FADN). It is a sample survey conducted every year by EU Member States on the basis of a common regulation and a harmonized methodology. Every EU country developed and is currently managing its own FADN – compliant database, whose standard is defined in the 'form and shape of farm return' (EU, 2015). Therefore, the FADN database includes a common dataset of mandatory fields but it also provides information that can be different in each country, including several orders of information, e.g. dealing with agrotechnology, market and sustainability.

The FADN database represents an important source both for policy makers and for researchers, indeed it has been already used e.g. for decision making, to assess CAP, to estimate farm efficiency, and compare production activities. However, the number of field descriptors, and observed values is making FADN a complex database (Hand, 2020), with consequences on the performances of any models making direct or indirect use of collected data.

The quality of a dataset is given by several aspects including Accuracy and Precision, Legitimacy and Validity, Reliability and Consistency, Timeliness and Relevance, Completeness and Comprehensiveness, Availability and Accessibility, Granularity and Uniqueness (see e.g. Harrington, 2016).

Most of these qualities can be measured by statistically-based metrics (Karr, 2006), also when having to do with heterogeneous data (Micic, 2017); however, such approaches are hardly useful to detect the level of detail of available data.

Data availability is at the base of statistical approaches adopted to derive indicators and technical parameters. However, the number of factors and levels available to compute such values depend on each other.

To the scope the concept of granularity has been explored here, meant as the number of factors and levels that determine the degree of aggregation to be used to produce statistically significant values.

To the purpose at first, we will have a glance at the FADN dataset and describe the methodological approach adopted, then we will describe the distribution of granularity and its effects on two case studies. Finally, we will draw the final conclusions.

2. Materials and methods

The FADN database

The FADN survey is carried out in each Member State by a liaison agency which in Italy is represented by CREA (Council for Research in Agriculture and Agricultural Economics analysis, research centre supervised by the Ministry of Agricultural, Food and Forest Policies). The survey, performed through a network of data collectors (experts of the agricultural sectors), is performed on a sample of agricultural holdings with an economic size of commercial (equal or more than 8,000 euro), which are selected on the basis of sampling plans established at the level of each Member State and according to guidelines provided by the European Commission. The sampling plan ensures the representativeness of the returning holdings as a whole and defines the number of farms to be selected by region, type of farming (ToF) and economic size classes, expressed in terms of standard output (SO), and also specifies the rules applied for selecting the holdings. The random sampling allows the extension of the results of the farms in the sample to the universe of the farms as a whole that is formed by the subset of the EU universe.

The information collected draw a portrait of farm's structure, their financial and economic aspects, environment, social issues, labour machinery etc. In particular, in the database it is possible to find general information related to farms (as Utilized Agricultural Area (UAA), economic size as standard output (SO), working units, kind of property, legal form, Gross production etc.). Other information is related to financial and economic aspects (as derived from accountancy as costs, investments, debts, value added, assets and liabilities, subsidies). In the FADN database it is also available some information related to social aspects (level of education, age of farmers, gender, labour etc.). Other information are linked to statistical aspects (information on sample and weights), environmental aspects (use of fertilizers and pesticides, use of water), and detailed information about land use (hectares of area dedicated to each cultivation) and livestock (number of heads of animal per species and categories). Some of the variables in the database are continuous (surfaces, number of heads, working hours, KW, subsidies received, etc.) others could be categorized in different classes (classes of UAA, classes of SO) or modalities (as altimetry: mountain, hill, plain) while others could be dichotomous (yes/no). Additional tables have been recently introduced in order to simplify specific research. The list of most relevant tables and a synthesis of their contents is given in Annex 1. A description of meta-data is reported in FADN documentation (RICA, 2021).

