



ROMANIA

Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217)

OUTPUT No. 7b

Report on advisory services provided to Recipient on the

Recommendations, best practices and guidance in developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units GAC2020

December 2021

Revised November 2022



*Project co-financed from the European Social Fund through the Operational Programme
for Administrative Capacity 2014-2020*

Disclaimer

This report is a product of the staff of the World Bank. The findings, interpretation, and conclusions expressed in this paper do not necessarily reflect the views of the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

This report does not necessarily represent the position of the European Union or the Romanian Government.

Copyright Statement

The material in this publication is copyrighted. Copying and/or transmitting portions of this work without permission may be a violation of applicable laws.

For permission to photocopy or reprint any part of this work, please send a request with the complete information to either: (i) the Romanian National Institute of Statistics (16 Libertății Blvd., District 5, Bucharest, Romania); or (ii) the World Bank Group Romania (31, Vasile Lascăr Street, 6th floor, Bucharest, Romania).

This report has been delivered in December 2021 and revised version in November 2022 under the Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217) signed between the Romanian National Institute of Statistics and the International Bank for Reconstruction and Development on September 17, 2019. It corresponds to Output 7b under the above-mentioned agreement.

Acknowledgements

This report was prepared under the coordination of Michael Wild, Senior Statistician, World Bank with the support of the local team of experts. The team would also like to express its gratitude to government officials of the National Institute of Statistics (INS), Florentina Gheorghe (General Director), Silvia Pisica (General Director and Project manager) and their team of specialists for their constructive collaboration.

Contents

Report on advisory services provided to Recipient on the.....	1
List of Figures.....	6
List of boxes.....	6
Abbreviations and Acronyms	7
Introduction.....	8
1. Methodology applied for geo-referencing	9
Map Control (Windows Edge and Google Chrome tested).....	11
Grid Types.....	11
Grid IDs.....	20
Download Formats	22
Download the reference grid	22
2. Data preparation and variables by type of farm.....	23
Organization of the GAC database.....	23
Update the database.....	24
3. Installing and configuring the application.....	26
Source code description.....	26
System requirements (hardware and software dependencies).....	26
Installation kit, installation procedure and configuration parameters.....	26
4. Recommendations and best practices.....	27
5. Annexes	29
Annex 1: Filtering of observations by selection from grid cell summary table.....	29
Annex 2: Data Download.....	34
Annex 3: GAC microdata	37
Annex 4: Variables for geo-references the agricultural holdings	41
Annex 5: R script for data matching and correction of GPS coordinates corresponding to the location of the holding	43
Annex 6: R script for transforming FOXCON files into postgres(-postgis db).....	50
Description	50
Load packages and Data	50
Cleaning & Imputation	51
take out I_KEY missings.....	54
Merging the data.....	54
Generating a spatial object from the data	54
Writing to DB	55
Loading the reference grid.....	56
Create spatial object with terra	56
Projecting the data to Lambert Azimuthal Equal Area (epsg 3035).....	56

Creating one grid for all data	56
Writing to DB	56

List of Figures

Figure 1 - Application start screen.	10
Figure 2 - Empty grid of 1 km.....	11
Figure 4 - Selection of variable for Aggregation type.....	12
Figure 4 - Bar height indicates aggregated value.	13
Figure 6 - Different resolutions for the selected variable (10km, 5km, 1km)	14
Figure 7 - Warning message informing user about exclusion of observation when creating the grid.....	15
Figure 8 - All Points	15
Figure 9 - Coordinates displayed over point selected.....	16
Figure 10 - Side Table with all variables (right).....	16
Figure 11 - The distance data allows for inspection of larger distances.....	17
Figure 12 - The distance data allows for inspection of smaller distances	17
Figure 13 - Dynamic Controls adjust to the minimum and maximum distance.	18
Figure 14 - Sub-setting of units by their maximum distance is possible.....	18
Figure 15 - APIA and Non-APIA units without and with grid.....	19
Figure 16 - ALL DATA and single grid cell.....	19
Figure 17 - Data Cube for a single feature	20
Figure 18 - Grid Cell Selection to identify singletons or other privacy violating cells.	21
Figure 19 - Selected Grid Cell for further inspection and follow up.	21
Figure 20 - Update database commands.....	25
Figure 21 - Table of grid cell counts to the right of the map.....	29
Figure 22 - Sort by highest	29
Figure 23 - Sort by lowest	30
Figure 24 - Selected grid cell.....	30
Figure 25 - Select medium size cell.....	31
Figure 26 - APIA units inside grid cell.....	31
Figure 27 - Non-APIA units inside grid cell	32
Figure 28 - 5 km grid cell size cell selection.....	32
Figure 29 - 10 km grid cell size cell selection.....	33
Figure 30 - Loading the data for a particular county creates the points file and the cell count files for download	34
Figure 31 - .zip file with points data and cell count.	35
Figure 32 - The population raster can be downloaded after it is visible on the map.....	35
Figure 33 - .zip data with points data, cell count and raster data.....	36
Figure 34 - QGIS display of raster data for different counties and different resolutions all in a single map...	36

List of boxes

Box 1 - The mean of GAC grid distribution tool.....	9
---	---

Abbreviations and Acronyms

APIA	Agency for Payments and Interventions in Agriculture
CASS	Computer Assisted Survey System
CRS	Coordinates Reference System
DPS	Data Processing and Storage
DTS	Territorial Statistics Directorates/Offices
EC	European Commission
EU	European Union
GA	Grid Application
GAC2020	General Agricultural Census 2020
GEOLOC	Geographical Location Software Application
GIS	Geographical Information System
GPS	The Global Positioning System
IFS	Integrated Farm Statistics
INS	National Institute of Statistics
INSPIRE	Infrastructure for Spatial Information in Europe
IT	Information Technology
RAS	Reimbursable Advisory Services
SDC	Statistical Disclosure Control

Introduction

The purpose of this report is to present **recommendations, best practices and guidance for developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units of the General Agricultural Census 2020 (GAC2020).**

This report is part of the deliverables under the Reimbursable Advisory Services (RAS) Agreement on Romania Capacity Building for Statistics (project No. P167217) and corresponds to Output 7b: *Report on advisory services provided to the Recipient on the Recommendations, best practices and guidance in developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units GAC2020.* The project is implemented by the National Institute of Statistics with support from the World Bank.

This report is structured in four sections and annexes.

The first section provides a description of the methodology applied for georeferencing farms enumerated during the data collection process of GAC2020, to display APIA and Non-APIA collected census data separately or together, to create grid-like statistical units with different cell sizes and answer to the requirements of the INSPIRE Directive as referred in the Regulation (EU) 2018/1091 regarding Integrated Farm Statistics and repealing Regulation (EC) no.1166/2008 and (EU) 1337/2011 and to analyze deviations between the original collected coordinates and the imputed coordinates.

A description of the organization of the GAC database, respectively the data sources used, the structure of the microdata and the preparation (cleaning data) process needed for the geo-referencing process is presented in the second section.

The third section presents details regarding the source code, the system requirements – hardware and software dependencies, the installation procedure and configuration parameters for proper installing and configuring the application.

A list with specific recommendations, best practices for the staff of INS to practice and optimally use the application is contained in the fourth section.

The annexes of the report provide information for the users of the application on the filtering of observations by selection from the grid cell summary table and the data download process, the GAC microdata and the list of GAC database variables used for georeferencing, the R script for data matching and corrections of GPS coordinates, and the R script for transforming FOXCON files into Postgres (postgis db) database.

1. Methodology applied for geo-referencing

The geo-referencing of the agricultural holding in grid type statistical units is the process of association between the location (GPS) of the agricultural holding and the corresponding grid code, according to the specific regulations^{1,2} relevant for integrated farm statistics, including GAC2020.

The methodological and IT solution applied for the geo-referencing of farms enumerated during the data collection process of the GAC, approaches farms differently: the farms that benefit from subsidies and are included in the APIA list, which are identified through GIS coordinates from the APIA source; and the small farms that are not included in the APIA list for which GIS coordinates have been collected according to the criteria of location.

Box 1 – The purposes of GAC grid distribution tool

The main task of the grid application is to produce the grid and the grid-codes for all agricultural holdings.

All visualizations are provided with the intend to support the INS staff in observing and examining data at level of county and grid cell.

The R source code including in-line documentation is provided to INS. This provides the opportunity for INS R experts to implement any further modifications or functionalities themselves as part of the institutional capacity building process.

The GA (Grid Application) allows to visualize large census data sets as well as corresponding spatial analytics either as points or as grid data sets. It is intended to support the work of statisticians with basic GIS knowledge. The GA for Romania focuses on following relevant aspects:

1. Display APIA and Non-APIA collected census data separately as well as together.
2. Create grid like statistical units with cell sizes of 1 km, 5 km, and 10 km of cell size.
3. Assign CELL CODE as required by INSPIRE directive to each of the census units, and for each of the different grid cell sizes.
4. Generate grid like units for better visualization and inspection which can also be used for exporting the grid data and is applicable to all provided numeric variables³.
5. Allow to visualize deviations between the original collected coordinates and the imputed coordinates.

