Journal of Geographical Systems
https://doi.org/10.1007/s10109-025-00472-5

ORIGINAL ARTICLE

A flexible approach for statistical disclosure control in geospatial data

Jon Olav Skøien[1] · Nicolas Lampach[2,3] · Helena Ramos[2] · Rudolf Seljak[2] · Renate Koeble[1] · Linda See[4] · Marijn van der Velde[5]

Received: 20 January 2025 / Accepted: 23 July 2025

Abstract

Due to confidentiality restrictions in releasing census and survey data, such as agricultural data from the European farm structure survey (9 million records), the data are aggregated to a coarse resolution (NUTS2 administrative regions) before public release. Even when other types of census data are released as grids, grid cells may be suppressed in locations where confidentiality rules have not been respected. Here, we present a method, implemented in the *R* package *MRG*, for creating multi-resolution grids that respect restrictions while maximizing the spatial resolution at which the data are disseminated. The method can be adjusted for different restrictions, it can create the same grid structure for a set of variables, and it allows for a contextual suppression of some grid cells (i.e., suppress if all neighbors are non-confidential, merge if several others are also confidential) if this results in a generally higher information content, a combination of features that has not previously been available. The method is exemplified with a synthetic data set.

Keywords Agricultural census · Confidentiality · Multi-resolution grid · *R* · statistical disclosure control · Variance estimation

Mathematics Subject Classification C8 · Q1 · Y91

The work was performed fully at JRC's Ispra site premises under the contracts FL1745JOS and FL1840RK).

✉ Marijn van der Velde
Marijn.VAN-DER-VELDE@ec.europa.eu

[1] ARHS Developments, Belvaux, Luxembourg

[2] European Commission, Eurostat, Luxembourg, Luxembourg

[3] University of Natural Resources and Life Sciences, Vienna, Austria

[4] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

[5] European Commission, Joint Research Centre, Ispra, Italy

Published online: 29 August 2025

1 Introduction

Across many domains, it is common to distribute data in the form of grids, where the grid cells represent sums or averages of the recorded values within the cells. These grids usually have a common resolution for all grid cells, independent of the number of records. This works well for many applications, but there are cases where we cannot or do not want to disseminate grid cell values unless they respect certain restrictions. These can be based on confidentiality (we cannot reveal information that might lead to identification of individual records), statistical reliability (we do not want to reveal information with too high an uncertainty) or other, more field-dependent restrictions.

Census and sample survey data are examples of data sets in which the distributed data must respect both confidentiality and statistical reliability restrictions. Historically, the solution has been to adopt a very conservative approach to data dissemination, resulting in a coarse aggregation level for publicly available data compared to raw data. Although the methods and software presented here are applicable to any type of census and survey data, the examples relate to the dissemination of data from the European agricultural census.

An agricultural census involves the regular and systematic collection of data on the structure of a nation's agricultural sector. The unit of data collection is the agricultural holding (farm), which is comprised of the parcels of land and livestock managed by a single entity, such as an individual, household, or a public or private sector organization, for the purpose of agricultural production. By collecting information at regular intervals over time, such as the size of the farm, crop and livestock production and agricultural inputs, any changes in the agricultural sector can be monitored as well as their impacts on food security and the environment (FAO 2017a).

Decennial agricultural censuses have been taking place since 1930 as part of the World Agricultural Census (Ribi Forclaz 2016), an initiative that has been continued by the Food and Agriculture Organization of the United Nations (FAO) since 1950. FAO (2017a, 2017b) provides countries with a recommended methodology that they can adapt within their own monitoring systems, including identification of essential variables that should be collected to ensure global comparability. The guidance also includes different modes of operation from a traditional census every ten years to a more integrated program of censuses and surveys, where a sample survey is used to collect data during years in between the decennial census, as well as a modular approach, which is used to collect more detailed information on specific areas of interest.

In the European Union (EU), a decennial agricultural census is conducted across Member States (MS) along with a sample survey every 3 to 4 years, referred to as the Farm Structure Survey[1] (FSS). Stipulated by Regulation (EU) 2018/1091 of the

[1] The name originated from the former Regulation 1166/2008 on the European farm survey (The European Commission 2008) and most users are familiar with this term. Current legislation (The European Commission 2018) amended the name to Integrated Farm Statistics (IFS), which is used less frequently, and therefore we opt to refer to it as the FSS in this paper.

European Parliament and of the Council of July 18, 2018, on Integrated Farm Statistics (The European Commission 2018), the data collection in the FSS follows a common methodology to produce comparable and representative statistics across Member States and over time. In addition, EFTA countries Iceland, Switzerland and Norway also participated in the 2020 census, covering more than 9 million farm holdings.

FSS data are used to assess the state of agriculture across the EU, monitoring trends and structural transitions of farms[2]. For example, Neuenfeldt et al (2019) used FSS data to determine the drivers of farm structure change, finding that past farm structure explains the largest amount of variation but other drivers such as environmental conditions, prices, subsidies and income also play a role. The data are also key inputs to the management and evaluation of the Common Agricultural Policy (CAP) in terms of its environmental, economic and social impacts, and as inputs to CAP reforms. In addition to the CAP, FSS data are valuable for other policy areas, including the environment, climate change, employment and regional development (e.g., Copus et al 2006; Einarsson et al 2020).

FSS data are a form of microdata, which refers to any data collected from a respondent in a census or survey (FAO 2017b). Agricultural census and survey data are also complicated as agricultural holdings can contain information related to commercial operations or sensitive personal data. Therefore, the release of census and survey data are subject to confidentiality legislation, stating that data about individuals or enterprises cannot be released or disclosed. Statistical disclosure control is the process by which national statistical offices ensure that any confidentiality legislation is applied (FAO 2017b; Eurostat 2019). Different methods of statistical disclosure are used including table redesign (some table values are aggregated), cell suppression (some values are completely omitted), and adjustment of values using different approaches such as rounding, controlled adjustment (replace values with 'safe' values), and perturbation (random noise is added to values) (Hundepool et al 2010; Fienberg and Jin 2009; European Commission 2021; Templ 2017; Quatember and Hausner 2013).

In the case of the FSS and to ensure that individual farms cannot be identified, the tables are first aggregated to coarse administrative levels (i.e., NUTS2, NUTS1 or even national level depending on the MS) before release by the EU's Statistical Office (Eurostat), curating FSS data for all Member States in the EU. However, there would be considerable value for policy design, policy impact assessment, and scientific research more generally, in having access to data at a finer spatial resolution. Moreover, with advances in technology and the increasing trend to provide open access to government data across many sectors, new methods for disseminating data from censuses and surveys are needed (Shlomo 2018).

Here, we present a methodology (implemented in an *R* package) that takes data collected at individual level, considers a set of confidentiality rules, and produces aggregate values for a multi-resolution spatial grid. The method can also apply a contextual suppression, where some grid cells with few records are suppressed if

[2] Available at: https://ec.europa.eu/eurostat/web/microdata/farm-structure-survey

neighboring grid cells can then be disseminated with a high resolution. We demonstrate the approach using the variables utilized agricultural area (UAA) and organic UAA for synthetic data from Denmark. Such an approach could also be adapted for releasing other individual census and survey-based data that are subject to legal rules of disclosure, or where a certain reliability is demanded for each grid cell. The method has been released in the *R*-package *MRG* on the Comprehensive R Archive Network (CRAN) to make the methodology available to other applications. The functionality has been developed with flexibility so that different restrictions than those relevant to the FSS data can be easily added.

2 Data

To provide a more detailed overview of the European survey on the structure of agricultural holdings, we first present the data collection framework and then describe the detailed topics and variables in the database. We also provide a synthetic data set for the Danish 2020 agricultural census along with the *R*-package, hands-on examples and guidance to produce the maps. In addition, we outline the confidentiality rules and quality assessment of the indicators that are implemented in the methodology that produces a high-resolution grid of the data.

2.1 European surveys on the structure of agricultural holdings

European surveys on the structure of agricultural holdings have been carried out since 1966, and they aim to provide statistical knowledge for the monitoring and evaluation of related policies, in particular the CAP as well as environmental, climate change adaptation and land-use policies. To reduce the burden on national administrations, Regulation (EU) 2018/1091 on integrated farm statistics provides a new framework by distinguishing between core and module variables[3], which vary in frequency and representativeness (The European Commission 2018). It is required that the information on the core variables (e.g., general structural agricultural variables) should cover 98% of the utilized agricultural area and 98% of the livestock units of each MS. The modules contain information on specific topics such as the labor force, animal housing, or irrigation, and can be carried out on samples of agricultural holdings by meeting the precision requirement laid down in Annex V of Regulation (EU) 2018/1091.

[3] The complete list and description of variables surveyed during the European agricultural census 2020 can be found in the Implementing Regulation (EU) 2018/1874 of November 29, 2018, on the data to be provided for 2020 under Regulation (EU) 2018/1091 of the European Parliament and of the Council on integrated farm statistics and repealing Regulations (EC) No 1166/2008 and (EU) No 1337/2011.