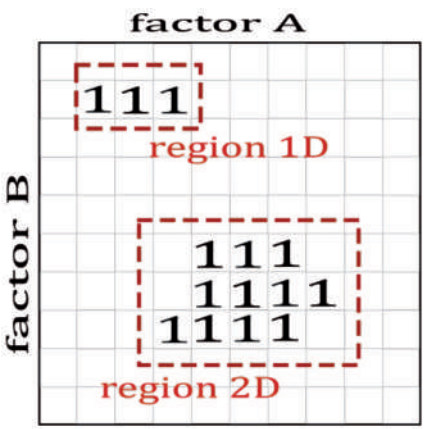
The FADN database is mainatined as a relational SQL-DB by the applicative GAIA that helps validation, gap filling and imputation. Tables could be linked directly (1-1) or after some elaborations (1-N) and they have a high level of redundancy allowing a prompt readability and giving the user the possibility to manage them independently – moreover codes accompanying descriptors increase their robustness.

A definition of granularity

Granularity, commonly referred to space or time data resolution, can also be referred to more general and abstract features. Each conventional Data-Base collects two kinds of information, alphanumeric/descriptive/categorical and numerical ones. Though every field can be used to extract records, categorical and numerical-discrete can be used to classify values, allowing each variable to be hosted by a (sparse) N-dimensional matrix (N being the number of factors).

Figure 1 display the case of a sparse 2D matrix representing the availability of data on 2 factors survey, where it is possible to identify the localised availability of data on levels 2,3,4 of factor A for some samples, and availability of several combinations of levels of factors A and B in another region, meaning that in the first region the matrix has a 1-dimension character, whereas it has a 2-dimensional character in the second one.

Figure 1 - Example of a sparse matrix representing data availability for 2 factors of 10 levels each



In terms of data collected from surveys (as FADN), each categorical data determines a dimension, and records may be split on the base of each of them over and over, till the sample size becomes too small (or empty) to compute reliable statistics for a given variable, or there are no levels to be compared as level are not equally populated – e.g. in FADN, farm managers are mainly males. Also, things could be different for each variable – information on farm activities (cropping/livestock) depends on farm specialisation. Granularity of each variable is described by the distribution of factors (local dimension of data matrix), levels (identifying the region of the matrix rich of data) and sample size characterizing the context of interest.

The analysis of granularity aims at producing a map of data availability which is preliminary to any statistical analysis including dimension-reduction (as Principal Component Analysis, Factor Analysis, discriminant analysis, etc..), aimed at describing a data-set which has not been designed on the base of a systemic view/model.

Exploring granularity distribution is a matter of combinatorics. As an example, if factors are *a,b,c,* possible combinations are represented by: *a, a-b, a-b-c, a-c, b, b-c, c*. In general, possible combinations of *k* factors can be obtained by: $\sum_{k=1}^{n}\left(\frac{n}{k}\right)$. Being factor levels given by (n_a, n_b, …), the number of possible levels for each combination is: $NL = \prod n_a \cdot n_b \cdot \ldots$ - namely, if n_a*=3,* n_b*=4,* n_b*=5*, the potential number of levels for combination *a-b-c* is $n_a \cdot n_b \cdot n_c$ = *60*. For more factors and levels combinations increase considerably generating a three graph whose exploration represents a well-known computational problem.

As the splitting is carried on, the sample size does not allow statistical analysis: e.g. organic farms are far less numerous than conventional ones and as increasing the number of factors, splitting make the size organic sample too small. As the majority of analyses performed on FADN are aimed at comparing variables related to different levels, a minimal size is required – in the following analysis a minimum sample size (*NMIN*) of 5 has been adopted.

In this study we focus the attention only on data collected on farm tables. Each farm has a maximum of 158 numerical variables, coming from 5 tables (FARMS, ENVIRONMENT, LAND-USE, BUDGET-CE and BUDGET-SP) some of them representing intensive values (indicators), while other are farm-wide (e.g. total surface, income, labour, or livestock units).