¹ D2.8.I.2 Data Specification on Geographical Grid Systems – Technical Guidelines; INSPIRE, 2014; Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

² EU Regulation 2018/1091 regarding Integrated Farm Statistics and repealing Regulation (EC) no.1166/2008 and (EU) 1337/2011; Eurostat Integrated Farm Statistics Handbook

³ The application allows to display all correctly provide variables as available through the upload/database connection. For viewing purposes, it also excludes observations which are below (above) the 5% or 95% range.

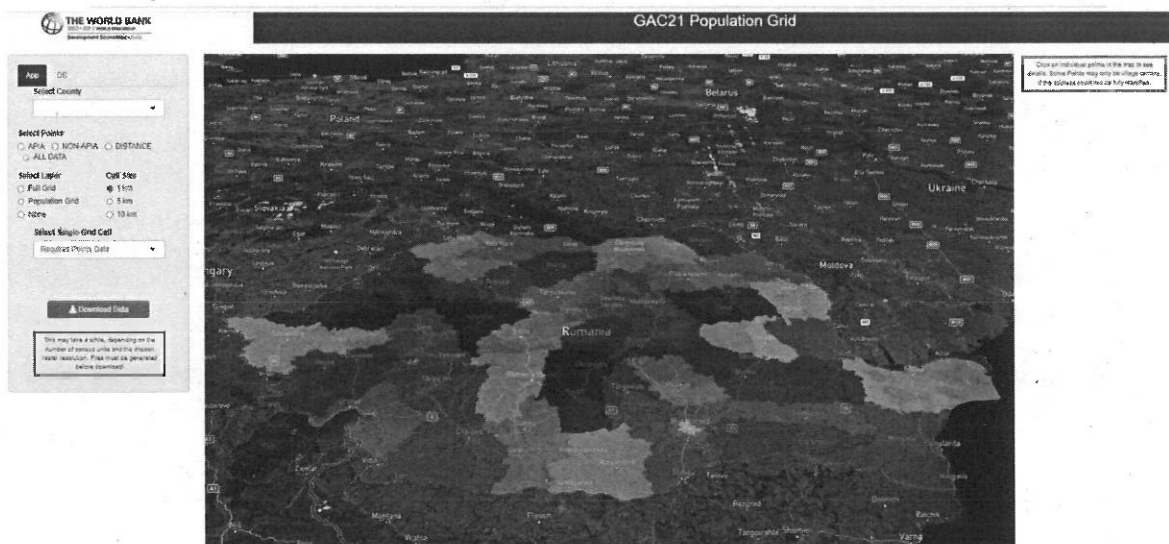
The application run on the INS own IT system and with connection to and use of their own PostGIS-Postgres⁴ database. The WB provided assistance for installation on the INS IT integrated system and offered the guidance to INS staff for proper use, installation and operation. The CRS (Coordinates Reference System) is the one recommended by the European INSPIRE⁵ directive which uses an equal area projection ETRS89 Lambert Azimuthal Equal Area, EPSG:3035. Based on the chosen CRS, the corresponding grid codes are created and assigned to each unit.

The application processes a lot of data and is dependent on the availability of resources that were designed and parameterized for the GAC purposes through Output 4.1b., the number and type of computational processes run by the INS IT system ("Analysis server") when the application is in use, and the number of concurrent users (recommended up to 3 who are running simultaneously the GAC data computation on grids) and by that the application may, sometimes, show screen freezes and other unexpected behavior. In such cases it is recommended to refresh the application and start over again.

Besides allowing the INS to fulfill their obligation towards the Eurostat according to EU Regulation 2018/1091 by creating grid like statistical units and assigning the corresponding cell codes to the data set, this application also allows to create a raster of 1000 (5000, 10,000) meter⁶ cell size which can be used for predictions of coverage, improvements of agricultural surveys during intercensal periods or any other analytical purpose as determined by future activities.

After opening the application in preferred browser⁷, the **first selection** to be made is always the COUNTY of interest. The reason for this is twofold, first, it allows the user a better visualization and, second, it reduces the workload and subsequently the processing time. Following the selection of the county are then the different operation types.

Figure 1 - Application start screen.



⁴ <https://postgis.net/>

⁵ Infrastructure for Spatial information in Europe; <https://inspire.ec.europa.eu/>

⁶ 1000 m is a fairly common cell size used for population projection/visualizations, however since the application is in R and the source code is publicly available, modifications for alternate cell sizes are possible.

⁷ MS Edge in its latest version has given best performance results, Google Chrome works well too, but may require more working memory

Map Control (Windows Edge and Google Chrome tested)

Map controls are based on the Web GL engine (which may not work on older browser versions or on table). The main controls are:

- i. Right Click/Move Mouse → move map
- ii. Right Click/Ctrl/Move Mouse → change viewing angle
- iii. Mouse Wheel +/- → zoom in/out

These should work in all contemporary standard web-browsers, but it has only been tested with Windows Edge⁸ and Google Chrome⁹ so far.

Grid Types

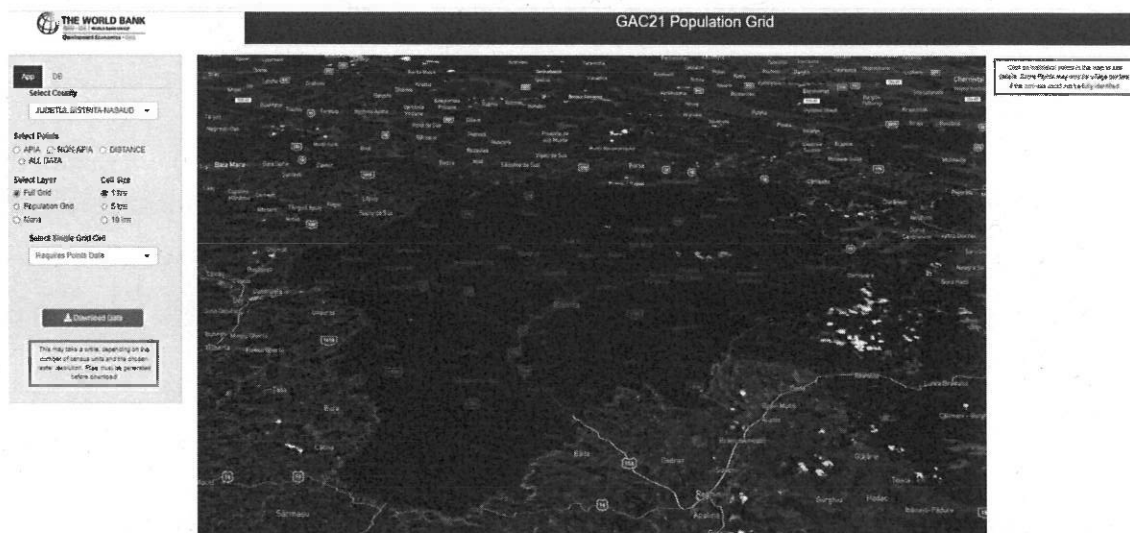
The application allows to (i) create an empty grid to overlay over the points data as well as (ii) aggregate the points data to different statistics (i.e., Total, Mean, Standard Deviation) for a selected variable at the cell level. These gridded population values can also be exported and used as inputs for other purposes, as well as for publicly accessible visualizations¹⁰.

1. FULL GRID

After selecting the county in SELECT COUNTY, select option Full Grid under SELECT LAYER. The default grid size is 1 km, if user wants to change this, should change it before selecting the SELECT LAYER option.

The FULL GRID allows user to visually inspect grid cells which are in violation of the minimum number of observations required for a public use data set, as well as to inspect cluster of observations, check distances by using the grid as a reference etc. When the population grid is created, the same cell size will be applied as in the empty grid produced in this section.

Figure 2 - Empty grid of 1 km



⁸ Version 94.0.992.50 (Official build) (64-bit)

⁹ Version 95.0.4638.54 (Official Build) (64-bit)

¹⁰ A very important aspect in the creation of public use/view data sets is the one of Statistical Disclosure Control, which should ideally be addressed with both sdcMicro, and sdcSpatial, both available in R. It is INS responsibility, that any identifying information is removed.

2. POPULATION GRID

After selecting the county select option Population Grid under SELECT LAYER. When selecting this option, a new drop-down menu SELECT VARIABLE FOR AGGREGATION will show up. This drop-down menu shows all the variable names in the dataset identified by the system as **numeric** except the main coordinates¹¹.

Selecting a variable creates a grid with CELL SIZE, and either calculates the: (i) total, (ii) the mean, (iii) or the standard deviation, over all the observed and valid values of the selected variable from all the observations within the corresponding grid cell. The calculation runs in depending on cell size¹².