2.1.1 The raw survey data

National data providers (i.e., national statistical institutes, ministries of agriculture or other governmental bodies) prepare the questionnaire, conduct the interviews, and complete the survey with additional information from administrative registers (e.g., wine, bovines, integrated information and the control system). The individual records at farm level are encrypted and transmitted to Eurostat via a secure system that implements an automated procedure to validate the content and structure of the microdata.

While an agricultural census is carried out every 10 years, sample surveys are administered during interim years. Table 1 summarizes the data collection for the last decade by highlighting the number of variables, the number of surveyed farms, the population covered and the number of countries participating in the survey rounds. During the 2020 survey campaign, more than 300 variables were collected from around 9.03 million agricultural holdings. In sample survey years such as 2016, 1.69 million agricultural holdings were surveyed, representing approximately 10.55 million farms at that time. It is worth mentioning that the lower sample numbers will give lower accuracy and quality of estimates from sample data compared to the agricultural census, particularly for variables that are nonzero only for a limited number of farms. Therefore, we have also introduced a reliability criterion for the indicators used in the production of the multi-resolution grid data which will also ensure comparability.

2.1.2 Synthetic data

We have derived a synthetic data set (Table 2) from the original 2020 agricultural census microdata which illustrates the methodology and the implicit trade-offs between the spatial resolution and disclosure of information. The data were generated using a hot-deck imputation procedure, which involves replacing missing information with a value from a similar record, known as the donor, within the same classification group as the original record, referred to as the recipient (Andridge and Little 2010; Ford 1983; Joenssen and Bankhofer 2012). Unlike other methods, the

Table 1 Data collection overview of the farm structure survey

Year	Type	Variables	Surveyed farms* (MM)	Population covered* (MM)	Countries
2010	Census	419	12.81	13.03	33
2013	Sample	358	1.73	11.04	30
2016	Sample	363	1.69	10.55	30
2020	Census	364	9.03	9.16	30

Note. * Covers all Member States, candidate, and EFTA countries for the respective data collection year.

Further details about the coverage can be found from the Eurostat-site (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Farm_structure_survey_(FSS))

Table 2 Overview of the variables selected for the synthetic data set

Variable code	Variable description	Type
COUNTRY	ISO code of the country name	Discrete
YEAR	Survey/census year	Continous
REGIONS	NUTS2 region	Discrete
GEO_LCT	Geographical location of the farm	Discrete
HLD_FEF	Main frame* (HLD_FEF=0) or frame extension (HLD_FEF=1)	dummy
STRA_ID_CORE	Stratification ID (any positive integer number). Stratification is mainly applied based on farm type, standard output and land size	Continuous
SAMPLE	Artificial sample created based on stratification (1=farm is included and 0=excluded)	Continuous Binary
EXT_CORE	Extrapolation factor of Core (principally the value is 1, but it can vary according to non-response adjustment or calibration)	Continuous
EXT_MODULE	Extrapolation factor related to the artificial sample	Continuous
FARMTYPE	Typology of the farm. This code list contains the types of farms described by their activities (e.g., raising cattle, cultivating arable crops). Farms are classified into different types according to their dominant activity: crops, livestock and mixed-farming (e.g. FT1_SO)	Discrete
SO_EUR	Classes of standard output (e.g KE_0, KE_LT2, KE2-5,..., KE_GE500)	Discrete
UAA	Utilized agricultural area of the farm	Continuous
*UAAXK*0000*_ORG*	Organic utilized agricultural area, excluding kitchen gardens	Continuous

synthetic data generated by this method contain only plausible values. To assess the quality rating system (i.e., the reliability), we created an artificial sample (*SAMPLE*) with the respective extrapolation factors (*EXT_MODULE*) based on stratification. The sample size consists of approximately one-third of the synthetic 2020 census for Denmark.

2.2 Disclosure control and quality rating

Official statistics are governed by a fundamental principle that protects the confidentiality of individuals or organizations and produces high-quality official statistics by masking sensitive information according to international and European law[4] (The European Commission 2009, 2018; Eurostat 2019; Trewin et al 2007). There is a legally binding obligation to employ appropriate aggregation and disclosure control before making spatial data sets accessible to the public (The European Commission 2018). Furthermore, the Implementing Regulation (EU) 2018/1874 defines a set of rules for disclosing information from European surveys on the structure of agricultural holdings collected at farm locations, including the use of the 1 km INSPIRE Statistical Units Grid for pan-European data. In addition to the standard rules for tabular data, a key requirement is that values can only be disseminated at a 1 km grid when the cell includes more than **ten** agricultural holdings. Alternatively, aggregating to a nested 5 km or larger grid size is required to satisfy the aforementioned requirement (The European Commission 2020).

A disclosure occurs when an intruder correctly finds or determines some values about an individual or organization from the data released. Duncan and Lambert (1989) differentiate between two types of disclosure risk: identity disclosure and attribute disclosure. While the former occurs when a record can be directly linked to an individual, the latter refers to the knowledge gained about an individual or organization from the attribute(s) in the data released. Statistical disclosure control (SDC) techniques are widely deployed to reduce the risk of disclosing private information at an acceptable level, while maximizing the utility of the data (Quatember and Hausner 2013; Templ 2017). From the two broad families of methods that exist, the perturbative method modifies the data prior to publication by adding random noise such as rounding to the nearest multiple of ten. Non-perturbative techniques reduce the amount of information by suppressing or aggregating the data. The optimal mixture of SDC should strike a balance between the mandatory privacy protection of the statistical output and the accessibility to the data at the highest available spatial resolution (Quatember and Hausner 2013).

In agreement with member states, Eurostat (2020) has provided a series of recommendations in the confidentiality charter for disclosure control. For the dissemination of aggregated tabular statistics, values must comply with the threshold[5] and

[4] Separate national laws (EU/EEA/EFTA) might contain stricter (or laxer) rules related to the disclosure of personal information.

[5] Suppression of cells representing less than four agricultural holdings.

dominance rule[6], and the statistical output must satisfy a quality criteria[7]. The quality criteria is only used for sample years, when stratified sampling and the use of extrapolation weights will introduce estimation errors. These rules also apply to the dissemination of gridded data.

Another important aspect that is receiving increasing attention is second-order confidentiality, which occurs when the value of a suppressed sensitive cell can be determined from neighboring cells or from other publicly available sources. In terms of gridded data, it is possible that cells become identifiable when both high-resolution gridded data and low-resolution NUTS data are published. Applying gap-filling methods to both data sets to impute the suppressed values would put the disclosure of private information at risk (Higgins and Scheiter 2012). This threat can be overcome by carefully choosing the size of the grid cells and the type of administrative regions for the dissemination of the data.

3 Methods

3.1 Multi-grid approach

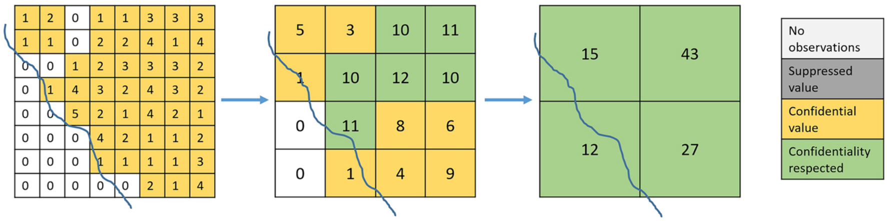
Several different methods can, with different advantages and disadvantages, be used to create a gridded data set that respects the confidentiality rules above. We will focus on grids with different resolutions, presented in the next subsections.

3.1.1 Gridding

From a point data set like the FSS data, an unlimited number of regular grids can be created. For the methods below, we first need to create a set of base grids of different resolutions. The first three resolutions are specifically mentioned in the EU regulations, which require these to be 1, 5 and 10 km (The European Commission 2020). There are no restrictions on coarser resolution grids. However, the methodologies below require the grids to have a hierarchical structure where the coarser resolution grids must be integer multiples of the higher-resolution grids. Coarser resolution grids could be 10, 20, 40, 80, and 160 km, or 10, 50, and 100 km. However, 10, 20, 50, and 100 km would not be possible as the 50 km grid includes 2.5 grid cells (in each direction) from the 20 km grid and is not an integer multiple. For geographical data, such as the FSS, it is recommended to use a projection with equal area properties to ensure that grid cells represent the same area across the entire grid. The FSS data are currently provided in the Lambert Azimuthal Equal Area projection (EPSG:3035). The predefined origin of the grid

[6] Suppression of cells when one or two contributors are dominant.

[7] The prediction errors can be estimated as a function of the sample size, population size, sampled values and possible stratification. The Integrated Farm Statistics Manual (Eurostat 2023, Section 4.6) requires that the relative standard error (coefficient of variation) of the estimate should be less than 0.35, otherwise the value is suppressed.