Tables include 13 non-redundant categorical data (factors), which are listed in table 1 together with the number of their levels. Four of them are dichotomous (yes/no) while others describe the same character with a different detail, - REGION (21) & AREA (5), ALTITUDE_3 (3) &

ALTITUDE_5 (5), TOF_4 (61) & TOF_2 (9). For the remainder of the study only the first one of each listed couple will be used. In table 1 they have been identified with a code, which are used in discussion below.

Table 1 - Factors of farm-based tables and their modalities

Code	Field	Description	Classes or modalities
A	REGION	Administrative Region or autonomous province (NUTS2 territorial units)	19 administrative regions and 2 autonomous provinces of Trento and Bolzano
	AREA	Grouping of Regions	North-West, North-East, Center, South, Islands
B	ALTITUDE_3	Identification code of the altimetric area a three types	Mountain, Hill, Plain
	ALTITUDE_5	Identification code of the altimetric area a five types	Internal mountain, coastal mountain, internal hill, coastal hill, plain
C	LFA	Less favoured area	Municipal territory not disadvantaged; Partially mountainous and partially disadvantaged municipal area**;** Totally mountainous and totally disadvantaged municipal area**;** Municipal territory with total or partial disadvantage due to depopulation; Municipal territory with specific disadvantages, partially or totally
D	TOF_4	Type of farming (61 levels)	Detailed levels of activities as defined in EU regulation 220/2015
	TOF_2	Type of farming (9 levels)	Specialist field crops; Specialist horticulture; Specialist permanent crops; Specialist grazing livestock; Specialist granivores; Mixed cropping; Mixed livestock; Mixed crops – livestock; Not classified
E	MANAGEMENT	Type of farm management	Direct with family members only; Direct with a prevalence of family members; Direct with a prevalence of extra-family; With wage earners; With only subcontracting; Other forms of management

Table 1 - Continued

Code	Field	Description	Classes or modalities
F	LEGAL_FORM	Farm legal form	Individual holding; Simple company; Company at collective name; Joint stock company (S.p.a.); Cooperatives (limited or unlimited liability); Other typology; Limited partnership (S.a.s.); Limited Liability Company (S.r.l.); Limited partnership by shares (S.a.p.a.); - Social cooperative; Other recognized and unrecognized association; Public authority
G	GENDER	Gender of farmer	Male/Female
H	SETTLEMENT	Method of settlement of the entrepreneur	Direct with family members only; Direct with a prevalence of family members; Direct with a prevalence of extra-family; With wage earners; With only subcontracting; Other forms of management
I	YOUNG	Presence of Young entrepreneur	Y = the entrepreneur is less or equal to 40 years old; N = over 40 years old
J	DIVERSIFIED	Presence of farms diversification activities	Y/N
K	ORGANIC	Presence of Organic farming	Y=organic; N=conventional
L	CLASS_UAA	Class of Utilized Agricultural Area	Less than 5 ha; 5 - 15 ha; 15 - 40 ha; more than 40 ha
M	CLASS_PS	Class of Standard Output (€)	4.000-8.000; 8.000-15.000; 15.000-25.000; 25.000-50.000; 50.000-100.000; 100.000-250.000; 250.000-500.000; 500.000-750.000; 750.000-1.000.000; 1.000.000-1.500.000; 1.500.000-3.000.000; more than 3.000.000 €

3. Results

The factors (*NF=13*) result in a number of possible combinations *NC = 8191* and to a potential number of levels *NL=16'114'775'040.*

To explore the combinations a recursive code has been developed in R, which has been run on FADN 2015 data-set. Because of the large number

of combinations, a stopping rule has been included on the base of subsample size (*NSTOP*).