Figure 3 - Selection of variable for Aggregation type

The screenshot shows a web application interface with a dark header containing 'App' and 'DB'. Below the header, there is a 'Select County' section with a dropdown menu currently showing 'JUDETUL BISTRITA-NASAUD'. Underneath, the 'Select Points' section has four radio button options: 'APIA' (selected), 'NON-APIA', 'DISTANCE', and 'ALL DATA'. The 'Select Layer' section has three radio button options: 'Full Grid', 'Population Grid' (selected), and 'None'. The 'Cell Size' section has three radio button options: '1 km', '5 km', and '10 km' (selected). The 'Aggregation Type' section has three radio button options: 'Mean' (selected), 'Total', and 'SD'. Below this is the 'Select variable for Aggregation' section with a dropdown menu showing 'C1P2A'. The 'Select Single Grid Cell' section has an empty dropdown menu. At the bottom, there is a 'Download Data' button with a download icon. A warning box at the very bottom states: 'This may take a while, depending on the number of census units and the chosen raster resolution. Files must be generated before download!'

The final result displayed and exported contains only values from non-empty cells. With mouse over it is also possible to display the corresponding cell value. The height of each bar is its relative value,

¹¹ This means if a provided variable contains **any** string, even only as placeholder like "###.###" the system will **not** identify the variable as numeric, and subsequently you won't be able to select it for processing.

¹² When the grid is re-calculated, a small window popping up at the lower right side is seen, meaning that when changing values, as depending on cell size, this may take quite some time.

the color is only for distinction purposes, therefore no scale is provided.

Figure 4 - Bar height indicates aggregated value.



In addition, it also adds the variable DIST (which is the calculated distance). The latter is also available for aggregation. After Aggregation, user will receive a 3D plot with the height and color indicating the sum of the value from observations within the grid cell boundaries.

The population grid created for visualization, can also be downloaded in a **GeoTIFF raster**¹³ format. Such population grids have several use cases, one of them is the combination with other geo-spatial data ideally provided in a similar format, like aerial photography or remote sensing data. Furthermore, it also provides some degree of statistical disclosure control, if aggregation follows strict guidelines on minimum sizes.

3. CELL SIZE

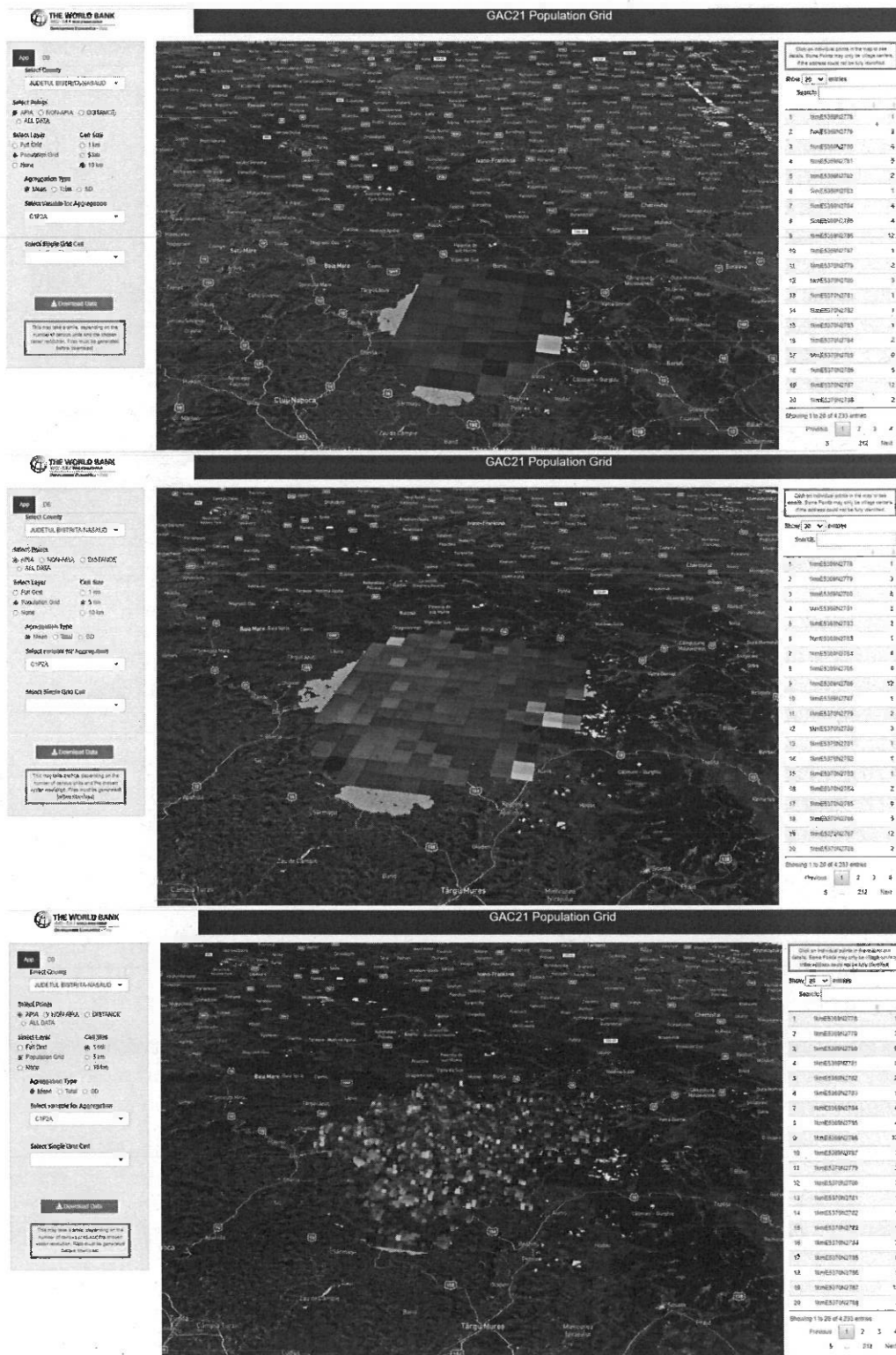
Cell size is allowed for 1000m, 5000m or 10000m. The cell size specifies the size of the FULL GRID cells and the size of the POPULATION GRID cells.

A smaller cell size requires a larger number of cells to cover the area, resulting in longer processing times and higher system requirements. It is therefore recommended to make sure computational resources have enough overhead.

Figures bellow show the different resolutions for the selected variable C1P2A for Bistrita-Nasaud.

¹³ [GeoTIFF, Revision 1.0 \(loc.gov\)](#)

Figure 5 - Different resolutions for the selected variable (10km, 5km, 1km)

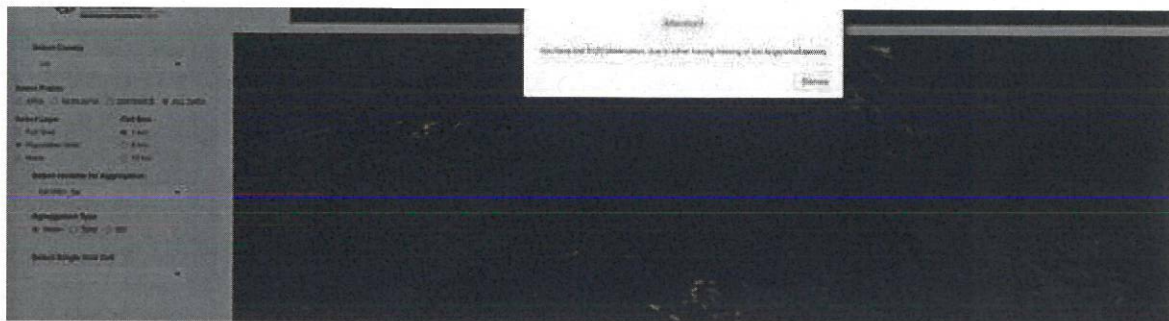


The processing environment should also be considered when selecting the resolution, as it can be seen above in the number of cells per resolution. Also, the downloaded raster file will be much larger, the smaller the resolution is.

The height of each bar is the corresponding statistics selected, i.e., mean, total, or the standard deviation of the selected variable. In this way any deviations and outliers can be easier caught by the eyeball. Placing the mouse pointer on any of the bars displays its height (i.e., the aggregated value, which is either mean, total or the standard deviation).

Since the application allows for selection and upload of any variables, for visualization purposes some extreme values may be dropped only for the mapping procedure. The application will tell user how many observations are not on the map.