1	2	0	1	1	3	3	3
1	1	0	2	2	4	1	4
0	0	1	2	3	3	3	2
0	1	4	3	2	4	3	2
0	0	5	2	1	4	2	1
0	0	0	4	2	1	1	2
0	0	0	1	1	1	1	3
0	0	0	0	0	2	1	4

5	3	10	11
1	10	12	10
0	11	8	6
0	1	4	9

15	43
12	27

Fig. 1 Example of gridded data, moving from a higher to a coarser resolution. The numbers represent the number of farms per grid cell. The line represents a border or coastline

in the code coincides with the false origin of the ETRS89-LAEA coordinate reference system (x=0, y=0) as specified in the Commission Regulation 1089/2010 on the interoperability of spatial data sets and services (The European Commission 2010). Also other projections can be used, and the base resolutions can be of any size, as long as they are integer multiples of each other.

If the data are to be disseminated as a regular grid, the confidentiality rules must be examined for each different grid level, and the final resolution will be the highest resolution at which the confidentiality rules are respected for all grid cells. This method is intuitive and will work well when the data are fairly well distributed over the domain of interest. However, if the density is considerably lower in some regions, then the resolution will need to be coarse for the entire data set.

Figure 1 shows a fictitious example in which the number of farms (numbers in each grid cell) have first been aggregated in 2*2 blocks to a lower-resolution grid. Assuming that ten farms are necessary for disclosing the information from a grid cell, none of the grid cells in the left panel will pass the confidentiality rules. In the second grid, 2*2 blocks of the original cells have been aggregated to larger cells. Here, the green cells respect the confidentiality rules, but the yellow do not, so this grid cannot be disclosed either. In the right panel, all grid cells respect the confidentiality rules. However, the data in the upper right grid cell have been aggregated to a coarser resolution than necessary, rendering this solution as suboptimal.

3.1.2 Value suppression

Another relatively simple approach is to suppress the values from grid cells where the confidentiality rules are not respected for the selected resolution. Figure 2 shows an example using the same fictitious data. The left panel shows a situation where all grid cells would be suppressed so no data could be released. Some grid cells are non-confidential and do not need to be suppressed in the central panel with lower resolution. There is no need to suppress any grid cells in the right panel of Fig. 2. This method can lead to a large number of empty cells if there is a significant difference in data density across regions. A major issue is that the total sum of farms will be lower than the actual number due to removed values.

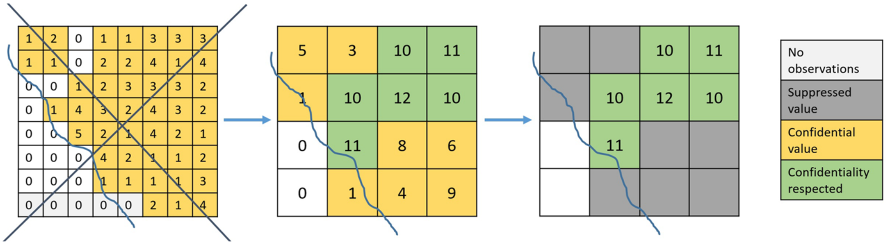


Fig. 2 Example of suppression of grid cell values that do not respect the confidentiality rules. The highest resolution grid cannot be used, as all values would have been suppressed with a limit of 10

3.1.3 Multi-resolution grid

A second option is to disclose information with a variable grid size, also referred to as multi-resolution grid or quadtree (Asim et al 2023; Behnisch et al 2013; Eurostat 2020; Lagonigro et al 2020). The idea here is that the resolution of the grid will vary according to the local density of the observations, and to ensure that the confidentiality rules are respected for all grid cells. An example of this is shown in Fig. 3 with the same fictitious data as above. However, when reducing the resolution toward the right panel, the four cells in the upper right corner are not aggregated, as they all have more than ten farms. Hence, it is possible to share the data with a higher resolution in this area and at a coarser resolution in the rest of the map.

The method is sensitive to islands and borders, where it might be difficult to include a sufficient number of farms, when a large part of aggregated grid cells do not include data. A general solution is to aggregate only up to a certain grid size, and then suppress grid cells that still do not respect the confidentiality rules.

There is also a second option for suppressing values. If a grid cell does not respect the confidentiality rules, it should be aggregated if most of the neighboring grid cells are also confidential. However, it is less optimal if a single confidential grid cell causes aggregation of many non-confidential grid cells. The package therefore has a suppression limit argument, where a confidential grid cell will only cause aggregation if its share of the value of a possible aggregated grid cell is above the limit. If the limit is 0.1, the grid cell with 1 in the lower left quadrant would not cause aggregation in Fig. 3, representing less than 10% of the

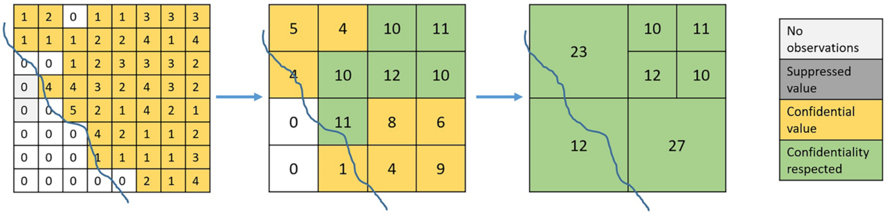


Fig. 3 Example of a multi-resolution grid, moving from a higher to a coarser resolution. The numbers represent the number of farms per grid cell. The line represents a border or coastline

value of the possible aggregated grid cell. Instead, it will be suppressed in the post-processing step.

The list below and Fig. 4 show the iterative process of producing a nested structure of multi-hierarchical grids satisfying a set of confidentiality rules and quality requirements. We denote the level of resolution $k \in K$ with $K = \{k_0, k_1, \ldots, k_m\}$ where k_0 is the highest resolution (1 km for FSS) and k_m the lowest resolution. The first iteration is i_1, the possible aggregation from k_0 to k_1 and continues until reaching the maximum level k_m ($i \in \{i_1, \ldots, i_m\}$). For each iteration, the following steps are evaluated (it is sufficient to pass one of the dominance rules):

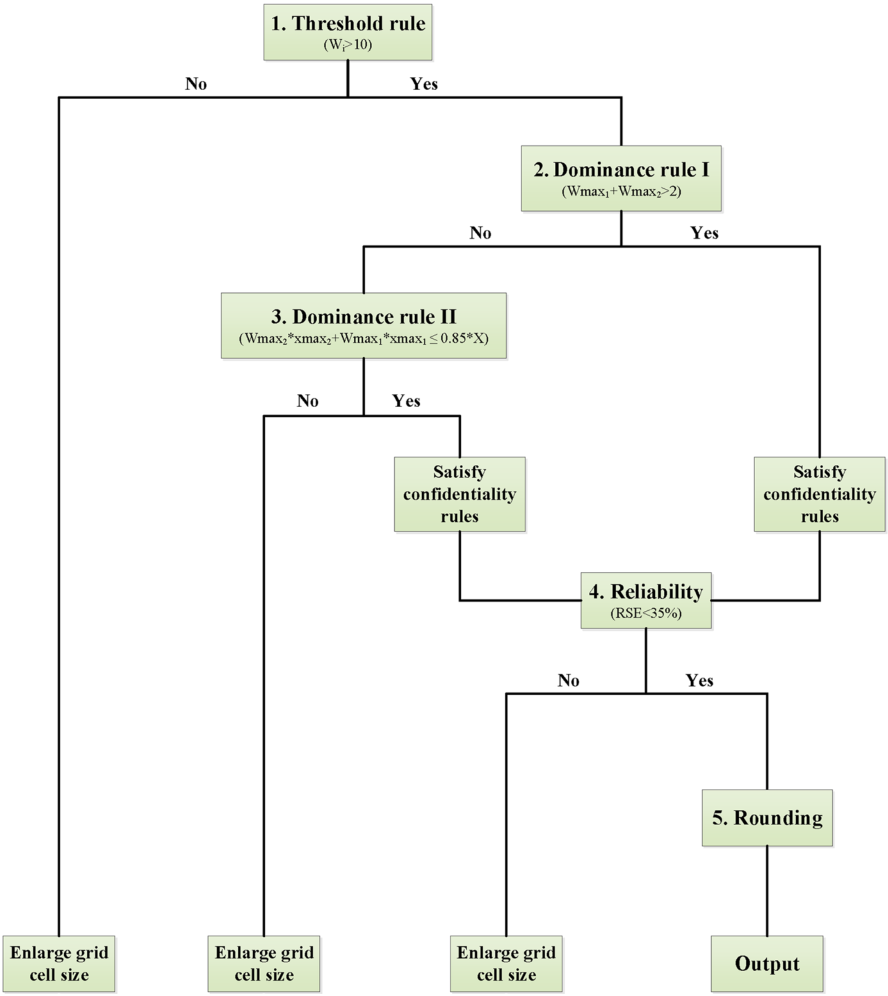


Fig. 4 Flowchart showing the rules that are applied for the release of Farm Structure Survey (FSS) data. Where the rules are not satisfied, the grid cell sizes must be increased in the next iteration, (unless their impact is smaller than a certain limit)

(i) Threshold rule: Is the aggregated extrapolated number of farms in grid cell l (W_l) for resolution k_i equal or more than ten ($W_l \geq 10$ with $W_l = \sum_{j=1}^{n_l} w_j$; n_l is the number of records in l).