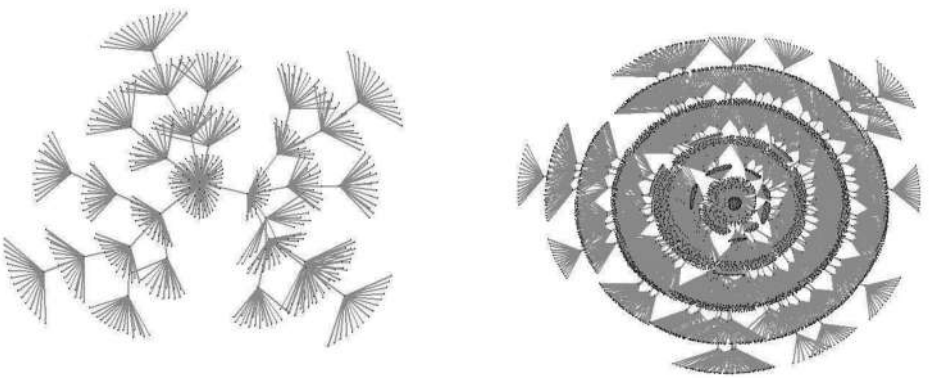
The computing time has been estimated as *t = 10 (a - b log10 NSTOP)* (with *a=7.47, b=1.81, R=0.999* on a core *i7* at *1.80GHz*).

The distribution of granularity, that is the combination of factors/levels/ sample size is written into a table including the possible factorial analysis that can be performed on FADN for FARM-based records with the selected sample size.

As each combination is deriving from a simpler one, the process of exploration of combination can be represented as a tree graphs that can be used to show a combination of factor levels that can be progressively added to a factorial analysis.

Figure 2 shows the graphs obtained with *NSTOP=5000 (572 edges, left side)*, and *NSTOP=2000 (4556 edges, right side).*

Figure 2 - Plots of tree graph[1] *derived for a threshold of 5000 records (left) and 2000 (right)*



As the threshold (*NSTOP*) decreases the graphs become more and more complex and hardly readable, however it can be easily seen that with *NSTOP=5000* a maximum of *5* factors can be explored, and with *NSTOP=2000*, 7 factors. Also, most of the branches are very selective - for each factor, few levels are selected for more complex combinations of factors.

1. Tree graphs have been generated by *igraph* (Csardi, 2006) and *ggraph* (Pedersen, 2021) R libraries.

Case studies

To make the method fully understandable, a closer look to the technique is given, starting from a table obtained for *NSTOP=50*, resulting in *513'795* combinations of factors/levels, that required a computation time of *6.9 hrs*.

1. Organic farms - In the last decades, interest in sustainable crop management has grown considerably, and analysing the differences between organic and conventional farming is an issue of considerable interest. However organic farm samples in FADN are not as rich as required for a thorough analysis. From the list of combinations obtained setting *NSTOP=50*, only *8'847* (*1.7%* of total) allow to have an adequate number of farms with both levels of ORGANIC (y/n) in the same context (combination of factors) - some of those combination, together with their numerosity, are reported in Table 2. Sample size shows that organic farms are far less than conventional (about *1/20*) on a national basis, however an analysis can be performed on almost every REGION (not A [2]). Also, for each region analysis can be performed just on some ALTITUDE (B) - for region A [1] only level B [1] (plain) is adequately populated. The same happen crossing regions with LFA (level C [2]), and few with TOF (D [2,9]), or MANAGEMENT (E [5]). Looking at the possibility to analyse the effect of ALTITUDE (B), not splitting the sample over the regions, samples are adequately large to analyse the combination with almost every LFA (C[2-6], and we can do the same combining YOUNG and DIVERSIFIED - in this case as both factors are dichotomous (y/n) splitting is reduced and the analysis includes all levels. As the combination includes more and more factors the levels included decrease and comparison can be done only for a few combinations of levels.