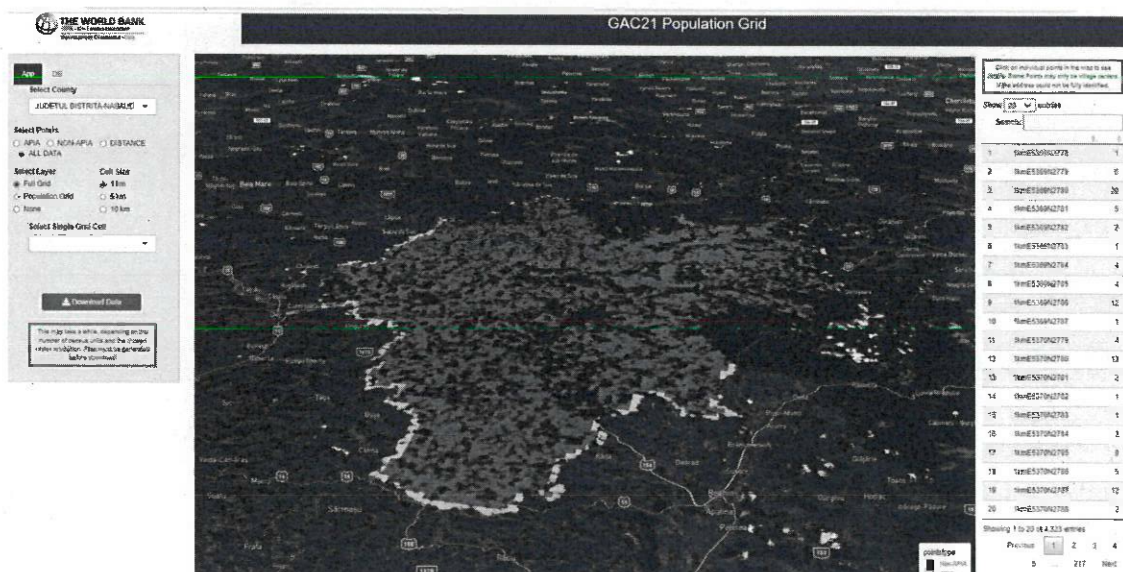
Figure 6 - Warning message informing user about exclusion of observation when creating the grid.



4. SELECT POINTS

This allows you to finally select and display the collected data. Selection is either the APIA data set, the NON-APIA data sets, the APIA data set with DISTANCE between original location and imputed location, or ALL DATA, combining APIA and Non-APIA datasets.

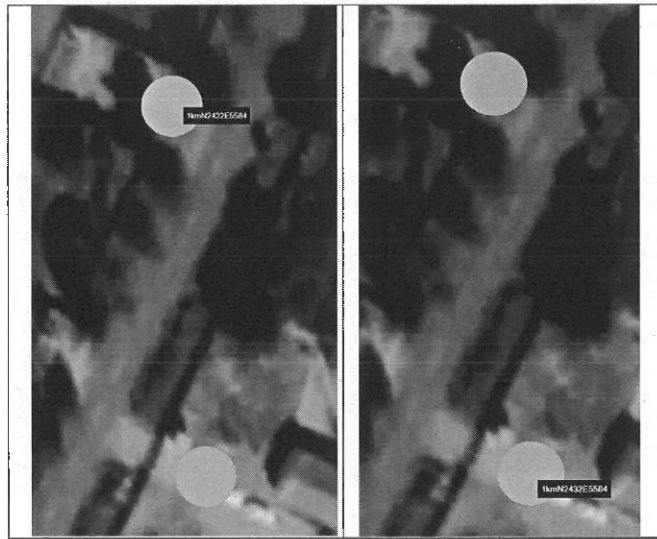
Figure 7 - All Points



The DISTANCE data is both data combined, and allows for the comparison of the imputed location and the actual location where the data was collected. It will be added to the set of variables and is displayed as point as well as gridded visualization, making it easier to spot any large deviations by the high of the grid cell.

Through **mouse over** it can be displayed the corresponding cell key, assigned to this census unit under recurrent coordinate system and as provided by the spherical coordinates in decimal degrees as well as the distance between the original and the imputed data.

Figure 8 – Coordinates displayed over point selected



By **clicking** on any point, it can also be displayed a table to the left, containing all variables for the corresponding observation. This means user can for example take the interview key and check further details in the Survey Solutions system.

Figure 9 – Side Table with all variables (right)



DISTANCE data allows for comparison of “old” and “new” coordinates. Since some of the coordinates had been imputed, this data set shows the old and the new coordinates on the same map and allows for a direct comparison.

Figure 10 - The distance data allows for inspection of larger distances

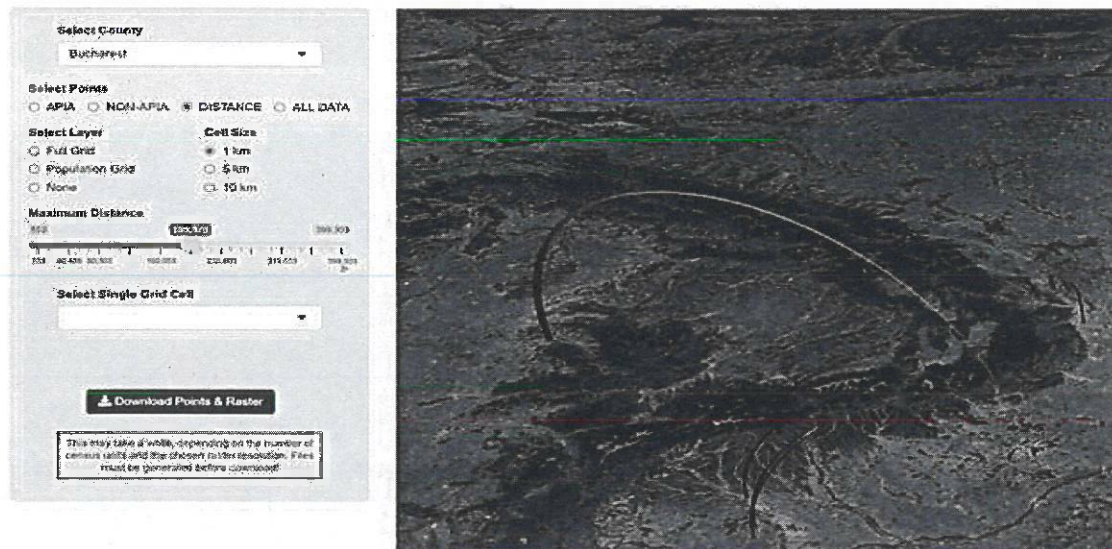


Figure 11 - The distance data allows for inspection of smaller distances



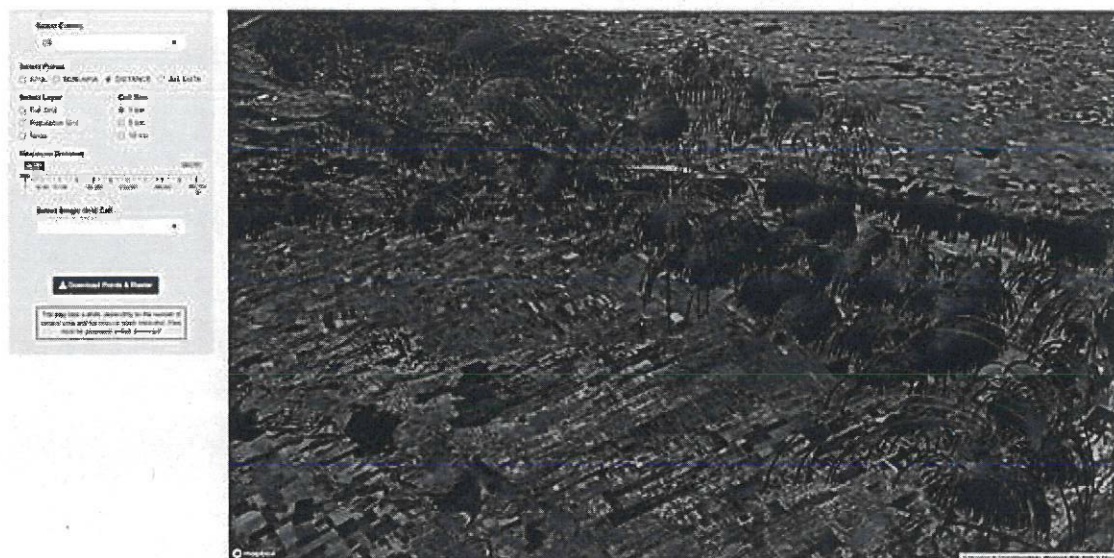
To facilitate any distance-based inspection and observe only a subset, the GA also shows a Slider Maximum Distance to regulate the maximum distance to display. This is in particular useful, with many units nearby, and only a few far away.

Figure 12 - Dynamic Controls adjust to the minimum and maximum distance.



Sub-setting the data on a particular Maximum Distance reloads the map and allows for a better inspection of these units.

Figure 13 - Sub-setting of units by their maximum distance is possible



ALL DATA allows for the direct visualization of APIA and Non-APIA units. This is in particular useful, when somebody wants to inspect the distribution in a narrower geographic area or within a grid cell, see figures below.