(ii) Dominance rule I: This rule is satisfied if, after ordering the variable of interest in descending order, the sum of the weights (w_{jmax_1} and w_{jmax_2}) of the two highest values (x_{jmax_1} and x_{jmax_2}) is greater than two ($w_{jmax_1} + w_{jmax_2} > 2$). (The weights in FSS are rounded before this step, so larger than 2 means at least 3.)

(iii) Dominance rule II: If the weighted sum of the two potential dominant contributors are less than or equal to 85% of the extrapolated aggregated value (X) of the grid cell ($w_{jmax_2} \times x_{jmax_2} + w_{jmax_1} \times x_{jmax_1} \leq 0.85 \times X$), then the confidentiality rules are satisfied.

(iv) Reliability of the results: The indicator is reliable if the estimated coefficient of variation for the grid cell at k_i is less than 35%, (will be disseminated with a warning if above 25%);

(v) Repeat: If not passing the threshold rule, the quality rule (if applicable), or one of the dominance rules, the grid cell will be aggregated with neighboring grid cells, and the steps above will be repeated (unless the value is below the suppression limit).

After the last iteration, and as a measure to add further perturbation to the disclosed information, all non-confidential extrapolated number of farms and extrapolated aggregated values of variables are rounded to the first significant digit if this digit is ≥ 3 and to the first two digits otherwise.

3.2 Implementation of the approach

The method has been implemented as a package in the environment *R* (R Core Team 2024). The package contains functions to create grids respecting the confidentiality rules when releasing survey/census data. Some similar functionality (i.e., the multi-resolution grid) is implemented in the packages *AQuadtree* (Lagonigro et al 2020) and *sdcSpatial* (de Jonge and de Wolf 2022). However, these packages lack the flexibility to include the dominance rule, and lack several other features in this package.

The functionality in the *MRG* package uses methods from the *sf* package (Pebesma 2018; Pebesma and Bivand 2023) and spatial analysis functionality from the packages *stars* (Pebesma and Bivand 2023), *terra* (Hijmans 2023) and *vardpoor* (Breidaks et al 2020). The gridding procedure, which is shown in Fig. 5, can be applied by calling the functions in the following subsections.

The processing time and the memory load will to a large degree depend on the problem at hand. However, some simple examples can indicate the computational burden, as shown in Table 3. All estimates in the table are from a Windows server with 64 GB RAM, without applying parallelization. Using a server is not always faster, an ordinary Windows desktop computer has been faster for some test cases.

The initial gridding can be very memory intensive. One of the intermediate objects depends on the bounding box itself, and this will create particularly large grids if European overseas territories are included. This size will be reduced with

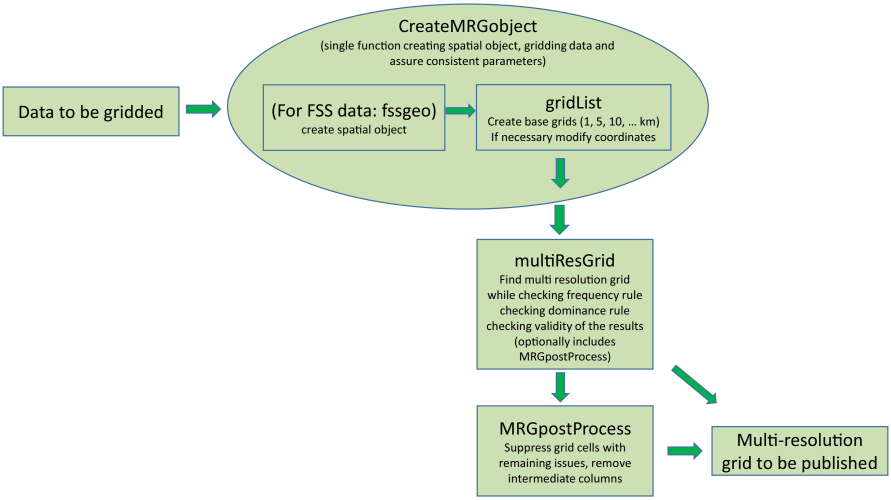

Data to be gridded
CreateMRGobject
(single function creating spatial object, gridding data and assure consistent parameters)
(For FSS data: fssgeo)
create spatial object
gridList
Create base grids (1, 5, 10, … km)
If necessary modify coordinates
multiResGrid
Find multi resolution grid
while checking frequency rule
checking dominance rule
checking validity of the results
(optionally includes MRGpostProcess)
MRGpostProcess
Suppress grid cells with remaining issues, remove intermediate columns
Multi-resolution grid to be published

Fig. 5 Flowchart of the procedure to produce the multi-resolution grid

Table 3 Data collection overview of the farm structure survey

Example	Sample size	Extent	Max memory	Time base gridding	Time MRG
1	9,030,000	Including oversea	3.5 GB	165 min	80 min
2	8,975,000	Continental Europe	1.3 GB	180 min	80 min
3	2,000,000	Continental Europe	0.6 GB	70 min	8 min

a smaller overall area. The size is also somewhat affected by the number of observations, mainly because more records will lead to fewer nonzero grid cells for the highest resolution. The initial gridding process is also the most time-consuming part for large data sets. This mainly depends on the number of records. The creation of multi-resolution grids is faster, after the initial gridding is done. This also depends somewhat on the number of records.

The initial gridding can be parallelized, although only to a certain extent. Memory consumption could be a problem, as many temporary objects have to be created in parallel. The third example in Table 3 took 10 minutes with three nodes on a Windows PC, whereas the time increased to 14 minutes with 6 nodes, due to the increased overhead.

The package is released on the Comprehensive R Archive Network (CRAN: https://cran.r-project.org/) under the GPL(>=3) License.

3.2.1 Prepare the data

The procedure starts with a *data.frame* with the variable(s) to be gridded, typically imported from a common file format (as a csv or Excel). This data set has to be converted to a spatial *sf*-type object (Pebesma 2018). The spatial information in the FSS data has a special format, encoded as a string with country name, the coordinate reference system, the resolution and the coordinates. The *fssgeo*-function will parse this string and create the spatial object to be used in further analyses. For other types of census and survey data, users will need to create an *sf*-object themselves.

3.2.2 Create a base grid

The procedure first needs a hierarchical set of grids with different resolutions, either using the *gridList* function or the *createMRGobject* function. The difference between them is that the first function will only create a list of the grids, whereas the second function will create an object that also includes resolutions, variable names, weight names, and parameters for the confidentiality rules in the object. This can then be used as an input to *multiResGrid*, instead of having to specify all the different parameters every time, and will for most users be more convenient. This will together ensure that the same data set and values are used consistently throughout the entire procedure.

The default resolutions (of 1, 5, 10, 20, 40, 80 km) follow the regulations up to 10 km, and aggregates 2*2 grid cells for lower resolutions, but the user can change these. If only the presence of a variable is of interest (such as number of farms), then no variable name is required for this function. For all other variables (such as the UAA, the number of livestock, etc.), the column name(s) should be a parameter of this function. If there is a weight associated with the variable (if some observations in the data are samples from a larger group), the column name(s) with the weights should also be added. If only one weight column is given, this will be applied to all variables. Otherwise, one weight should be given for each variable.

Problems will arise if the observations fall exactly on the border between grid cells, as it will not be clear which cell they belong to. This is the case for FSS data, where the coordinates have been mapped to the corners of a 1 km grid. If using *gridList*, the coordinates should be adjusted before calling the function (for example, with *st_jitter* or *locAdjFun*). The same modified coordinates will then also have to be passed as a parameter to the *multiResGrid* function. If creating an *MRG*-object, the adjustment of the coordinates can instead be done through a variable *locAdj*. The function will then shift the locations away from the borders. The value can either be one of *LL* (lower left, and default for FSS data), *LR* (lower right), *UL* (upper left), or *UR* (upper right), where the value refers to where in a grid cell the coordinate is located. Alternatively, the value can be *jitter*, where the function will add a jitter to the coordinates, randomly distributing the records to either side of a border.

3.2.3 Create multi-resolution grid

The multi-resolution grid is created using the function *multiResGrid*. Most of the parameters for this function were mentioned in the methods section, with default values reflecting the standards for the FSS data. The parameters should be included as a part of the *MRG*-object if *createMRGobject* was used above. Some of the parameters not already mentioned include:

- the choice of whether to apply the confidence rules to all variables individually (*confrules* = *"individual"*), or only look at the first variable,
- the parameter *suppresslim*, which indicates if grid cells with a value less than the *suppresslim* share of an aggregated grid cell should be left unaggregated, and rather be suppressed at a later stage,
- the possibility to add another function (*userfun*) that tests other criteria for other applications of the method (see more details below), and
- a logical variable *postProcess*, which indicates if post-processing should already take place in this function (the default is *TRUE*), or if the user wants to examine the raw data before post-processing in a separate function.