Table 2 - Excerpt from granularity table with records allowing to compare organic and conventional farms. The first column reports combination of factor & level, the right ones sample sizes of the two levels of factor ORGANIC (y/n)

Combination of factor[level]	**Organic**	**Conventional**
K	465	8582
vs regions		
A[1]-K	23	515
A[3]-K	19	356
…		
vs regios & other factors		
A[1]-B[1]-K	19	379
A[1]-C[2]-K	13	281

Table 2 - Continued

Combination of factor[level]	Organic	Conventional
A[1]-D[2]-K	7	57
A[1]-D[9]-K	9	177
A[1]-E[5]-K	13	170
…		
vs altitude & LFA		
B[1]-C[2]-K	96	1706
B[1]-C[3]-K	31	402
B[1]-C[4]-K	33	480
B[1]-C[5]-K	98	1234
…		
vs young & diversified		
I[1]-J[1]-K	316	6683
I[1]-J[2]-K	59	793
I[2]-J[1]-K	75	944
I[2]-J[2]-K	15	162
…		
vs region, altitude & LFA		
A[1]-B[1]-C[2]-K	13	281
…		
vs more factors		
A[14]-B[3]-C[2]-D[9]-F[5]-G[2]-I[1]-J[1]-K	8	60

2. Farm Income - Economic sustainability is the foremost important aspect for a farmer, and FADN is expected to be a source of important information to develop market analysis and economical performances of different sub-sectors. An interesting analysis could be aimed at detecting which aspects CLASS_PS (M, see table 1) may be related to. To the scope from the whole set, combinations are selected with at least two levels of CLASS_PS for the same combination (context). The selection includes 13% (66'584) of total combinations. While 1/4 (15'620 combinations) involving REGIONS (A, all levels), 1/2 (30'068 combinations) ALTITUDE (B), 1/2 (28'672 combinations) LFA (C[2-6]), 1/4 (25512 combinations) TOF (D[2-3,5,8-9]), 1/2 (27'770 combinations) MANAGEMENT (E), 1/2 (29'583 combinations) LEGAL FORM ([5,7,12]), 1/2 (31'075 combinations) GENDER (G), 1/2 (27'876 combinations) SETTLEMENT (H), 1/2 (30'721 combinations) YOUNG (I), 1/2 (29'720) DIVERSIFIED (J), 1/2 (24'012 combinations) ORGANIC (K),

and 5/8 (36'219 combinations) UAA (L). However not every level is equally represented. If the 4 levels of L are comparably populated in terms of combinations (6732, 7430, 9369, 12688), for ORGANIC, 23'343 combinations are related to conventional farms and only 668 combinations are related to organic ones. Of the latter ones only 147 combinations include the UAA (L [2-4]) - on the other hand conventional farms can be combined with every UAA combination (2835, 3105, 3894, 5081).

The whole sample (9'024 farms) can be used for a general study though it appears that 7 to 11 PS levels are covered out of the 16 possible - the first 3 rows reported in table 3 show the regions with different PS class coverage. As combinations become more complex, the modalities available may decrease, as can be seen in the 4th row of table 3, or maintain its range, as in one of most complex combinations reported in 5th row.

Table 3 - Excerpt from granularity table with records allowing to compare farms with a different class of Standard Product. The first column reports combination of factor & level, the second the corresponding sample size for the available levels (minimum 2) of factor CLASS_PS

Combination of factor[level]	**CLASS_PS**															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
A [1]	–	–	–	47	78	147	119	96	34	11	–	–	–	–	–	–
…																
A [6]	–	–	–	123	15	116	12	94	54	54	19	24	8	–	–	–
…																
A [13]	–	–	14	44	65	19	131	128	48	23	7	6	6	–	–	–
…																
A [1] -B [1] – C [2]	–	–	–	27	53	95	59	36	14	–	–	–	–	–	–	–
…																
D [9] -E [5] -F [5] -G [2] -H [7] -I [1] -J [1] – K [1]-L[4]	–	–	7	22	12	33	6	33	5	21	11	8	15	–	–	–

4. Conclusions

In the research, not every data-set is designed with a system view point and with a model in mind, and observed variables and factors are identified by different criteria, including economical aspects. Nonetheless such data-bases, including the FADN, collect a large amount of information, allows to obtain indicators and technical coefficients of relevant importance. However,

such values cannot be obtained for every context (e.g., crop or livestock data are only recorded in specific farm types, or some regions).