Figure 14 - APIA and Non-APIA units without and with grid

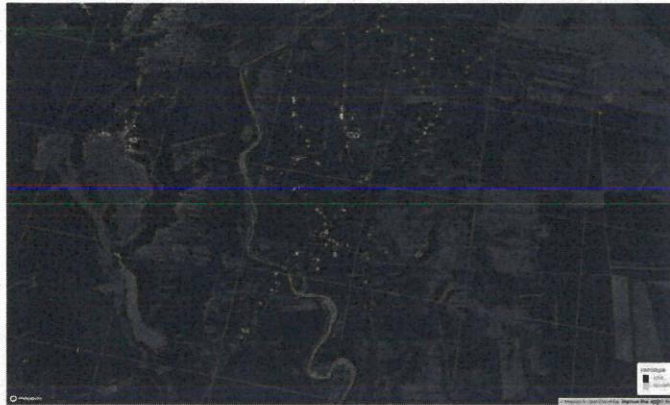


Figure 15 - ALL DATA and single grid cell

Select County:

Select Points: ☐ APIA ☐ Non-APIA ☐ Display ☐ All Data

Select Layer: ☒ All Data ☐ 500 m ☐ 1000 m ☐ 1500 m

Select Single Grid Cell:

This tool will create a shapefile for the number of selected units and the resolution selected. The data will be downloaded to your computer.



Grid IDs

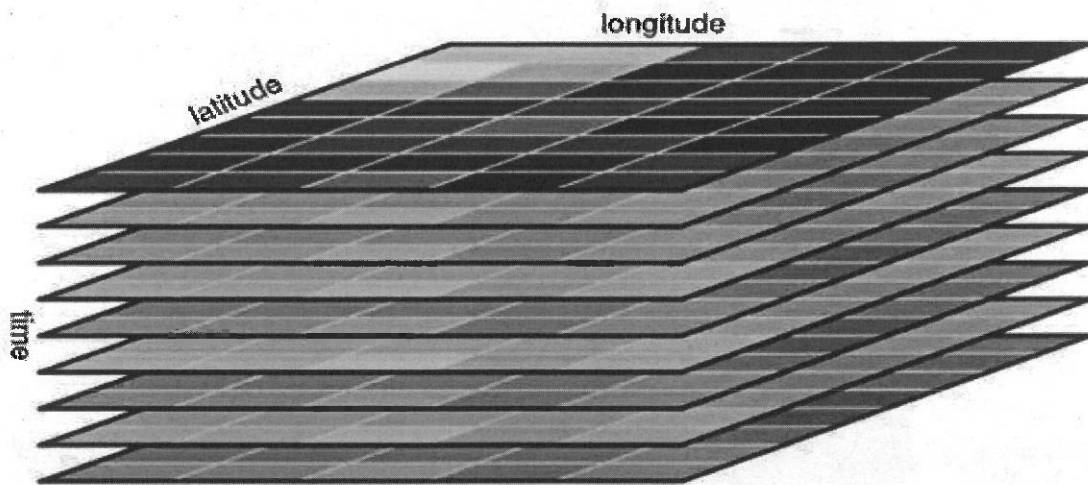
Grid Identifiers (IDs) are produced in line with INSPIRE directive (according with EU Regulation 2018/1098) and allow to organize (aggregate) census units at the level of a predefined distance. The first step is therefore the projection of the coordinates into the coordinate space, which is the corresponding CRS. This coordinate space was intended to facilitate the production of geo-referenced statistics and forms the foundation for the Unified European Grid Coding system has the following attributes:

- easy to manipulate,
- hierarchical,
- based on a Unified European Grid Coding System,
- based on units of equal area,
- adopt ETRS-LAEA

While reporting 1 km grid values and codes as part of INS obligation towards Eurostat, it is also very useful, to continue using the grid system for other sizes as provided from the system. This allows to efficiently store geo-referenced data.

Since standard population-based census units are considered to be stationary, such a grid code is created once, and can then be used in subsequent data collection activities, like surveys. Geo-spatial analysis and visualization is facilitated when storing data in this way.

Figure 16 - Data Cube for a single feature



Source: from <https://keen-swartz-3146c4.netlify.app/datacube.html>

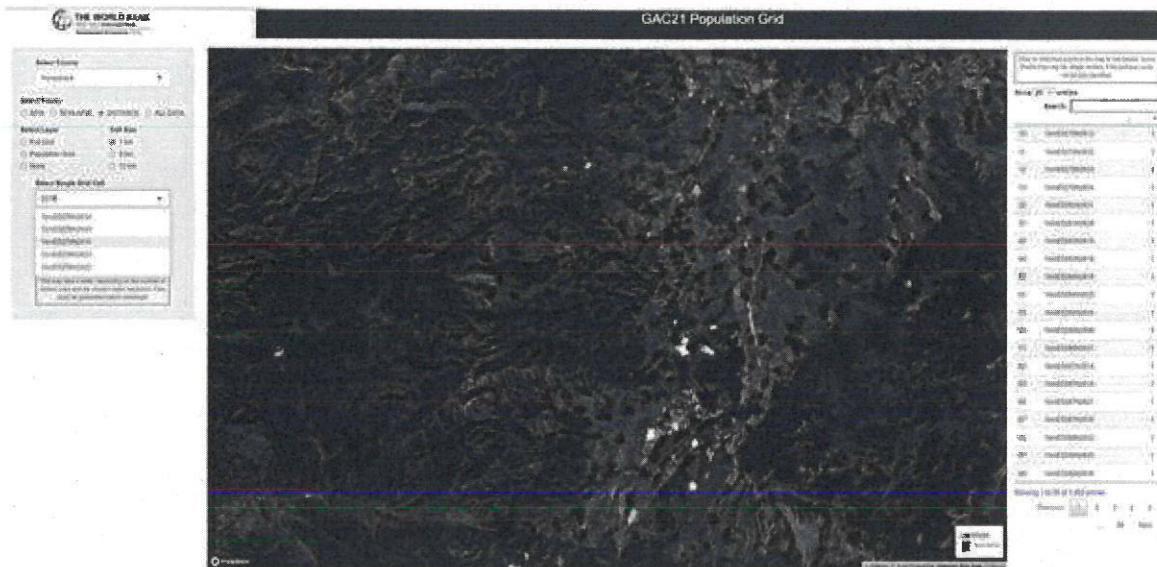
The different layers indicate the collected data over time for the different units and aggregated to its grid cell values. It is an efficient way to store spatiotemporal data¹⁴. The codes are thus generated only once and in line with its underlying population grid (i.e., 1km, 5km or 10km).

Subsequent data processing activities can aggregate tabular observations by grid code and immediately report results at the cell level, instead of preprocessing the information in its spatial output. In addition, such data format also facilitates analytic purposes, survey sampling designs and spatial predictions.

¹⁴ Spatial Data Science with Applications in R, <https://keen-swartz-3146c4.netlify.app/>

Since one of the relevant privacy checks is the number of units per grid cell, the app displays a table of grid cell codes, with their counts, which can be ordered (i.e., from smallest to largest) or searched.

Figure 17 - Grid Cell Selection to identify singletons or other privacy violating cells.



The value can then be used in the drop-down grid cell selection on the left-hand side, to select such a grid cell for identification.

Figure 18 - Selected Grid Cell for further inspection and follow up.



Download Formats

Current download formats are a zip file which contains.

- a) a .csv file containing the units data and their assigned grid codes (two files .csv within zip archive)
- b) a .tif file following the GeoTIFF 1.1, 2019 convention (including CRS); the tiff file appears only if variables are selected for aggregation, and it depends on the selection (see Figure 29 and Figure 31) <https://earthdata.nasa.gov/esdis/eso/standards-and-references/geotiff>.

Additional remarks:

- The resulting grid can be exported and used for other analytical purposes (e.g. download the grids by county).
- The application uses deck.gl, which is open source for the maps. Check the website for details.
- The application requires a Mapbox license key (<https://www.mapbox.com/>), which is free under certain restrictions. If no mapbox key is provided, the application still produces the grid, however will not show any visualizations.

Download the reference grid

The reference grid can be downloaded directly from the DB section and produces the grid file, delivered as an ESRI shape file and as used by EUROSTAT. It contains the grid for all the country as a single shape file.

2. Data preparation and variables by type of farm

Organization of the GAC database

The organization of the GAC database, according to the questionnaire implemented in Survey Solutions, that includes the information needed for preparing the geo-referencing, resides in:

1. "Identification Data" contains the variables of the SuSo questionnaire related to the identification data of the agricultural holding;
2. "Areas, Livestock, Animal housing, and other" contains the variables of the SuSo questionnaire regarding the areas, animals and other data of the agricultural holding;
3. "Labor Force" contains the variables of the SuSo questionnaire related to the labor force of the agricultural holding;
4. "Sales" contains the variables of the SuSo questionnaire related to the sales of own production of the agricultural holding.

The collected data from GAC were exported from Survey Solutions then transformed into micro-data of GAC, according to the description elaborated together with INS Team. The table of correspondence between variables is detailed in the **Annex 3**.