3.2.4 Post-processing

If not done as part of the *multiResGrid* function, run the *MRGpostProcessing* function, which will check that all grid cells respect the confidentiality rules and suppress values from those cells that do not. This function will also round the variables according to the rounding rule, where the default is rounding to the first significant digit if this digit is ≥ 3 and to the first two digits otherwise.

3.2.5 Joint aggregation or merging of multi-resolution grids

Some indicators can be a function of two or more variables, and these require a common grid. An example is the ratio between a variable and the UAA of a grid cell. The procedure can create a common multi-resolution grid for several variables, assuring that all of them respect the confidentiality rules for each grid cell. If one grid has already been created and cannot be recomputed, it is possible to create a similar grid of the second variable by passing the first grid as a parameter to the *multiResGrid*-function (keeping the resolution of the first variable), or to create a second grid and merge them with *MRGmerge*. *multiResGrid* works well if the density of the second variable is similar or higher than the first variable, but will create a grid with a high number of suppressed values of the density is lower.

3.2.6 Other features

There is also an additional feature in the package, where a user-defined function *userFun* can be passed to the gridding function. This is useful if a user wants different rules than the ones implemented. An example for the FSS data could be to restrict some of the variables in a grid cell to a maximum value. FSS data are reported at the administrative location, which in some cases can be in a municipality center rather than the location of the parcels. In those cases, the *userFUn* could flag a grid cell for aggregation if the total UAA of the grid cell is more than the area of the grid cell itself. The input of the function must be similar to the functions for the confidentiality rules. Further details are given in the help file for the multi-resolution grid function.

The package also includes a *print* method for *MRG*-objects and a plotting function (*MRGplot*), based on *ggplot2*. This function makes it easier to visualize the multi-resolution grids, whereas *ggplot2* is still necessary for more advanced plotting.

4 Results

A synthetic data set for Denmark (included in the package) was used to demonstrate the procedure, representing an agricultural census like the FSS. A subset was used to represent an agricultural survey, collected in between census years. The procedure was then applied to actual UAA data from the 2020 FSS for entire Europe.

4.1 Gridded FSS data

A hierarchical set of gridded values is the base for the multi-resolution grid, and was created with the function *gridList*, after *fssgeo* created a spatial object and modified the coordinates so they are not exactly on the grid lines. Figure 6 shows the number of farms per grid cell for different grid cell sizes for the synthetic data set: 1, 5, 10, 20, 40 and 80 km. First, we only use the frequency rule (i.e., minimum of ten farms in a grid cell) as the confidentiality rule. We cannot see much in the 1 km grid, but none of the grid cells reach the confidentiality limit of ten farms. It is still difficult to visualize the individual grid cells in the 5 km grid, but approximately 80% of the 2000 grid cells have more than ten farms. For the 10, 20, 40 and 80 km grids, there are 52, 11, 1 and 0 grid cells with less than ten farms, respectively, meaning that the 80 km grid is the highest resolution at which it is not necessary to suppress any values.

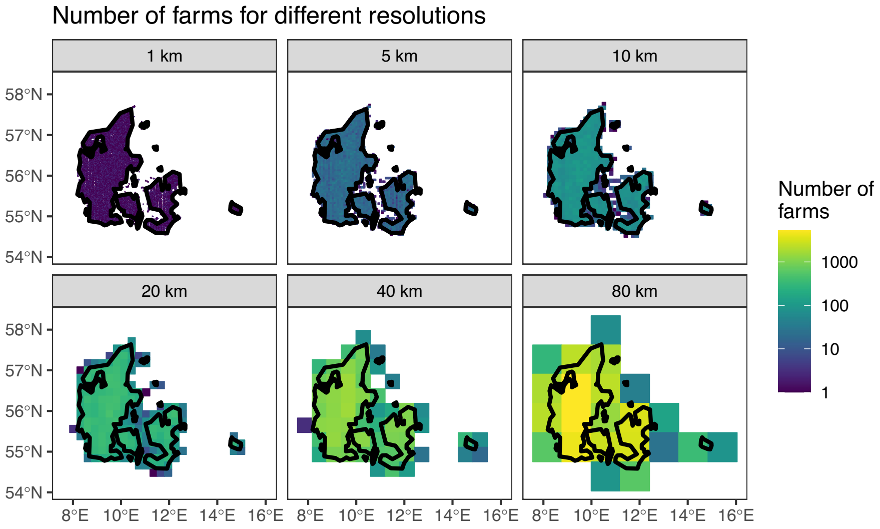


Fig. 6 Number of farms per grid cell for different grid cell sizes for Denmark (synthetic data)

It is the coastal grid cells that make it necessary to continue aggregating until the grid cells are 80 km. Both in the 40 and 80 km grids, some of the coastal grid cells are partly or mainly in the sea, resulting in the need to aggregate to the same resolution for the entire data set. This example was presented with the entire grid cells to illustrate the border effect. However, in the rest of this section, we will clip the grid cells to the coastlines, which gives a better visual representation of the agricultural activity.

4.2 Multi-resolution grids of FSS data

4.2.1 Gridded farm density

With the set of grids from the previous step, we can run the *multiResGrid* function to create a multi-resolution grid, only using the threshold rule (at least ten farms) in the first example. If the observations are added to the gridding procedure, it is also possible to apply the dominance rule, in this case based on the UAA. The dominance rule is used by default if the observations are provided.

Figure 7 shows the multi-resolution grids for the synthetic data for Denmark. In the top left panel, only the threshold rule was applied (minimum ten farms). The

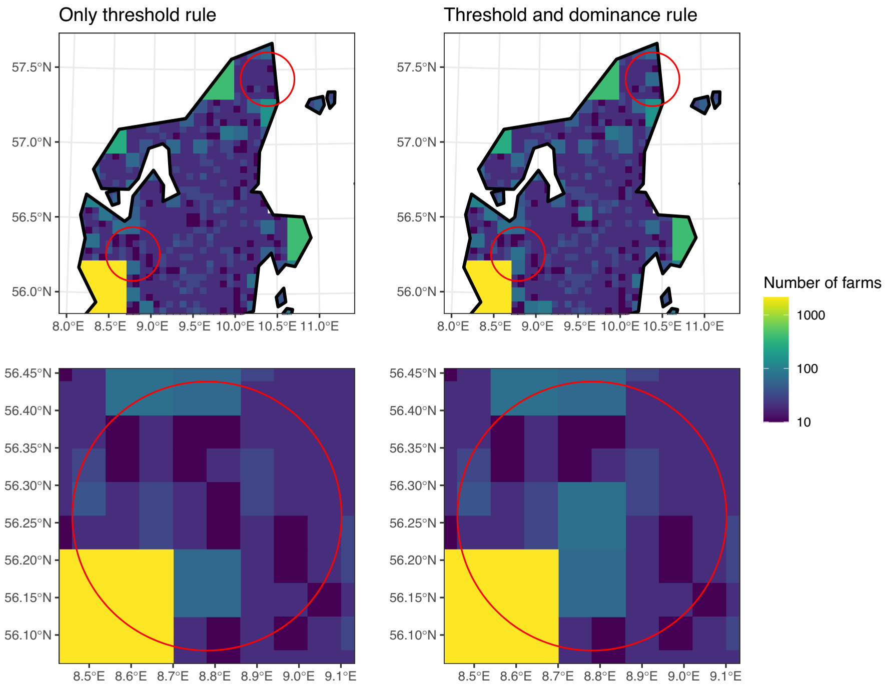


Fig. 7 Number of farms per grid cell for different grid cell sizes for Denmark (based on synthetic data) with different confidentiality rules employed. The lower two panels are zoomed in on the bottom left circle

majority of the grid cells (1186) have a resolution of 5 km but there are also 144 with a resolution of 10 km, 27 with 20 km, 8 with 40 km, and 1 with 80 km. It is challenging to see the 5 km grid cells, but we can observe that most of the larger grid cells are on the coastline. Most of the grid cells have 10–50 farms, but there are 19 grid cells with more than 100 farms, and one of them with 2080 farms.

The panel on the top right side of Fig. 7 shows the result when the dominance rule is also applied. The difference between the two is small in this case. There are 42 fewer 5 km cells, 7 more 10 km cell, and 2 more 20 km cells. This difference is caused by some large farm farms/producers in the grid cells that had to be aggregated. Some differences can be noticed inside the circles. The two lower panels are zoomed in on the bottom left circle, showing how four 5km grid cells become a single 10km grid cell.

4.2.2 UAA and organic UAA

The example above only looked at the number of farms, but gridded farm variables will be of more interest. Two examples are the UAA and the organic UAA. These results are shown in Fig. 8. First, one can notice that the UAA depends on the grid cell size, which is typical for variables that are summed. An alternative would be to present the UAA as UAA/km^2 for each grid cell. Second, the grid cells are the same size as the left panel in Fig. 7 because this is just another variable from the same underlying input data.

The map of the organic UAA (Right in Fig. 8) differs, with much larger grid cells. This is because there are considerably fewer farms with organic farming. Only one grid cell can be disseminated at 5 km, whereas the majority are 10 km (91) or 20 km (67). Then, there are 18, 2, and 1 grid cells of 40 km, 80 km, and 160 km, respectively. There are totally 180 grid cells in this map with organic farms.

