EUROPEAN COMMISSION
EUROSTAT

Directorate E – Sectoral and regional statistics
E.1 – Agriculture and fisheries

10 January 2025

Geospatial data from agricultural census: Methodological note

1. Introduction

Combining geographical information and official statistics on the farm structure helps to unveil developments of various aspects related to agriculture, allows evaluating the impact of policy measures on farming at more local levels, and improves the monitoring of the agricultural sector. Yet, public access to these new data sources at European Union (EU) level has been hampered owing to the lack of method ensuring data confidentiality and assuring the quality of estimated indicators.

Data on farm structure are collected from respondents in a census or survey which can contain information related to commercial operations or sensitive personal data. To release census and survey data, the application of statistical disclosure control methods is required to reduce the risk of disclosing private information at an acceptable level[1].

The EU Regulation 2018/1091 on integrated farm statistics put forward the commitment to make highly resolved spatial information from agricultural censuses and surveys available to users by establishing a harmonised approach to safeguard confidential information and guarantee the quality aspects for the data dissemination.

A joint project between Eurostat and Joint Research Centre was launched in 2022 to develop a flexible approach[2] by combining the Quadtree-based method[3] with suppression method to

[1] Templ, M., 2017. Statistical disclosure control for microdata. Cham: Springer
[2] Skoien et al. 2024, Flexible Approach for Statistical Disclosure Control in Geospatial Data.
[3] Behnisch, M., Meinel, G., Tramsen, S., Diesselmann, M., 2013. Using quadtree representations in building stock visualization and analysis. Erdkunde 67, 151–166.

maximise the utility of the data and at the same time to reduce of the risk of disclosing private information from individual units.

The methodological note is based on Skoien et al. (2024) and more detailed information can be found in their article.

To protect data confidentiality, a high-resolution grid is produced from geo-reference data with a minimum size of 1 km nested in grids with increasingly larger resolution based on statistical disclosure control methods (i.e threshold and concentration rule). While the method overcomes certain weaknesses of Quadtree-based method by accounting for irregularly distributed and relatively isolated marginal units, it also allows creating joint aggregation of several variables.

The method is illustrated by relying on synthetic data of the Danish 2020 agricultural census for a set of key agricultural indicators, such as the number of agricultural holdings, the utilized agricultural area, and the number of organic farms. The need to assess the reliability of indicators is demonstrated when using a sub-sample of synthetic data followed by an example that presents the same approach for generating a ratio (i.e., the share of organic farming).

Although, the methodology is specifically demonstrated using agricultural census data, such an approach could also be adapted for releasing other microdata based on census or survey whereas data confidentiality, privacy or reliability is required to release the data at higher spatial level of resolutions.

2. Data

2.1. Farm structure survey

Farm structure surveys in the European Union have been carried out since 1966, and the results 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. The survey is separated into core and module variables[4], 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 utilised agricultural area and 98% of the livestock units of each MS. The modules contain information on specific topics, such

[4] The complete list and description of variables surveyed during the European agricultural census 2020 can be found in the EU Implementing Regulation 2018/1874 of 29 November 2018 on the data to be provided for 2020 under EU Regulation 2018/1091 of the European Parliament and of the Council on integrated farm statistics and repealing EU Regulation 1166/2008 and EU Regulation 1337/2011.

as the labour 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 EU Regulation 2018/1091.

2.1.1. The raw survey data

The actual data collection is done by national data providers (i.e., national statistical institutes, ministries of agriculture or other governmental bodies), who prepare the questionnaire, conduct the interviews, and complete the survey with additional information from administrative registers (e.g., wine, bovines, integrated information and 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 micro data. For the first time during the 2020 agricultural census, Eurostat introduced an automated error detection procedure, leading to higher quality statistics. While an agricultural census is carried out every 10 years, sample surveys are administered during interim years. A substantial volume of information was collected during the 2020 survey campaign, which was comprised of more than 300 variables from around 9.03 million agricultural holdings. In sample survey years such as 2016, 1.69 million agricultural holdings were surveyed, which at that time represented approximately 10.55 million holdings. 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. 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

For practical purposes, we have derived a synthetic data set from the original Danish 2020 agricultural census micro data to avoid any malicious disclosure of sensitive information. This data set is used in the examples below. Whereas the values are different from the true values, their distribution mimics the distribution of the true data.

2.2. Statistical disclosure control and quality rating

Surveys collect data on an individual basis, but these cannot be disseminated without protecting the confidentiality of individuals and organisation. Official aggregated statistics can only be disseminated if the values do not reveal sensitive information according to international and European law[5].

[5] Eurostat, 2019. Quality assurance framework of the European statistical system.

At a higher spatial resolution, there is a legally binding obligation to employ appropriate aggregation and disclosure control to render spatial data sets accessible to the public (The European Commission, 2018). Furthermore, the EU Implementing Regulation 2018/1874 defines a set of rules for disclosing information from European surveys on the structure of agricultural holdings collected at farm location 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 condition.

For the dissemination of aggregated tabular statistics, key elements include mandatory compliance with the threshold[6] and dominance rule[7], and the statistical output must satisfy certain quality criteria[8]. These rules also apply to the dissemination of spatially explicit data on a grid.

Lastly, a quality control of the gridded value is necessary. This is usually not an issue when (almost) the entire population has been sampled, as in census years, but in sample years, the use of larger extrapolation weights will introduce prediction errors to the gridded values. 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 recommends that the relative standard error (coefficient of variation) of the estimate should be less than 0.35, as otherwise, the information cannot be disclosed.

In the context of data analysis, it's crucial to verify the accuracy of gridded values, especially when working with sample populations rather than complete datasets. This is particularly relevant in years when only a subset of the population is surveyed, as this will introduce errors in the gridded data. By accounting for factors such as sample size, total population, and any stratification used, it's possible to estimate the magnitude of these prediction errors.