The **microdata** are structured by 5 sections, splitted in 42 files each, one per each county of Romania (*nn* - counties codes), as follow:

- ID*nn* files - contains the variables related to general information of the agricultural holding;
- ALT*nn* files - contains the variables for Animal housing; Manure management, and others Organic farming; Other gainful activities, Rural development, and others;
- FM*nn* files - contains the variables for Labor force;
- SA*nn* files - contains the variables for Land use; Livestock,
- VZ*nn* files - contains the variables for sales,

The key variables which link these 5 types of files, corresponding to the described 5 sections are:

- i_key/interview_key
- Number of the folder (C1P1a/ HLD_ID) combined with Number of the questionnaire within the folder, C1P2a/ HLD_ID2, (Map + Form). The first 2 digits of the map represents the county SIRUTA code, even if the information could be redundant with theSirjud/C01P01_3.

The data preparation for geo-references the agricultural holdings on grid consists in combining the information from 4 data sources, according to the accuracy of the GPS information, as follow:

- microdata from ID*nn* files (3225908 records), which represents the GAC collected data, organized as microdata;
- APIA shape files (ipa_2020_changed_centroids.zip - 6046848 records), which represent an administrative data source, provided by APIA, containing the GPS location of each parcel of the agricultural holding with high precision (centroid of the parcel);
- APIA areas (APIA_2020_suprafete_sept_2021.xlsx - 1001440 records), which represent an administrative data source, provided by APIA, used for identification of APIA

Code for the persons who don't have APIA Code in GAC collected data (being registered in APIA), but they have ISU, C1P311 (C01P04_1d) or C1P321 (C01P04_2b);

- centroids of localities ("Sate_centroide_latlon.csv") (12959 records), which represents the database with the GPS location of the centroids for each locality in Romania, used for replacing the NA values for GPS location for agricultural holdings collected with errors or not collected and that could not be corrected with data from other sources.

The collected values for the variables C1P2B_1 (GPS_Latitude) and C1P2B_2 (GPS_Longitude) are checked according to the geographic limits of Romania, Nord, South, Est and West (N = 48.265083, S = 43.621100, E = 29.7, W = 20.261759). The non-valid values were changed to NA.

The C1P2B_1 (GPS_Latitude) and C1P2B_2 (GPS_Longitude) data were corrected based on APIA data, as the coordinates of the holding must correspond to the location of the holding, not to the place where the interview was done, (APIA shape files), according to these steps, in this order:

- for all the APIA codes (FARM_ID) which is match one-to-one to the collected data (C1P1B), the collected coordinates were replaced with the APIA coordinates which correct correspond to the location of the holdings (38236 records);
- for all the APIA codes (FARM_ID) which is match one-to-many, one in collected data (C1P1B), many in APIA data, the collected coordinates were replaced with the APIA coordinates (627520 records), based on the agricultural holding having the largest land area, which also correct correspond to the location of the holding, $\max(\text{area_decla})$;
- for all the APIA codes (FARM_ID) which is match many-to-many, many in collected data (C1P1B), many in APIA, the collected coordinates were replaced with the APIA coordinates (67612 records), matching all records where $\text{SIRCOM} = \text{sirsup_cod}$, on largest area ($\max(\text{area_decla})$);
- for all the APIA codes (FARM_ID) which is match many-to-many, many in collected data (C1P1B), many in APIA, the collected coordinates were replaced with the APIA coordinates (129 records), identifying the APIA code by the ISU code from "APIA_2020_supraf- ete_sept_2021.csv" file (key=CNP_CUI APIA CODE = IDF; key=C1P311 APIA CODE = C1P1B);
- replace NA (as the coordinates were missing from different reasons) from C1P2B_1 and C1P2B_2 with the centroids of the locality from "Sate_centroide_latlon.csv" file, by key=SIRUTA_SUP (C1P2B_2=xcoord, C1P2B_1=ycoord). (18188 records).

The database variables for geo-references the agricultural holdings are presented in the **Annex 4**. The R script for data matching and correction of GPS coordinates is presented in the **Annex 5**.

Update the database

The purpose of updating database is to support further changes of it in terms of content or variables obtained from other (structural) surveys or for additional data analysis.

As the data is available at INS in FoxPro export files, the data is preprocessed to meet contemporary data standards. This is done by the script shown and documented in **Annex 6**.

Figure 19 - Update database commands

App

DB

Update Database

⬇️ Update Database with new data

Updating the database will run a script in the background, which reads the data from the provided FoxPro database and stores it in a PostgreSQL database. This is only required, if the FoxPro database export has been updated. For more details see the corresponding chapter in the report

Download Grid

⬇️ Download Reference Grid

This allows you to download the national reference grid, including population counts in a single file.

MapBox Key

Provide MapBox Key

Your mapbox key

The background maps require a mapbox key (<https://www.mapbox.com/>). Without the key, the map visualization won't work. However you can also process the data, without any visualization. For more details see the corresponding chapter in the report

If any new data is added, it needs to be placed on the original directory on the server, and subsequently the Update Database with new data button needs to be activated. Important to note in this respect is that the database schema must not change. This means variable names and format needs to be the same even in case of any new data.

3. Installing and configuring the application

Source code description

The source code contains the standard R shiny application set-up, consisting of:

- **app.R** file which contains all both the UI and the Server file,
- **helpers**: a directory of helper functions
- **data**: a directory of static data files
- **www**: a directory of web resources, like logo and css style sheets.

System requirements (hardware and software dependencies)

The standard set-up as recommended in the Output 4.1b: *“Report on advisory services provided to Recipient on the Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)”* is sufficient. No extra hardware requirements are necessary.

For running the application, a valid R installation is required as described here: <https://www.digitalocean.com/community/tutorials/how-to-install-r-on-ubuntu-18-04>

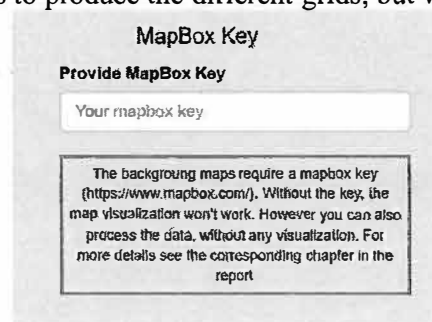
As well as a valid shiny open sources installation as described here: <https://www.rstudio.com/products/shiny/download-server/ubuntu/>

For all of them default configuration is sufficient.

Installation kit, installation procedure and configuration parameters

As the hardware and software components of the integrated INS system are available, the World Bank performed the installation and configuration for all software components (programs, licenses, scripts, etc.) necessary for the proper functioning of the application and they are available on INS “analytics server” - link [xx.x.xxx.xx/rga-grid/](#).

For the maps a free Mapbox subscription is required. A valid subscription can be obtained here: <https://www.mapbox.com/>. After the Mapbox key is obtained it will be included in the space indicated in application (see below). However, the application also works without a mapbox key. In this case the application allows to produce the different grids, but will not display any visualizations.



The screenshot shows a web form titled "MapBox Key". Below the title is a label "Provide MapBox Key" and a text input field with the placeholder text "Your mapbox key". Below the input field is a rectangular box containing the following text: "The background maps require a mapbox key (<https://www.mapbox.com/>). Without the key, the map visualization won't work. However you can also process the data, without any visualization. For more details see the corresponding chapter in the report".

4. Recommendations and best practices

The recommendations presented below are in line and incorporate best practices for developing, maintaining, and exploiting applications that are managing high amount of data in a georeferencing process. The recommendations are focusing on data, on organizational and capabilities related aspects. The purpose of applying them is the optimal use of application with GAC2020 collected data. However, the recommendations should be considered in the future, when using the application in another context.

Regarding the data, the following are recommended:

- The assignment of the correct GPS is part of the data collection (or the frame building exercise prior to the census) and its organization, for which were also made some recommendations, i.e., only take coordinates at place of holding, pre-load APIA coordinates etc. The application deals with the data from this point onwards.
- Data needs to be provided in the same way as outlined by the R script in the Appendix 5 about all the transformation steps.
- Frame Building should be done with dedicated software solutions since this will facilitate further processing. MS Excel and similar solutions are not recommended for such an important task. Survey Solutions can also be considered for such internal tasks.
- A stronger cooperation with administrative data producers to allow access to their administrative data (see lack of administrative data sources from the ANCPI - Agency for Cadastre and Real Estate Advertising) and a better knowledge about the quality of their data (see APIA as source of data for the georeferenced of agricultural holdings).
- A harmonized (spatial) data infrastructure will address several issues encountered during post processing. This also requires an integrated IT infrastructure.
- The organization into grid like units can then be done with any GPS coordinates, it does not require specific coordinates, important to note here. Depending on the local data environment, is recommended storing the location in the local geo-spatial projection (CRS), and only transform it to spherical (decimal) coordinates when required. This facilitates storage(decimal vs. integer) as well as any space-based calculations, like distance.
- The collected coordinates could be verified to be inside the polygon of the locality and adjusted accordingly. Such a procedure can be automated in future activities.
- If coordinates are preloaded, Survey Solutions allows to geo-reference the questionnaire. In this way INS can make sure, that the holding coordinates are at least within the locality polygon.
- In the latest dataset were **72,705 records** with no interview key, and completely empty rows. as imputed from APIA administrative source of data. As they were not collected the interview key is not available. To still being able to use these records in the applications, a synthetic interview key was generated.
- Item non-response should follow clear and transparent imputation procedures. A methodology and applicable documentation for imputation procedures are recommended and should be in place at the beginning of the census or survey for statistics production.
- Geolocations during field data collections should be taken at the location of the agricultural holding, and not at the location where the interview is carried out. Imputation of the

location should also follow clear and transparent rules.