The presented methodology is aimed at defining a technique for a preliminary exploration of data-sets based on identification of regions of data with a potentially higher density of information, challenging the issue of mapping information contained in the FADN database as a preliminary step for further investigations. Several orders of problems have been faced - together with the factorial & combinatorial aspects.

The method used to perform a farm-based analysis put in evidence a large heterogeneity of factors and levels that witnesses the existence of specific data 'patches' or clusters - in terms of granularity it means that a fine-grained texture characterizes only specific combinations of factors/levels.

The work proved to be able to increase the awareness about effective data availability as a preliminary analysis to queries performed on a relational database as FADN, which can be considered valid for any survey-supplied data.

The approach is expected to be useful to FADN management boards, to increase homogeneity of data granularity by optimising farm sampling or rearranging survey entries (in respect of FADN rules).

Further on, the way to display results has been challenged to make the analysis and result readily comprehensible. The possible combinations of records rich enough of information to enable statistical analysis is so huge to make static tree diagrams only useful to have a glance at granularity distribution and complexity, not to a direct browsing of information.

Future directions include the possibility to use interactive visualization tools to navigate combinatorial graphs.

Annex 1 - List of most relevant tables in FADN database

- FARMS reports for each holding information on location at several levels of space granularity, organization, management, profile, economic aspects, resources availability and usage. FARMS is considered the main table - its records could be matched directly (1-1) with some other tables such as ENVIRONMENT, SAMPLE, BUDGET-CE, BUDGET-SP and indirectly (1-N) with every other table.
- ENVIRONMENT collects information useful to understand farm environmental conditions, including altitude, slope, soil texture, water availability and type of irrigation. Latitude and Longitude has been considered for a range of years but definitely removed (in 2017) because of the spatial resolution of FADN (Lat/Lon only referred to farm administrative address). Other databases, (e.g. the one managed by AGEA, linked to satellite imagery and aircraft survey), are expected to supply a parcel-level land-use detail.
- SAMPLE collects information related to the representativeness of selected farms on the land-use. FADN adopts the principle that farms should be sampled with the aim of representing a country level universe. Such strategy allows to obtaining an integrated survey structured unit able to increase considerably record reliability, and allowed, from 2003, to give to each farm a weight estimating its representativeness on a national and regional[2] basis, which is obtained from three variables: region, economic size (since 2010 expressed in Euro) and type of farming, following the Neyman methodology [23]. Each year-farm entry, coded by a series of identifiers, reports weight (from universe and sample size) allowing to scale up each farm data to get an estimate of its territorial relevance.
- CERTIFICATION table records information on type of certification and its object (the farm or a given surface) e.g. denomination of origin or geographical indication.
- SUBSIDIES table collects information detailed by type of subsidy, type of policy, duration, amount[3].
- BUDGET_CE collects most terms (63) of farm accountancy with different aggregation criteria (see Figure 1).
- LABOUR COSTS table collects details on costs related to different kinds of labourers (e.g. external, family) as number of people, working days, salary.
- BUDGET_SP collects most of terms (44) estimating farm capital.
- BUILDINGS collects information about estates by building typology and ownership, including number of buildings, size, age and value.
- MACHINERY collects information about machinery power by type of machine and ownership, including power, age, and value.
- LAND-USE reports the surface utilized of main surface types of cultivations (arable, permanent crops, pasture, horticulture, woody crops, woodland, tares – woodland and woody crop surfaces are usually not included in cropland).