4.2.3 Suppressing insignificant grid cells

The parameter *suppresslim* can be used to suppress some grid cells instead of aggregating, as described in Sect. 3.2.3. Figure 9 shows the effect for different values of *suppresslim*. When *suppresslim* = *0* (upper left panel of Fig. 9), there are some large grid cells marked with circles that disappear in the following panels (as the value

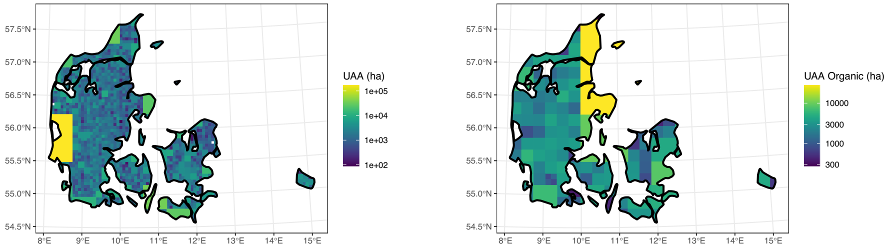


Fig. 8 Utilized agricultural area (UAA) (left) and organic UAA (right) per grid cell for Denmark (synthetic data)

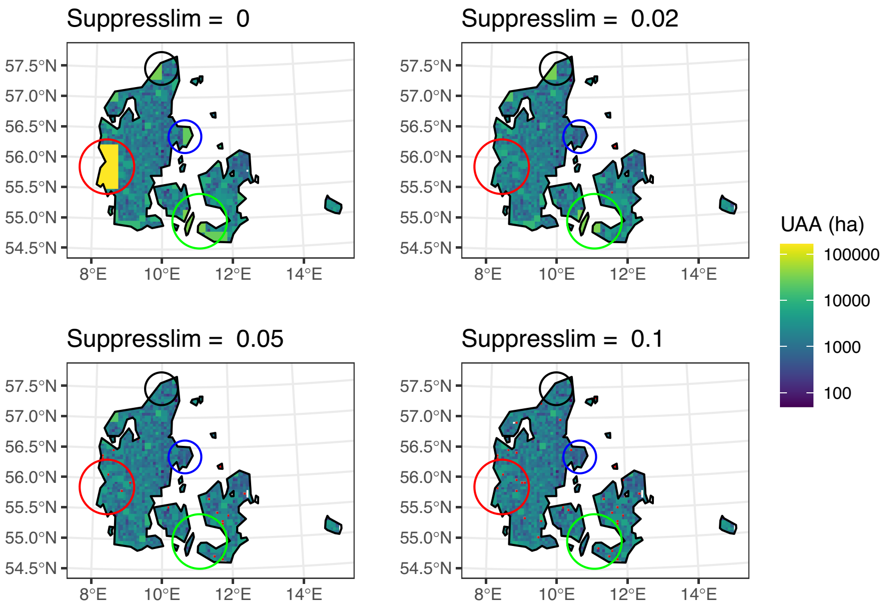


Fig. 9 Utilized agricultural area (UAA) per grid cell for different grid cell sizes for Denmark for different values of *suppresslim*. Suppressed grid cells are shown in red. The circles highlight regions mentioned in the text

provided to the *suppresslim* function increases). Large grid cells inside the red and blue circles disappear already with *suppresslim = 0.02*. The ones in the black and green circles disappear with *suppresslim = 0.05*, and the grid cells within the green circle are further reduced in size for *suppresslim = 0.1*. The suppressed grid cells (red squares) are barely visible for the lowest value of *suppresslim*, whereas there are considerably more (and larger) grid cells suppressed for *suppresslim = 0.1*. Table 4 shows grid cell sizes for different values of *suppresslim* (including *suppresslim = 0.2*). The numbers in brackets shows the number of suppressed grid cells for each resolution.

As the value of *suppresslim* increases, the number of large grid cells decreases. For example, the largest grid cell for *suppresslim = 0* is 40 km, whereas 20 km is

Table 4 Distribution of grid cell sizes for different values of *suppresslim* with the number of suppressed grid cells in brackets

Resolution (km)	*suppresslim*				
	0	0.02	0.05	0.1	0.2
5	1144 (0)	1337 (18)	1459 (47)	1591 (90)	1774 (165)
10	151 (0)	178 (3)	170 (5)	147 (6)	101 (7)
20	29 (0)	18 (0)	15 (1)	10 (1)	7 (2)
40	8 (0)	5 (0)	2 (0)	1 (0)	0 (0)
80	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)

the largest for *suppresslim = 0.2*. At the same time, the number of smaller grid cells increases considerably. There are 1774 grid cells of 5 km for *suppresslim = 0.2*, whereas there are 1144 for *suppresslim = 0*. The number of large grid cells (20 and 40 km) go down from 29 and 8, respectively, to 7 and 0. However, increasing values of suppresslim also leads to suppression of an increasing number of grid cells, in most cases small ones. The total number of suppressed grid cells are 21, 51, 97 and 174, respectively, for the different values in the table. If we look at the percentage of farms and UAA that are not part of the final map, this ranges from 0.3% - 3.7% of the total number of farms, and 0.001% – 2.5% of the total UAA, with the highest values for *suppresslim = 0.2*.

4.2.4 Demonstrating the need for reliability checks

Here, we show the effect of the reliability checks when gridding survey data as opposed to census data. For survey data, collected in a stratified approach, weights are assigned to each stratum based on subsampling rates. If a record has a high weight, confidentiality rules are met, but the value may not be reliable.

The results are illustrated in Fig. 10, which shows four maps of gridded synthetic data from Denmark, a subset of the data set above. The top panels of Fig. 10 show the maps without reliability, whereas it was included in the procedure for the bottom panels. The left panels show the number of records (the actual number of farms in the survey), whereas the right panels show the weighted number of farms. The tiny

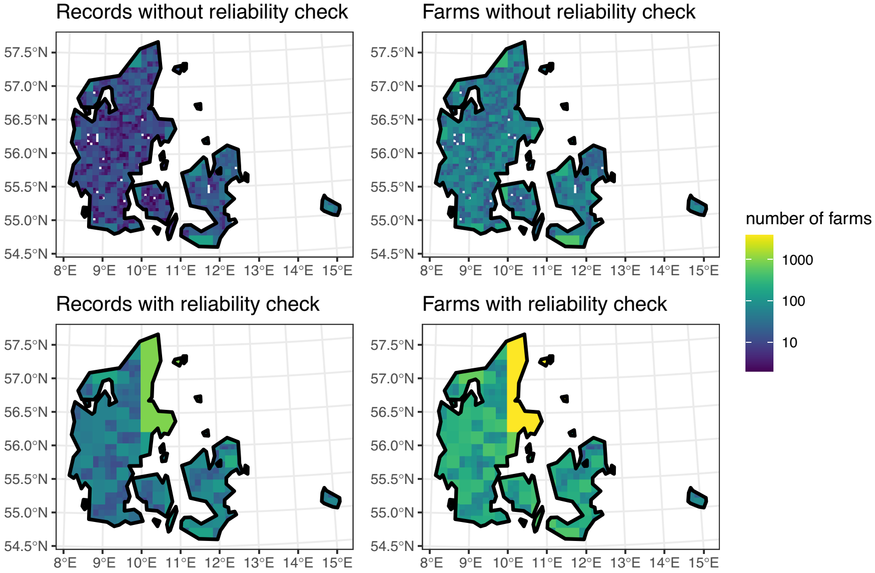

Records without reliability check
Farms without reliability check
Records with reliability check
Farms with reliability check
57.5°N
57.0°N
56.5°N
56.0°N
55.5°N
55.0°N
54.5°N
8°E
9°E
10°E
11°E
12°E
13°E
14°E
15°E
number of farms
1000
100
10

Fig. 10 Multi-resolution grid of the number of farms for a synthetic Farm Structure Survey (FSS) data set for Denmark, with the reliability check on the bottom and without on the top. Number of records refers to the actual farms in the survey, whereas the number of farms refers to the weighted number of farms

grid cells that contain just a few farms with large weights have mostly disappeared, producing a smoother and more realistic map. The result is considerably fewer grid cells, based on more records. Only six grid cells have less than ten records, with three records as the fewest. Note that the reliability check is applied as an integrated part of the iterative process, but it is not applied by default due to its computational burden.

4.3 An example of producing a ratio

Figure 11 shows the gridded total UAA and gridded organic UAA in the upper panels, together with the gridded organic share in the lower panel. The variables have been gridded jointly, and we can see that the grid cells are the same for both of them. The gridding procedure was done with *suppresslim = 0.05*, which resulted in the suppression of two grid cells in the southern part of Denmark and on the island to the east. Using the synthetic data, we can see that the concentration of organic farming is higher in the south of the country, although this pattern may differ when actual data from the agricultural census are used.