When interpreting gridded data, a commonly accepted benchmark is the relative standard error, also known as the coefficient of variation. According to guidelines outlined in the Integrated Farm Statistics Manual, estimates should have a relative standard error of 0.35 or lower to

---

6 Suppression of cells representing less than four agricultural holdings.

7 Suppression of cells when one or two contributors are dominant in the cell covering a certain percentage of the total value of the grid cell.

8 Data accuracy is evaluated based on sampling errors that can be estimated from the sample itself using the standard errors of the estimated values. If the coefficient of variation of the estimated values is larger than 35%, the cell is usually suppressed.

ensure the data is reliable and accurate enough for disclosure. If this threshold is not met, the data may not be suitable for public release due to concerns over its precision.

3. Method

This section explains the harmonised approach to release grid data from agricultural census by complying with the statistical quality requirements.

3.1. Multi-resolution grid

Our flexible approach is based on Quadtree method[9] which forms grids of different resolutions in hierarchical structure.

To achieve this, the resolution of the grid needs to adapt to the local density of observations, ensuring that confidential data is not compromised. This involves creating a series of base grids with varying resolutions to accommodate different levels of detail. The European Union's Implementing Regulation 2018/874 provides specific guidelines for the initial three resolutions: 1 km, 5 km, and 10 km. However, there are no restrictions on coarser resolutions, and a hierarchical structure is recommended to facilitate data analysis and visualization. In a hierarchical grid system, coarser resolution grids must be integer multiples of higher resolution grids. For instance, a 10 km grid could be used as a base, with coarser resolution grids of 20 km, 40 km, 80 km, and 160 km. Another example might be a 10 km base grid with coarser resolution grids of 50 km, 100 km, and 200 km. But, it's not possible to create a grid hierarchy where a coarser resolution grid is not a whole number multiple of a higher resolution grid.

[9] Behnisch, M., Meinel, G., Tramsen, S., Diesselmann, M., 2013. Using quadtree representations in building stock visualization and analysis. Erdkunde 67, 151-166.

Figure 1: Fictitious example of gridding based on multi-resolution approach

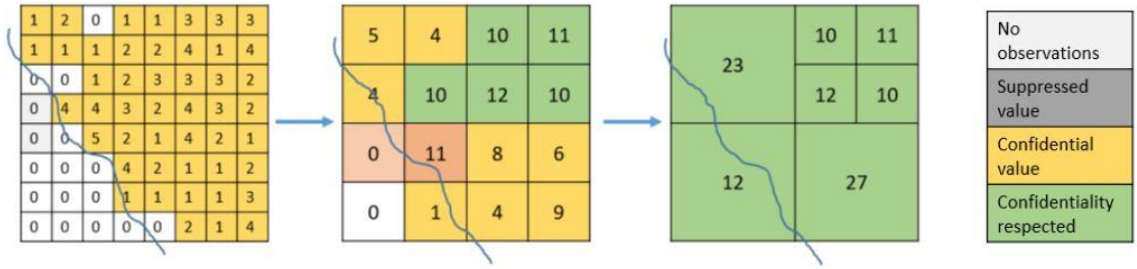
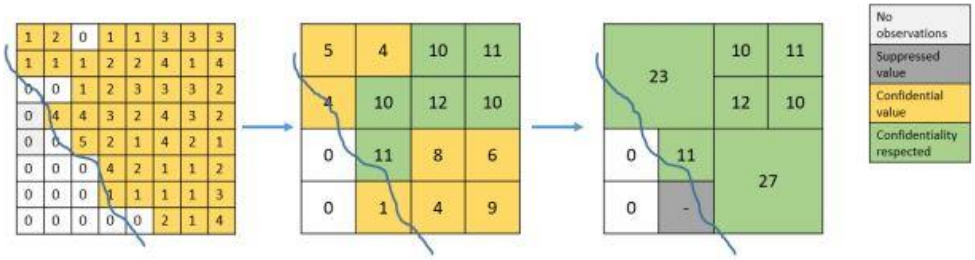


Figure 1 shows an example of the creation of multi-resolution grids. In the left panel, the numbers represent the number of holdings per grid cell. If 10 holdings are necessary for disclosing the information from a grid cell, none of the grid cells will pass the confidentiality rules. In the second grid, 2× 2 blocks of the original cells have been aggregated to larger cells. Here several of the cells (green) respect the confidentiality rules, but others (yellow) do not, so this grid cannot be disclosed either. In the last grid in Figure 1, all grid cells respect the confidentiality rules. However, we can see that the notice that the data in the lower left grid cell quarter could have been disseminated at the intermediate resolution if it weren't for the grid cell with only one farm.

The issue above is particularly causing issues for islands, coastlines, and borders, where empty areas make it difficult to include enough holdings, even for relatively large grid cells. Therefore, one can consider the option of stopping at a maximum resolution to suppress the values from grid cells that still do not respect the confidentiality rules, or to also suppress some of the smaller grid cells that would not contribute much to the total value of an aggregated grid cell. This means that the sum of the values from the grid cells will be lower than the total number of holdings, but considerably fewer than if the suppression was done for a particular grid size. An example of this is shown in Figure 1 with the same fictitious data and the same aggregation as in Figure 1. The grid cell with value 1 in the lower left block of the right panel is suppressed as it represents less than 10% of the value of the possible lower resolution grid cell.

Figure 2: Combining the multi-resolution method with suppression approach.



We have, therefore introduced the possibility for the user to set a minimum share of a grid cell value relative to the possible lower resolution grid cell before it is necessary to aggregate.