- To have consistency in the variable I_KEY, as best practice and recommendation for future works, an interviewkey is generated for all records without key (NA). Any corrections have to take place on the source data. Data cleaning practices implemented should avoid adding the ID_variables when they are missing.

One recommendation is about the skills and capabilities of specialists working with the application in the view of future developments as needs evolve and the preparation of the adequate database. In this respect, is very important to integrate permanent knowledges on R and R Studio, Shiny programing languages and GIS within the team/division of INS that manage the agriculture statistics, as a permanent duty and practice. This could be implemented for other statistics departments of INS, also, as this is a practice applied in peer-institutions at EU level and has been observed during recently exchanged of experience (in Poland and in The Netherlands).

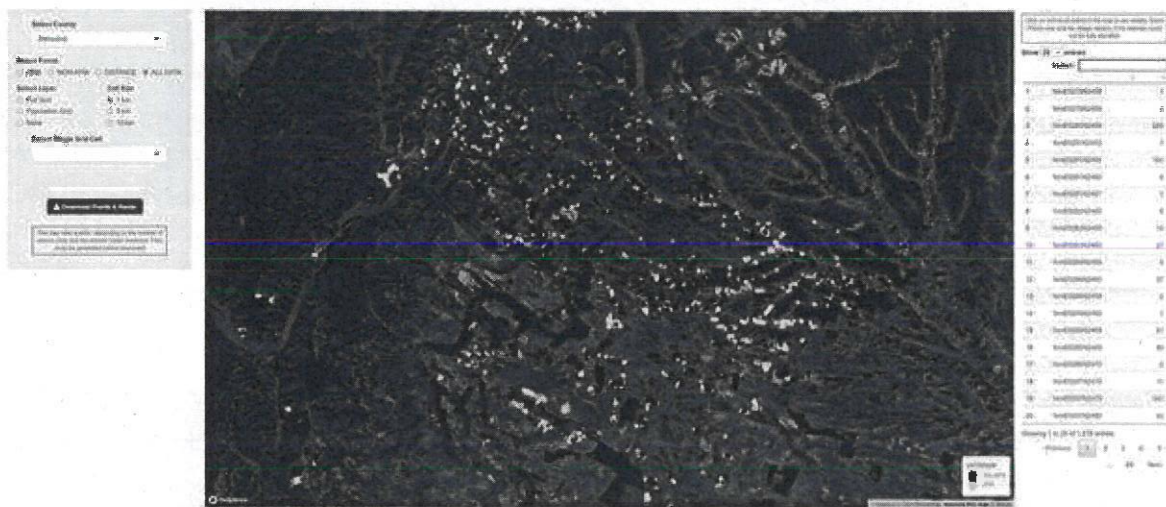
The gaining of capabilities and skills could be done through trainings and on-the-job practice (the list of training courses is provided within “*Development of a multi-annual training needs plan (2021-2024) for INS including STDs based on training needs*” of Output 12 under the same RAS Agreement) either through hiring graduated staff with background in statistics and IT. This approach could avoid the shortage of specialists or the high workload that INS faces within IT department.

5. Annexes

Annex 1: Filtering of observations by selection from grid cell summary table

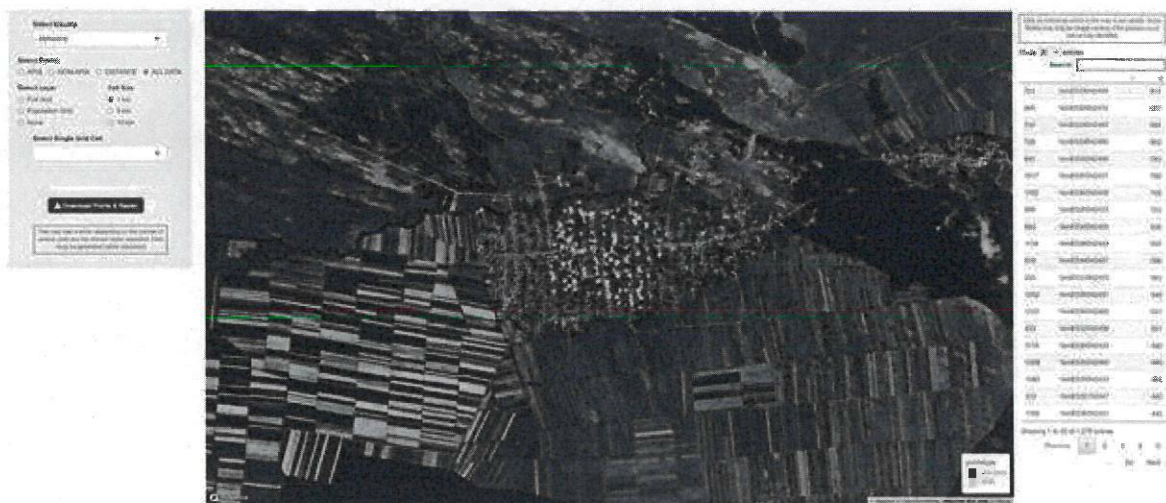
Select Single Grid Cell already allows the user to select individual grid cell by its grid cell code. To also enable a more detailed inspection of unit counts per grid cell, and identify singletons, but also other numbers of units per cell, the application also allows the user to select any line from the table to the right, which shows the number of grid cell counts, and is created after Select Points option has been selected.

Figure 20 - Table of grid cell counts to the right of the map



The cell count table can be sorted by the highest number of units per cell, ...

Figure 21 - Sort by highest



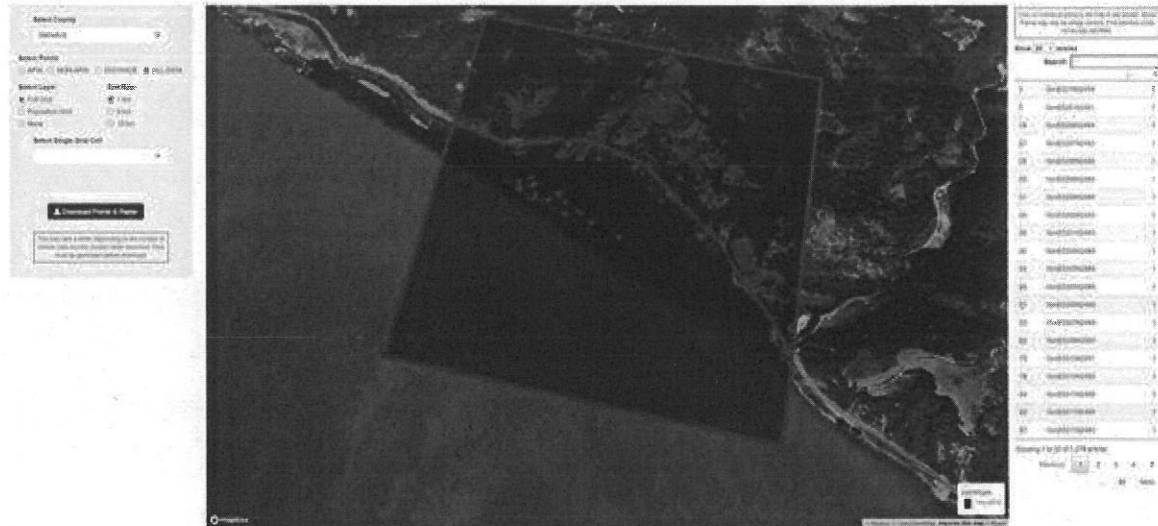
Or by lowest number of units per cell.

Figure 22 - Sort by lowest



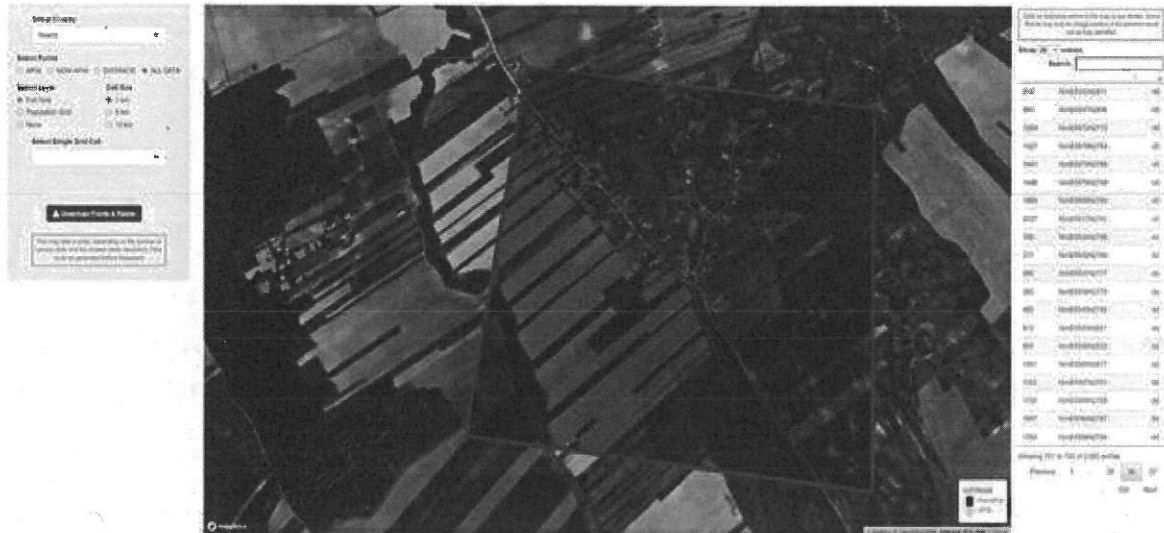
Also, the corresponding grid cell can be added.