4.4 Producing European-wide estimates

The overall aim is to disseminate agricultural variables at a European scale. Since the resolution of a European-wide grid cannot be conveyed with sufficient detail

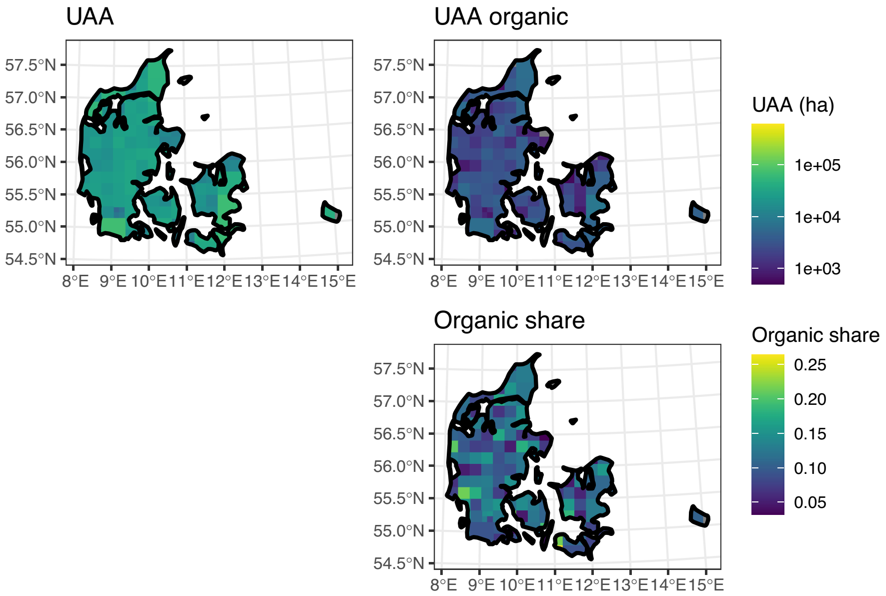


Fig. 11 Utilized agricultural area (UAA) and organic UAA (ha per grid cell) for Denmark (synthetic data), and the ratio between the two (share)

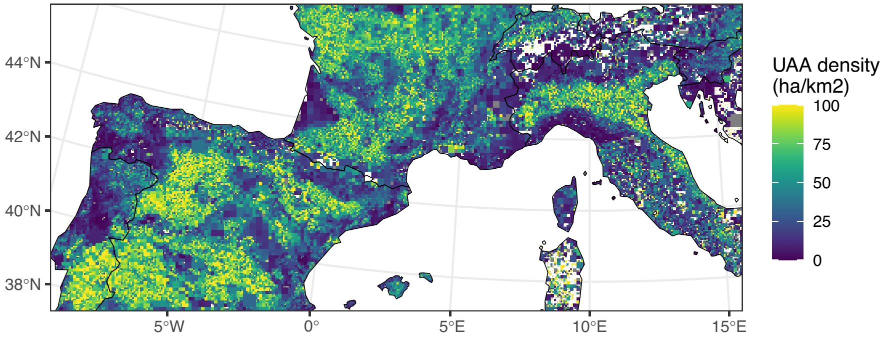


Fig. 12 Utilized agricultural area (UAA) density per grid cell (ha/km^2) for a region in southwest Europe, based on 2020 Farm Structure Survey (FSS) data. Suppressed grid cells are shown in gray, and white grid cells have no farms

Table 5 Distribution of grid cell sizes for the region in Fig. 12

Resolution (km)	Total number of grid cells	Suppressed grid cells	Non-confidential grid cells
1	7445	3478	3967
5	26108	490	25618
10	4824	92	4732
20	355	5	350
40	17	2	15

in a figure, we provide an example of the density of UAA per grid cell in Fig. 12 for a region in South-Western Europe, based on the 2020 census data from the FSS. It shows the recorded hectares per km^2. This can be seen as a surrogate for the percentage of agricultural land in a grid cell (there are 100 ha in one km^2). In reality the number of hectares recorded in FSS can be higher, especially for smaller grid cells, as all the agricultural land of a farm is recorded at its administrative location, used for the gridding. The color scale has been limited to 100, and the effect of administrative boundaries can somehow be noticed through the “salt-and-pepper” like distribution of colors in the most densely cultivated regions.

We can see how the grid cell sizes vary between different regions. Mostly areas with a high density of farms also have high-resolution grid cells. This is the case for the Po Valley in Italy. In France, we can recognize some of the areas for wine and distilled alcohol (Bordeaux, Loire, Armagnac, Cognac), in Portugal the Antelejo region, and in Spain Castile-León, Castile???La Mancha and Extremadura. Larger grid cells can be found in regions with lower density of agriculture, such as in high mountains (the Alpine region traversing France, Italy and Switzerland, and mountainous regions in Spain and Portugal) and larger forested areas (and mostly hilly regions) along the Mediterranean coast of Spain, France, and Italy.

Table 5 gives an overview of the grid cells in this image. The majority of the grid cells are 1 or 5 km, whereas there are also many with a size of 10 and 20 km. There are considerably fewer that are 40 or 80 km.

5 Discussion

The multi-resolution gridded solution presented here represents a step change in the way that the rich amount of information on the farming sector in Europe, collected by EU Member States and Eurostat in agricultural censuses and surveys, could be released in the future. Using this approach, the information content is maximized and released at the locally highest resolution possible while respecting the confidentiality regulations as specified in EU laws as well as guidelines set by Eurostat and agreed with the Member States. In contrast, other countries outside of the EU are still much stricter in their dissemination of agricultural census data. For example, the US Department of Agriculture releases data at county level, which is similar to NUTS2 regions in Europe (USDA NASS 2024). In Canada, one-third of data were not disclosed in the 2016 agricultural census, which employed suppression of data. For the 2021 Census, Statistics Canada has switched to the use of random tabular adjustment, which makes changes to individual cells to ensure data protection (Statistics Canada 2023b). However, the size of the areas for which data are released must be a minimum of 25 square km in area and contain more than 16 farms or the areas are merged with adjacent zones (Statistics Canada 2023a). Moreover, comparability of the 2021 agricultural census data with previous censuses will be impacted (Statistics Canada 2023b). In the UK, the Edinburgh Data and Information Access (EDINA) releases agricultural census data at 2, 5, and 10 km grids (Macdonald 2004). However, with a single grid size, the data are less reliable and/or suppressed in areas where the disclosure requirements are not met (Khan et al 2013). Hence, the suggested approach could be used and adapted by other statistical services that disseminate agricultural census and survey data (such as farm accountancy data) to meet their specific disclosure requirements. Given the versatile and flexible implementation of our approach, the methodology could easily be expanded to other statistical domains where sensitive information on individuals or enterprises is collected, such as population, migration, business, and labor force statistics.

However, there are also limitations with the multi-resolution gridded approach. The examples provided in the paper were for continuous variables. Categorical variables such as farm type, irrigation methods, or other gainful activities will require transformation into dummy variables, and these classes will need to be treated individually. Secondly, the reliability check demonstrated in the Sect. 4.2.4 is integrated into the iterative process that produces the multi-resolution grid but it is not applied by default as this process is computationally time-consuming. Finally, creating multi-resolution grids of a ratio requires a different calculation for the estimation of variance in the reliability check, which is currently neglected. We hope to address this in a future upgrade of the package. In the mean time, the recommendation would be to use a lower reliability limit for variables that are to be used in a ratio computation, as the CV of the ratio is higher than the CV of each of the variables.

It is also important to consider specific effects that can occur when working with certain types of data. In this article, we used areal data (agricultural parcels, main buildings of the farm, location of the main agricultural activity) that was first summarized as point data before being aggregated into grids. However, this process can introduce uncertainty, as entire parcels or parts of parcels may be attributed to a point in a neighboring grid cell. As a result, the gridded values may not accurately reflect the true distribution of the data. For instance, it is possible for the utilized agricultural area (UAA) in a grid cell to exceed the size of the grid cell itself. It is essential to note that this issue arises from the conversion of areal values to point values, rather than the gridding process per se. Assessing the uncertainty associated with this process is challenging, and out of scope for this study, but we plan to investigate these effects in more detail in a forthcoming project.

Zero values can be interpreted in slightly different ways depending on how they occur. A zero value for a variable means that the variable was not observed in that grid cell. However, if the variable is usually only occurring in a share of the farms, it will not be clear if this for example means zero organic farms out of zero farms in total, or if it means that there are zero organic but potentially several conventional farms (sometimes referred to as true zero). How to deal with this will depend on the ones who are sharing the data.

A potential issue with second-order confidentiality may arise when regional multi-resolution grids are published separately from national or local grids. For instance, Germany releases agricultural data on a 5 km grid, with slightly different confidentiality rules than those applied to the FSS data[8]. In theory, it might be possible to deduce a suppressed cell in the German dataset by subtracting the values of the corresponding 10 km grid cells from the FSS data. However, since the German dataset uses classes and the FSS data are rounded prior to publication, the likelihood of identifying confidential information is extremely low.