Figure 3 displays the iterative process of producing a nested structure of multi-hierarchical grids satisfying a set of confidentiality rules and quality requirements. The confidentiality rules are evaluated in the following order, where the threshold rule must be passed, whereas it is sufficient to pass one of the dominance rules:

1. **Threshold rule**: If the aggregated extrapolated number of agricultural holdings in grid cell l (w_l) for resolution k_0 in iteration i_1 is less than ten ($W_l < 10$ with $W_l = \sum w_j$), then the grid cell size must be enlarged to k_1 and the confidentiality rules for the new grid cell will be assessed in iteration i_2.

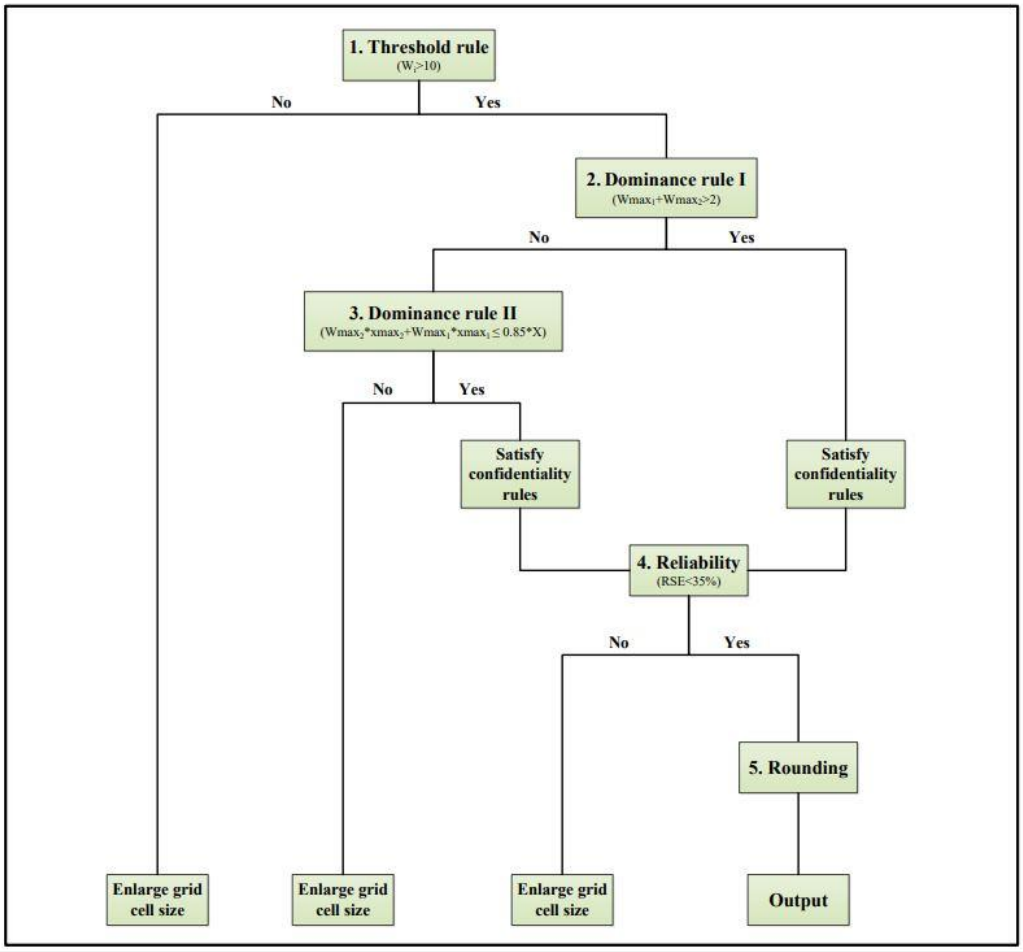
2. **Dominance rule I**: If, after ordering the variable of interest in descending order, the sum of the weights w_{jmax1} of the highest value x_{max1} ($W_{max1} = w_{jmax1} \times x_{max1}$) and of the second highest value x_{max2} ($W_{max2} = wjmax2 \times xmax2$) is greater than two ($W_{max1} + W_{max2} > 2$), then the dominance rules are satisfied for the grid cell at k_0 and the reliability of the results needs to be assessed. (Note that the weights are rounded before this step, so larger than 2 means at least 3). Otherwise, **Dominance rule II** needs to be satisfied.

3. **Dominance rule II**: If the two potential dominant contributors are less than or equal to 85% of the extrapolated aggregated value of the grid cell ($W_{max2} \times x_{max2} + W_{max1} \times x_{max1} \leq 0.85 \times X$), then the confidentiality rules are satisfied; otherwise, the grid cell size must be enlarged to k1 and the confidentiality rules for the new grid cell will be assessed in iteration i_2.

4. **Reliability of the results:** If the coefficient of variation (Relative Standard Error (RSE)) for the grid cell at k_0 is less than 35%, then the indicator is reliable (to be disseminated with a warning if above 25%); otherwise, the grid cell size must be enlarged to k_1 and the confidentiality rules for the new grid cell will be assessed in iteration i_2.

5. **Rounding:** After the last iteration, and as a measure to add further perturbation to the disclosed information, all non-confidential extrapolated number of holdings and extrapolated aggregated values of variables are rounded to the nearest multiple of ten.

Figure 3: Flowchart of the iterative process to produce multi-resolution grid based on confidentiality rules and quality criteria.



3.2. Joint aggregation

Some indicators might be a function of two or more variables. One such scenario arises when an indicator is calculated as a ratio of two variables, such as the ratio of Var1 to Var2. In this case, both variables must be treated as sensitive data and their grid cells must be protected accordingly.