2. Hereafter “regions” are NUTS2 territorial units

3. In CERTIFICATION and SUBSIDIES, objects may be represented by the whole farm or the specific activity.

- SURFACES collects data on each farm surface – for each cadastral quality and slope class, collects the number of fields, their altitude, total surface (detailed ownership, irrigated) and estimated value. Data however do not indicate spatially explicit land-use patterns due to the non-spatial nature of FADN requirements.
- PLANTS collects economical detail on permanent crops by variety and training system.
- CROPS include both economical (detail of costs and income) and agrotechnological information by species, cropping system including use of resources. About measure units, the EU suggests using quintals and hectares when possible.
- WATER-USE collects information by species and cropping system of irrigation details (days, avg daily hours, water usage and if combined to fertiliser).
- FERTILIZERS collects information by species and cropping system and typology of product together with distributed amount and N, K and P content.
- PESTICIDES collects information by species and cropping system and typology of product, indicating measure unit, position, class of toxicity of: unit price, distributed amount, distributed value, crop surface.
- LIVESTOCK collects information by livestock typology, attitude, and technical information (Head Units, milk units, milk production) and economic data (e.g. SGP, TGP, GPS, FRP, …).
- ANIMALS reports information by animal species and category, management, reporting prevalent attitude of: number of heads, weight, age, lifespan, value.
- PRODUCTS records for each product, type of warehouse, cropping system, identifying measure units: production, together with initial and final inventories, acknowledged, sold and transformed amounts and values.
- SERVICES reports for every activity, offered service, use of renewable energies: size, annual capacity (e.g. customers).
- LABOURERS collects for every worker (personal information is omitted) by type of job, specialization, country of origin, and sector of activity: number of males and females, hours, hours machine, working days and third party.
- PERSONNEL add registry information including gender, family member or relative, management role, level of study, professionalisation, external job, external income, year of birth, year of enrolment.

References

Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal - Complex Systems*, 1695 pp. -- https://igraph.org.

EU (2015). COMMISSION IMPLEMENTING REGULATION (EU) 2015/220 of 3 February 2015 -- https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0220&from=EN.

Hand, D.J. (2020). Dark Data: Why What You Don't Know Matters. Princeton University Press, doi: 10.2307/j.ctvmd85db.

Harrington, J. (2016). *Data Quality. in Relational Database Design and Implementation* (Fourth Edition) (pp. 509-520). Morgan Kaufmann, doi: 10.1016/B978-0-12-804399-8.00025-9.

Karr, A.F., Ashish, P.S. & Banks, D.L. (2006). *Data quality: A statistical perspective. Statistical Methodology*, *3*(2), 137-173.

Micic, N., Neagu, D., Campean, F. & Habib Zadeh, E. (2017). *Towards a Data Quality Framework for Heterogeneous Data*, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

Pedersen, T.L. (2021). An Implementation of Grammar of Graphics for Graphs and Networks, 143 pp. -- https://cran.r-project.org/web/packages/ggraph/ggraph.pdf.

RICA (2021). -- https://rica.crea.gov.it.

Concetta Cardillo
CREA, Council of Research in Agriculture and analysis of agricultural economics, Centre of Policies and Bio-economics
CREA-PB - 00198 Rome, Italy
E-mail: concetta.cardillo@crea.gov.it
Concetta Cardillo is Researcher at CREA and she is involved in several research groups, her key qualifications are: Sampling procedures and statistical methodologies related to agricultural and environmental statistics; Database management and data elaboration of FADN, FSS, Agricultural Census and other relevant surveys on agriculture and environment; Multivaried analysis, in particular "Multiple correspondences" and "Cluster analysis"; Calculation and review of Standard Output coefficients through the collection and elaboration of statistical information from different sources; Microeconomic research and analysis of impact of agricultural policies on the basis of statistical surveys; Community typology.

Giuliano Vitali
Department of Agricultural and Food Sciences (DISTAL), Alma Mater Studiorum-Università di Bologna
UNIBO-DISTAL - Viale Giuseppe Fanin, 44 - 40127 Bologna, Italy
E-mail: giuliano.vitali@unibo.it
Giuliano Vitali is researcher at UNIBO - Dept. of Agricultural Science and Food Technology. As an agro-physicists he worked on several fields of agronomy included soil physics, agro-meteorology, and environmental data analysis, plant – crop and farm modelling, and recently on farm data analysis – included FADN data.