6 Conclusions

In this paper, we presented a method for creating a gridded layer of varying resolutions that maximizes the information content at an aggregated level while respecting confidentiality rules and the recommendations for data disclosure from Eurostat. The R package includes several features that have not been a part of previous methods for producing multi-resolution grids, such as contextual suppression, joint gridding of several variables, and the possibility for additional user-defined restrictions.

The next steps are to apply the method to produce a set of key agricultural indicators from the agricultural census and survey data for Europe, which can be used to better understand agricultural systems across Europe and to identify what drives the adoption of different agricultural practices. The release of grids for analyzing change over time will be more challenging as the multi-resolution grids will need

[8] https://agraratlas.statistikportal.de/

to be spatially consistent if meaningful comparisons are to be made. Methods for ensuring both spatial and temporal consistency will be added in the future.

This method is the start of what could be generalized into an on-demand web processing service that would allow users to select the variables of interest and produce multi-resolution grids without requiring large labor resources from Eurostat while respecting all the required confidentiality measures. Such a service could also result in the considerable uptake and use of high-resolution European agricultural survey and census data that up to now has only been possible at a highly aggregated resolution.

Funding No additional funding (all authors paid directly or indirectly by the European Commission).

Data Availability Statement The example data in the manuscript used for all figures (except for Figure 12) is available through the R package MRG, published on the Comprehensive R Archive Network (CRAN).

Declarations

Materials availability NA.

Code availability The code is available through the R package MRG, published on the Comprehensive R Archive Network (CRAN).

Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use) NA.

Ethics approval and consent to participate NA.

Consent for publication All authors have given their consent.

References

Andridge RR, Little RJ (2010) A review of hot deck imputation for survey non-response. Int Stat Rev 78(1):40–64

Asim K, Schorlemmer D, Hainzl S et al (2023) Multi-resolution grids in earthquake forecasting: the quadtree approach. Bull Seismol Soc Am 113(1):333–347. https://doi.org/10.1785/0120220028

Behnisch M, Meinel G, Tramsen S et al (2013) Using quadtree representations in building stock visualization and analysis. Erdkunde 67(2):151–166. https://doi.org/10.3112/erdkunde.2013.02.04

Breidaks J, Liberts M, Ivanova S (2020) vardpoor: Estimation of indicators on social exclusion and poverty and its linearization, variance estimation. Riga, Latvia, https://csblatvia.github.io/vardpoor/, r package version 0.20.1

Copus A, Hall C, Barnes A et al (2006) Study on Employment in Rural Areas. Available at: https://www.napier.ac.uk/~/media/worktribe/output-246104/serareport1ruraleu272006pdf.pdf. Tech. rep

de Jonge E, de Wolf PP (2022) sdcSpatial: Statistical Disclosure Control for Spatial Data. https://CRAN.R-project.org/package=sdcSpatial, r package version 0.5.2

Duncan G, Lambert D (1989) The risk of disclosure for microdata. J Bus Econ Stat 7(2):207–217

Einarsson R, Pitulia D, Cederberg C (2020) Subnational nutrient budgets to monitor environmental risks in EU agriculture: calculating phosphorus budgets for 243 EU28 regions using public data. Nutr Cycl Agroecosyst 117(2):199–213. https://doi.org/10.1007/s10705-020-10064-y

European Commission (2021) Statistical disclosure control. In: European Business Statistics Manual: 2021 Edition. Publications Office, LU, https://doi.org/10.2785/50198

Eurostat (2019) Quality assurance framework of the european statistical system. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf

Eurostat (2020) Integrated Farm statistics manual - 2020 edition. Tech. rep., Eurostat, https://doi.org/10.2785/03054, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-20-009

Eurostat (2023) Integrated Farm statistics manual - 2023 edition. Tech. rep., Eurostat, https://wikis.ec.europa.eu/display/IFS/Integrated+Farm+Statistics+Manual+%7C+2023+edition

FAO (2017a) World programme for the census of agriculture 2020. Volume 1 Programme, concepts and definitions. No. 15 in FAO Statistical Development Series, Food and Agriculture Organization of the United Nations, Rome, oCLC: 1091006801

FAO (2017b) World programme for the census of agriculture 2020. Volume 2 Operational guidelines. No. 16 in FAO Statistical Development Series, Food and Agriculture Organization of the United Nations, Rome, oCLC: 1091006801

Fienberg SE, Jin J (2009) Statistical Disclosure Limitation For Data Access. In: Liu L, Özsu MT (eds) Encyclopedia of Database Systems. Springer US, Boston, MA, p 2783–2789, https://doi.org/10.1007/978-0-387-39940-9_1046

Ford BL (1983) An overview of hot-deck procedures. Incomplete Data Sample Surv 2(Part IV):185–207

Higgins SI, Scheiter S (2012) Atmospheric CO2 forces abrupt vegetation shifts locally, but not globally. Nature 488(7410):209–212. https://doi.org/10.1038/nature11238

Hijmans RJ (2023) terra: Spatial data analysis. https://CRAN.R-project.org/package=terra, r package version 1.7-46

Hundepool A, Domingo-Ferrer J, Franconi L et al (2010) Handbook on statistical disclosure control. ESSnet on Statistical Disclosure Control

Joenssen DW, Bankhofer U (2012) Hot deck methods for imputing missing data. In: Perner P (ed) Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, pp 63–75, https://doi.org/10.1007/978-3-642-31537-4_6

Khan J, Powell T, Harwood A (2013) Land use in the UK. Available at: https://seea.un.org/content/land-use-uk

Lagonigro R, Oller R, Martori JC (2020) AQuadtree: an R package for quadtree anonymization of point data. R J 12(2):209–225

Macdonald S (2004) Counting cows and cabbages ??? Web-based extraction, delivery and discovery of geoReferenced data. IASSIST Quarterly Spring:5–14. https://iassistquarterly.com/public/pdfs/iqvol281macdonald.pdf

Neuenfeldt S, Gocht A, Heckelei T et al (2019) Explaining farm structural change in the European agriculture: a novel analytical framework. Eur Rev Agric Econ 46(5):713–768. https://doi.org/10.1093/erae/jby037 (https://academic.oup.com/erae/article/46/5/713/5183522)

Pebesma E (2018) Simple features for R: Standardized support for spatial vector data. R J 10(1):439. https://doi.org/10.32614/RJ-2018-009

Pebesma E, Bivand R (2023) Spatial data science: with applications in R, 1st edn. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780429459016

Quatember A, Hausner MC (2013) A family of methods for statistical disclosure control. J Appl Stat 40(2):337–346

R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.r-project.org

Ribi Forclaz A (2016) Agriculture, American expertise, and the quest for global data: Leon Estabrook and the First World Agricultural Census of 1930. J Glob Hist 11(1):44–65. https://doi.org/10.1017/S1740022815000340

Shlomo N (2018) Statistical disclosure limitation: new directions and challenges. J Priv Confid 8(11):25–69. https://doi.org/10.29012/jpc.684

Statistics Canada (2023a) Census consolidated subdivision (CCS). Dictionary, Census of Population, 2021. Available at: https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-eng.cfm?ID=geo007

Statistics Canada (2023b) Guide to the Census of Agriculture, 2021. Available at: https://www150.statcan.gc.ca/n1/pub/32-26-0002/322600022021001-eng.htm

Templ M (2017) Statistical disclosure control for microdata. Springer, Cham

The European Commission (2008) Regulation (EC) 1166/2008 of the european parliament and of the council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing council regulation (EEC) No 571/88. Off J L321:14–34

The European Commission (2009) Regulation (ec) no 223/2009 of the european parliament and of the council of 11 March 2009 on european statistics and repealing regulation (ec, euratom) no 1101/2008 of the european parliament and of the council on the transmission of data subject to statistical confidentiality to the statistical office of the european communities, council regulation (ec) no 322/97 on community statistics, and council decision 89/382/eec, euratom establishing a committee on the statistical programmes of the european communities. Off J L87:164–173

The European Commission (2010) Commission regulation (EU) No 1089/2010 of the European Parliament and of the Council of 23 November 2010 implementing directive 2007/2/ec of the european parliament and of the council as regards interoperability of spatial data sets and services. Off J L323:11–102

The European Commission (2018) Regulation (EU) 1091/2018 of the European Parliament and of the Council of 18 July 2018 on integrated farm statistics and repealing Regulations (EC) No 1166/2008 and (EU) No 1337/2011. Off J L200:1–29

The European Commission (2020) Commission implementing regulation (EU) 2018/1874 of 29 November 2018 on the data to be provided for 2020 under Regulation (EU) 2018/1091 of the European Parliament and of the Council on integrated farm statistics and repealing Regulations (EC) No 1166/20. Official JOurnal L306 Novermber 2018:14–19

Trewin D, et al (2007) Principles and guidelines of good practice for managing statistical confidentiality and microdata access. UNECE United Nations Economic commission for Europe: http://www unece org/stats/documents/tfcm/1epdf

USDA NASS (2024) Census of Agriculture. Available at: https://www.nass.usda.gov/Publications/AgCensus/2022/index.php