However, when creating separate grids for each variable, it's unlikely that they will align perfectly, leading to discrepancies and potential data quality issues. To address this challenge, it's often more effective to grid the variables jointly, rather than separately. This approach involves merging the data and applying a consistent resolution to both variables, replacing high-resolution grid cells with low-resolution ones, if necessary, to ensure that the confidentiality rules are respected while maintaining the integrity of the data.

3.3. Post-processing of the data

Following the creation of multi-resolution grids, additional post-processing steps can be taken to further ensure compliance with confidentiality regulations. Although the preceding procedures aim to safeguard sensitive data, there are still potential vulnerabilities that can arise when working with multi-resolution grids.

Specifically, the lowest resolution grid may not provide adequate protection for individual data points, and some grid cells may remain exposed due to their proximity to high-resolution cells. To address these concerns, supplementary post-processing methods can be employed to reinforce confidentiality measures and prevent potential data breaches.

The post-processing step will therefore suppress the values in cells that do not pass the confidentiality rules.

4. Results

This section presents the flexible approach by using key agricultural variables of a synthetic dataset from the Danish 2020 agricultural census. To illustrate an agricultural survey, a subset is used to demonstrate the need to apply additionally reliability assessment.

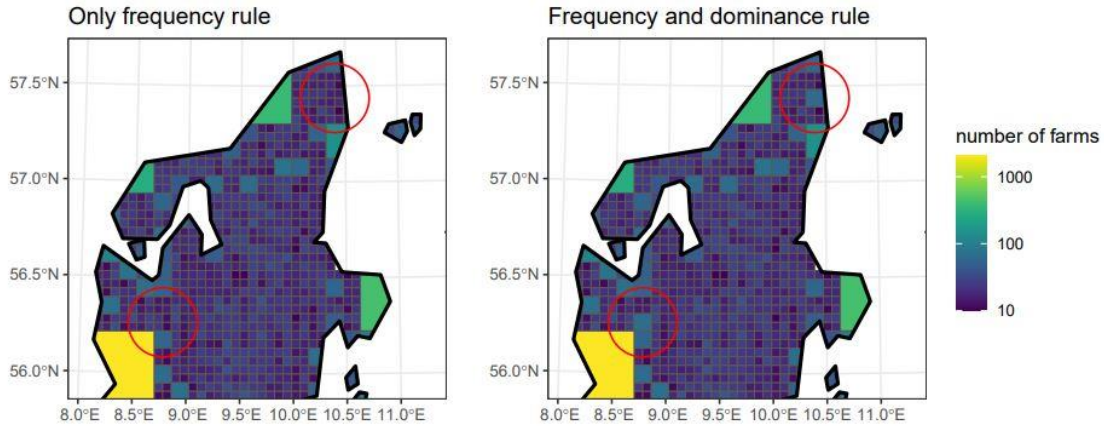
4.1. Multi-resolution grid of key agricultural variables

Displayed in Figure 4 is the comparison of multi-resolution grids for different confidentiality rules applied for the number of agricultural holdings of the synthetic data. In the left panel, only the frequency rule was applied, ensuring that all grid cells have at least 10 holdings.

Most of the grid cells in this panel (1174) have a resolution of 5 km but there are also 143 with a resolution of 5 km, 29 with 20 km, 8 with 40 km, and 1 with 80 km. We can notice that most of the larger grid cells are on the coastline.

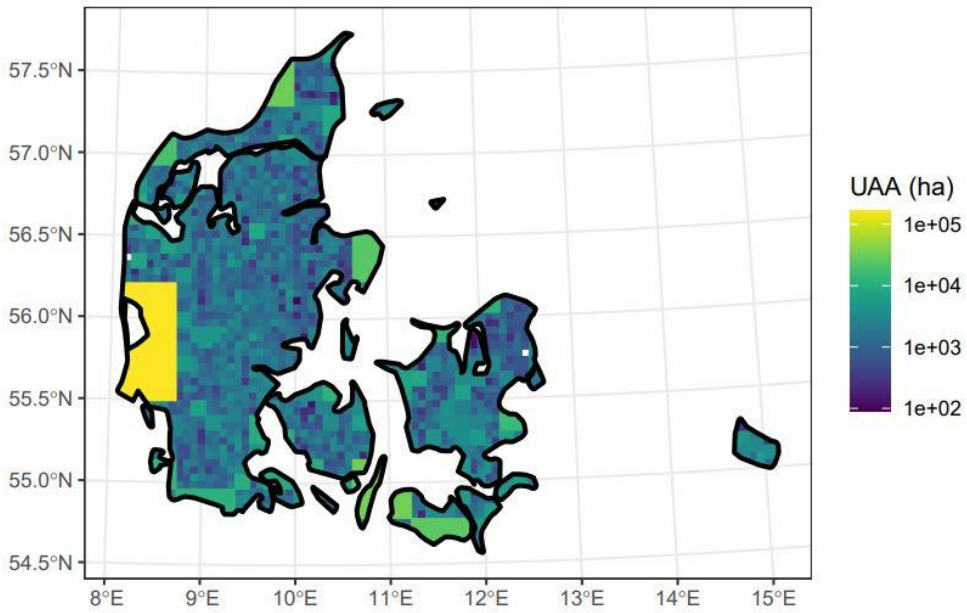
Many of the smaller grid cells include10-50 holdings. However, there are 19 grid cells with more than 100 holdings or more. One of them has 2078 holdings.

Figure 4: Comparison of applied confidentiality rules for multi-resolution grids of the number of holdings.



The right panel of Figure 4 shows the map after adding the dominance rule. There is only a small difference between the two: 12 fewer 5 km cells, 1 fewer 10 km cell and 2 more 20 km cells. The aggregation is caused by some large farm holdings/producers that were too dominant within the grid cell. The circles show where smaller grid cells have been merged because of the dominance rule.

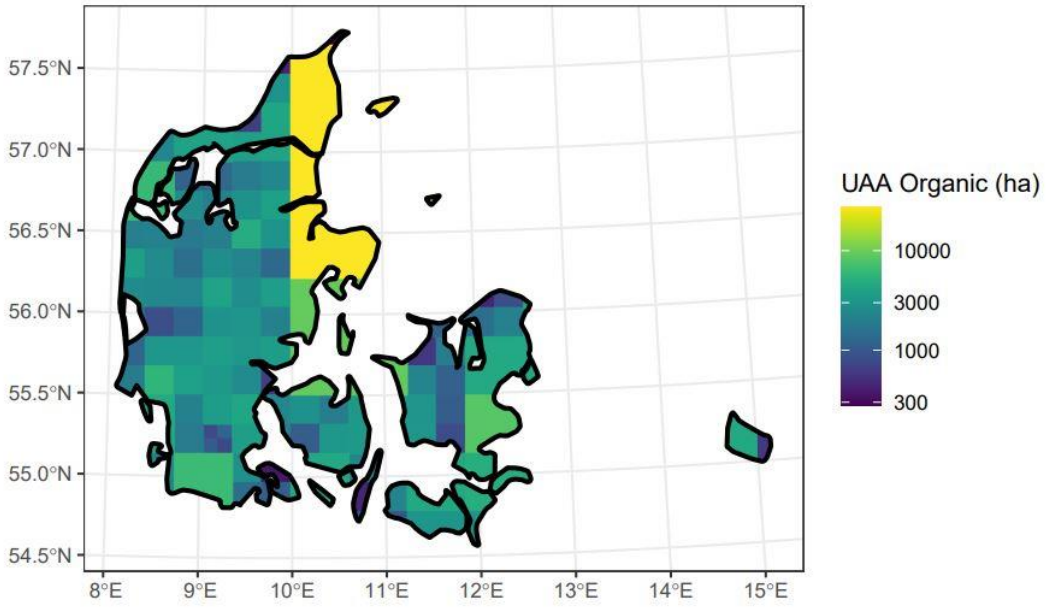
Figure 5: Multi-resolution grid of UAA.



The map of the organic UAA displayed in Figure 6is different with considerably larger grid cells. This is because there are considerably fewer holdings with organic farming in this data set. Only

one grid cell is non-confidential at 5 km, whereas most 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. The map of organic farming has a total of 180 grid.

Figure 6: Multi-resolution grid of organic UAA



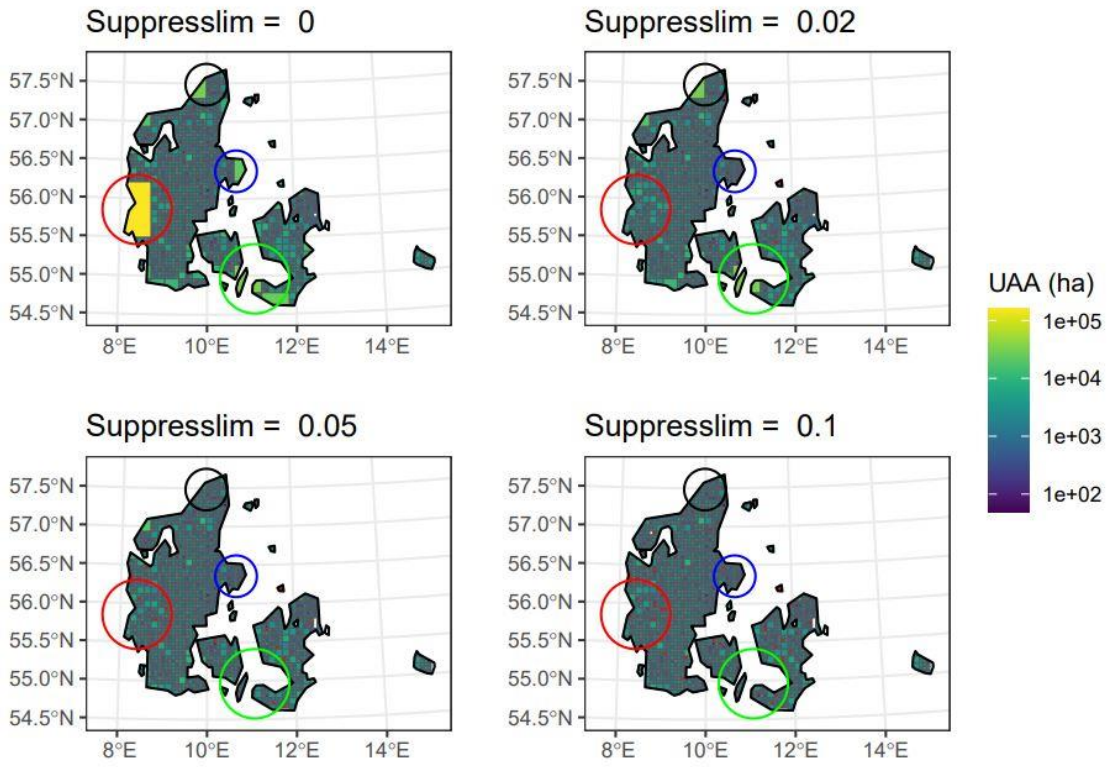
4.2. Suppression of insignificant grid cells

The large grid cells on the coast will mask the details just inside the coastline, and will in most cases be unwanted. Instead, we can suppress some of the smaller grid cells with a few holdings instead of aggregating them with grid cells that already have a high number of holdings. The effect of different suppression thresholds on UAA are demonstrated in Figure 7.

Varying the suppression rule has a direct effect on the grid cells that should be merged and generates different outcomes. The ones inside the red and blue circles disappear already when applying a threshold of 0.02 (meaning the grid cell accounts for less or equal 2% of the total aggregated value). The ones in the black and green circles disappear with 0.05, and the grid cells within the green circle are further reduced in size for 0.1. The suppressed grid cells (red squares) are barely visible for the lowest threshold, whereas there are considerably more (and larger) grid cells suppressed for the largest suppression threshold.

It can be concluded that the number of large grid cells decreases with increasing suppression threshold value.

Figure 7: The effect of suppression rule on the outcome



4.3. Demonstrating the need for reliability checks

When creating multi-resolution grids from survey data, it is essential to consider the reliability of the gridded values. Unlike census data, which covers the entire population, survey data is often collected in a stratified approach, where the selection of surveyed holdings is based on various criteria such as geographic location, farm size, economic size, and crop types. In this context, the reliability of the estimates depends on the weighted number of holdings and weighted values.

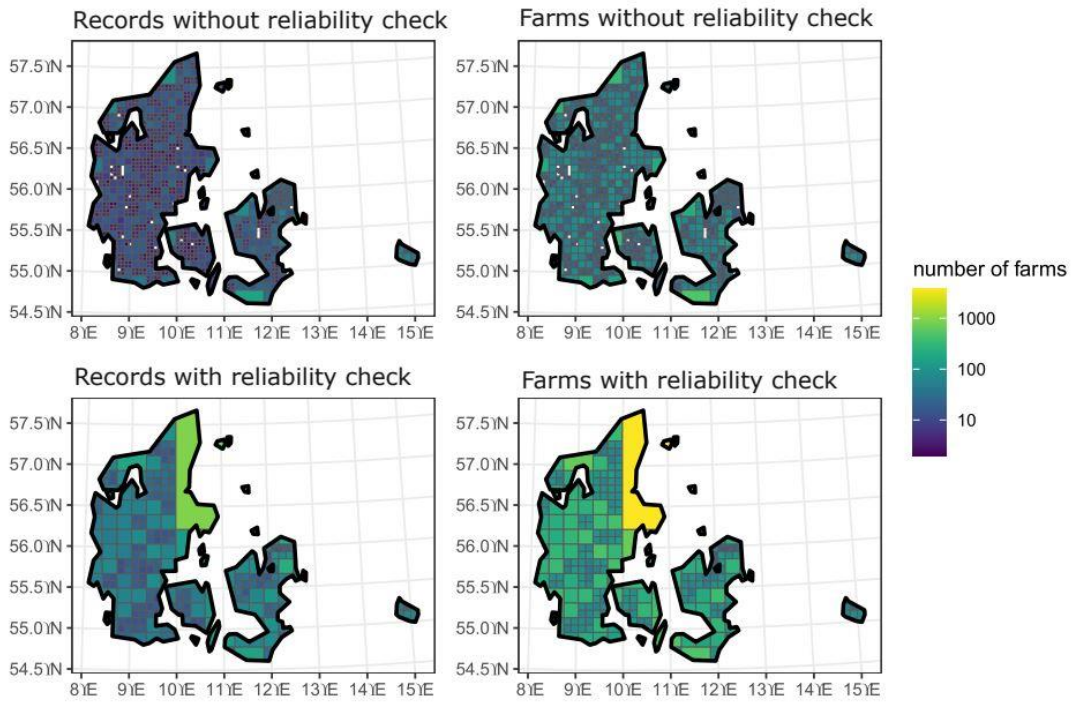
A grid cell value can be considered reliable with fewer than 10 records, but the threshold depends on the population size and variability within the recorded values. However, high-weighted observations can lead to unreliable estimates, as the grid cell values might be based on a single or very few observations.

This is demonstrated in Figure 8, which compares the effects of considering and not considering reliability checks on gridded synthetic agricultural survey data from Denmark. The top panels show the data without reliability checks, while the bottom panels demonstrate the results with reliability checks applied. A notable difference is observed in the number of grid cells, with the

reliability check procedure leading to a smoother and more realistic map. In particular, the tiny grid cells with few holdings but large enough weights to pass confidentiality rules have largely disappeared, reducing the number of grid cells based on more accurate records. Notably, only 6 grid cells have less than 10 records, with the lowest count being 3 records.

Although the reliability check is an important component of the multi-resolution grid production process, it is not applied by default due to its computationally intensive nature.

Figure 8: Gridding the number of farms with and without reliability treatment.

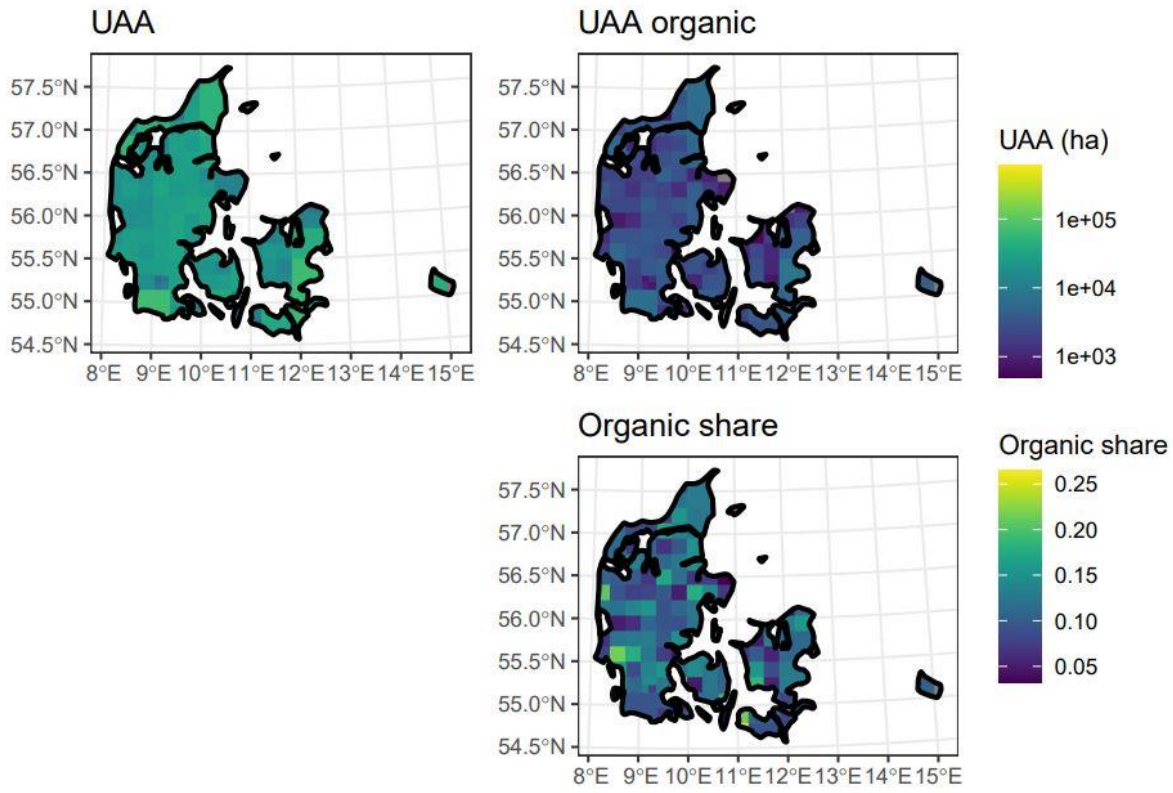


4.4. An example of producing a ratio

It is only possible to make ratios if the maps have the same resolution. Therefore, a meta grid is computed by jointly gridding several variables implying that the confidentiality rules are respected for all variables for each grid cell. This has been done for organic and total UAA in Figure 9, where the upper panels show the gridded values of the two variables, whereas the ratio (organic share) is shown in the lower panel. We can see that the grid cells are the same for all grids. To produce the ratio or the share of organic UAA, the gridding procedure was performed with the suppression threshold value of 0.05, which resulted in the suppression of three grid cells. The figure indicates that the concentration of organic farming is higher in the

south of the country for the synthetic dataset, although this pattern may differ when actual data from the agricultural census is used.

Figure 9: Producing a ratio from gridding UAA and organic UAA.



5. Discussion and conclusion

The new methodology marks a significant shift in the way agricultural census data can be released at high spatial resolutions while maintaining confidentiality. This approach maximizes information content and releases it at the highest possible resolution, adhering to EU laws and Eurostat guidelines.

In contrast, other countries outside the EU are more restrictive in their data dissemination. For instance, the United States Department of Agriculture releases data at the county level, similar to NUTS2 regions in Europe[10].

[10] USDA NASS, 2024. Census of Agriculture.

In Canada, one-third of data were not disclosed in the 2016 agricultural census, which employed suppression of data. Canada's 2016 agricultural census suppressed one-third of the data, while the 2021 Census employed random tabular adjustment to ensure data protection, albeit with limitations on data release areas and potential comparability issues[11]. In the UK, the Edinburgh Data and Information Access (EDINA) releases data at 2, 5, and 10 km grids, but with less reliable data in areas where disclosure requirements are not met[12].

[11] Statistics Canada, 2021. Guide to the Census of Agriculture.
[12] Khan, J., Powell, T., Harwood, A., 2013. Land use in the UK.

Annex

The geospatial data from 2020 agricultural census are presented in form of a statistical atlas of European farming based on the contextual indicators of the common monitoring evaluation framework of the Common Agricultural Policy for the period 2014 to 2022. Users can download the data files in csv format that contain the latitude and longitude of lower left corner of the grid cell, the grid cell sizes, and the relevant agricultural variables to construct the indicator. Note that the coordination reference system '3035' is applied in all datasets, and that the latitude and longitude coordinates as well as the resolution have to be multiplied with 1000 before recreating the polygons. The associated R package MRG includes a function (MRGfromDF) which simplifies the conversion from csv-files to polygons that can be stored as shapefiles.

Additionally, we also provide the data as GeoPackages (in the .gpkg format) by using '3035' coordination reference system.

The context indicators are organised into three broad categories: structural components, demographics of farmers and agricultural production method. Several indicators are part of each component, and each indicator corresponds to a dataset with specific codes[13] which are explained in Tables 1 to 3. There are some additional issues to be considered, summarized here:

1. All original variables are in capital letters. There are also some additional variables with names in camel case (alternating capital and small letters) which describes derived variables in the GeoPackage files.

2. In some cases, the additional variables have been created by division on the actual area of a grid cell. For coastal areas, the grid cells were first intersected with a European map downloaded from GISCO (with a relatively coarse resolution 1:20 million). This would create smaller polygons on coast lines, which are then used in the division. The exported grid will include the entire grid cell (as some users will maybe like to intersect with a different country polygon data set), but the areal value from the intersection is included in the geopackage.

3. In general, large areal average values should be treated with care. These could have some possible causes:
    a. The intersection above created too small polygons, inflating the value/area.

[13] Further information on the variables collected during the agricultural census can be found in the Integrated farm statistics manual, 2020 edition.

b. Some regions have submitted holding locations at the centre of an administrative region instead of the 1 km grid. This will inflate both the total and the average value for the grid cell which includes the centre.

4. Some of the variables that describe a share of something have been calculated from post-processed estimates (non-suppressed values are rounded). As a result, some estimated shares, particularly from grid cells with few farms, will have some uncertainty added to their values, which could give some surprising effects. As an example, whereas the share of grassland, arable land and permanent crop land should sum to one, this is often not the case when the shares have been calculated from the post-processed estimates. The magnitude of the deviation will depend on the number of farms in the grid cell.

5. The collected microdata is supposedly collected on a one km grid (lower left corner). However, this is not always the case. Some regions (such as Sicily in Italy and large parts of Croatia) appear to have submitted the data with the location of the centre of an administrative region instead. This will give high values for the grid cell with the centre, and zero values around. Unfortunately, there is no good way to correct for this, except for requesting better data in the next survey.

6. Almost all data sets include the following columns, which are then not included below (unless it is the column is the main feature of the indicator):

    a. ID – an identification number of the grid cell
    b. res – the resolution of the grid cell (one sided length)
    c. area – the estimated land area of the grid cell – relevant for coastal grid cells, and grid cells on the border towards countries not included in the data set. This value is just an indication, as lakes are included in the area. The exact value is partly depending on the resolution of the coastline data set (Resolution “01” of the GISCO data set)
    d. UAA – the utilized agricultural area in hectares
    e. NUTS2 – the nuts2 region based on overlaying the grid with the GISCO-map of NUTS2 regions. The match is not exact, as a grid cell can be completely within one NUTS2 region or split between 2 or more regions. This should therefore only be used as an indication.

    The reason why UAA can be found in different indicator sets, is that the grids are different between the different data sets. The grid cell size is a function of both the UAA and the other variables of the data set. As an example, there are more smaller grid cells in the

C18a indicator (UAA) than in the C17 indicator (Agricultural holdings) because the latter also includes the Standard output, where some large values might trigger the dominance rule where the UAA might not be confidential. The area is included for most indicators, although it is only necessary for some of the values but is added in case users want to further explore the data.

Table 1: Overview of contextual indicators and meta data of the datasets for context indicators related to structural elements.

Agricultural holdings

Dataset name	**Variables**	**Labels**	**Unit**
C17	HOLDING	Number of holdings	Number
C17	UAA	Utilised agricultural area	Hectare
C17	SO_EURO	Standard output	Euro
C17	HoldingsPerKm2	Holdings per square km	Number/km2

Agricultural area

Dataset name	**Variables**	**Labels**	**Unit**
C18a	UAA	Utilised agricultural area	Hectare
C18b	ARA	Arable land	Hectare
C18b	AraShare	The arable land share of UAA	Percentage
C18c	PECR	Permanent crops	Hectare
C18c	PecrShare	The permanent crops share of UAA	Percentage
C18d	J0000T	Permanent grassland	Hectare
C18d	GrassShare	The grassland share of UAA	Hectare
C18e	Q0000T	Fallow land	Percentage

Dataset name	Variables	Labels	Unit
C18e	FallowShare	The fallow land share of UAA	Percentage

Livestock units

Dataset name	Variables	Labels	Unit
C21a	LSU	Livestock units	Livestock units
C21a	LSUdensity	Livestock units per hectare UAA	Livestock unit per hectare
C21b	BOVINE	Bovine animals	Heads
C21b	BovineDensity	Bovine per hectare UAA	Heads per hectare
C21c	PIGS	Number of Pigs	Heads
C21c	PigsDensity	Pigs per hectare UAA	Heads per hectare
C21d	SHEEP	Sheep	Heads
C21d	SheepDensity	Sheep per hectare UAA	Heads per hectare
C21e	GOATS	Goats	Heads
C21e	GoatsDensity	Goats per hectare UAA	Heads per hectare
C21f	POULTRY	Poultry	Heads
C21f	PoultryDensity	Poultry per hectare UAA	Heads per hectare

Table 2: Overview of contextual indicators and meta data of the datasets for context indicators related to farmers' demographics.

Age structure of farm manager

Dataset name	Variables	Labels	Unit
C23	HOLDING	Number of holdings	Number
C23	Y_LT40	Number of farms with the age of the farm manager below 40 years old	Number
C23	Y_LT40_Share	Share of farms with the age of the farm manager below 40 years old	Percentage
C23	Y_GE65	Number of farms with the age of the farm manager equal to or higher than 65 years	Number
C23	Y_GE65_Share	Share of farms with the age of the farm manager equal to or higher than 65 years	Percentage
C23	AGE_MANAGER	The total value of the age of farm managers in the grid cell	Number
C23	AGE	The average value of the age of farm managers in the grid cell	Average number

Agricultural training of farm manager

Dataset name	Variables	Labels	Unit
C24	HOLDING	Number of holdings	Number
C24	BASIC	Basic agricultural training	Number

Dataset name	Variables	Labels	Unit
C24	BasicShare	Share with basic agricultural training	Percentage
C24	FULL	Full agricultural training	Number
C24	FullShare	Share with full agricultural training	Percentage
C24	PRACT	Practical agricultural training	Number
C24	PractShare	Share with practical agricultural training	Percentage
C24	Control	Sum of shares – deviation from zero indicates rounding effects	Number

Gender gap

Dataset name	Variables	Labels	Unit
C_sex	FEMALE	Number of holdings with female managers	Number
C_sex	MALE	Number of holdings with male managers	Number
C_sex	FemShare	Share of female managers	Percentage
C_sex	HOLDING	Number of holdings	Percentage

Table 3: Overview of contextual indicators and meta data of the datasets for context indicators related to agricultural production methods.

Organic farming

Dataset name	Variables	Labels	Unit
C19	UAAXK0000_ORG	Utilised agricultural converted and under conversion to organic farming excluding kitchen gardens	Hectare
C19	OrgShare	Share of organic farmland relative to UAA	Percentage

Irrigation methods

Dataset name	Variables	Labels	Unit
C20	UAA_IB	Irrigable utilised agricultural area	Hectare
C20	IrrShare	Share of irrigable land relative to UAA	Percentage