Figure 23 - Selected grid cell



And any cell can be selected on any page of the table.

Figure 24 - Select medium size cell



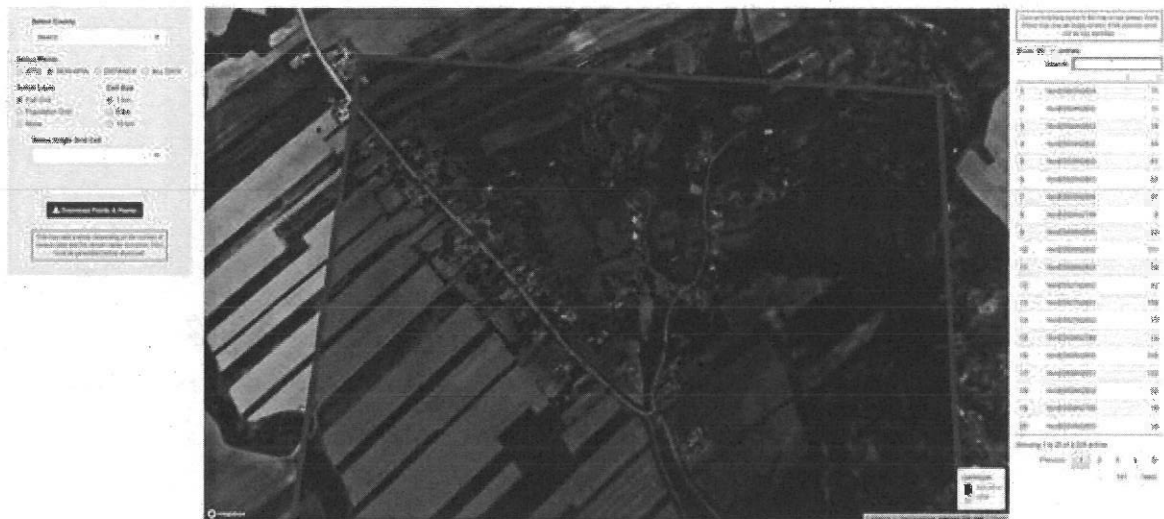
Even the points data set can be switched between different Select Points layers showing Apia data ...

Figure 25 - APIA units inside grid cell



Or Non-APIA units for the same grid cell.

Figure 26 - Non-APIA units inside grid cell



And certainly, it is also possible to select from different base grids, either like above in 1 km cells, or as shown below with 5 km grid cells,

Figure 27 - 5 km grid cell size cell selection



and, also with 10 km cell size.

Figure 28 - 10 km grid cell size cell selection



In this way a more detailed (spatial) monitoring **during and after** data collection is possible. All the created resources, i.e., grid cells can be exported if desired and sent to the Survey Solutions CASS for use in further data collection activities¹⁵. This may for example be required in case of post enumeration surveys or other similar activities. Some processes can even be automated however do require a clear definition of automation tasks.

¹⁵ Currently we only export the files described in the main manual, however any output produced internally can also be exported.

Annex 2: Data Download

Commonly the final step in a spatial transformation as provided by this application, is to send the data to the next processing step. In an integrated data processing environment this processing step could either result in:

- storage in a data base or
- further processing either for
 - data verification
 - coverage checks,
 - other monitoring purposes.

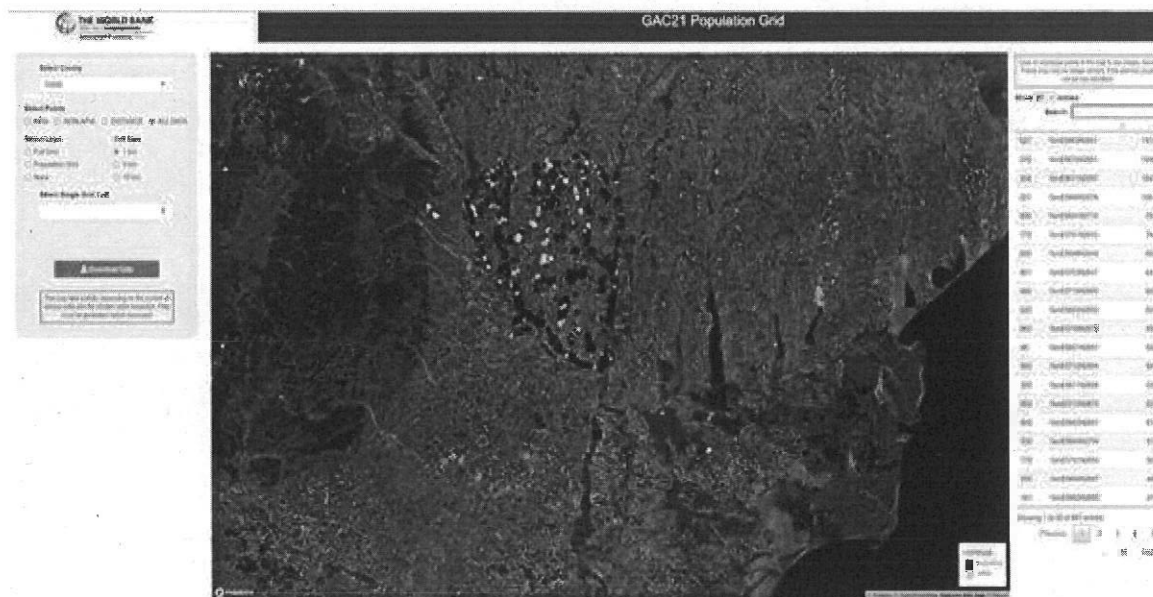
If in any of the further processing steps additional **primary data collection** is involved either by web, phone or in person, then required support files, like i.e., identification data or even cell boundaries can be exported, and even at resolutions as low as 100m¹⁶.

Depending on the selected area, the resolution and the input data, the file creation and compression process may take some time.

The points data

The first data set available for download, after selecting the County in Select County is the points data.

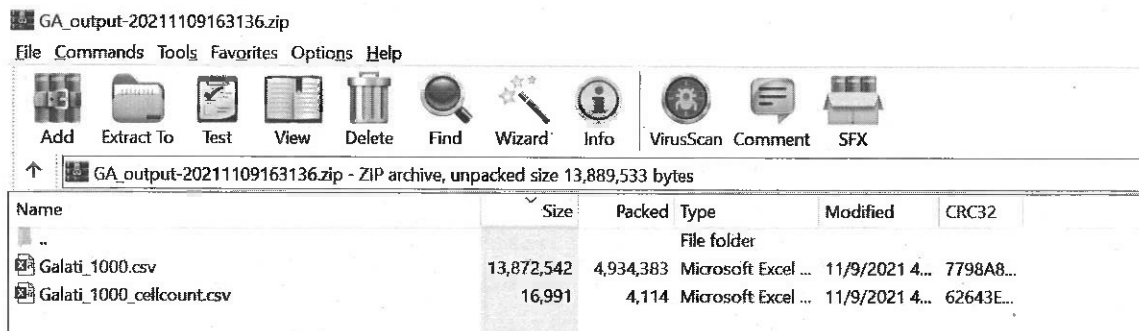
Figure 29 - Loading the data for a particular county creates the points file and the cell count files for download



16100 m is used in several machine learning applications on gridded population data. It would therefore be possible to use this publicly available data sets in conjunction, or also with other population raster data.

Opening the .zip file gives the following directory view:

Figure 30 - .zip file with points data and cell count.



The Cell Count data

This file just contains the same data as displayed on the right-hand side of the application, which is the count of census units, by grid cell, **for the selected resolution only**.

If it is required to download the data, and the codes at different resolutions, should be required a separate download each time is changed any of the resolution. The points data and the cell count data are always included after the data for a specific county has been selected. There should be noted that for the whole country the reference grid is downloaded, only (see Download reference grid, under Section 1).

The raster files

The raster file is only included if, from the options available under Select Layer, the option **Population Grid** is selected.

Figure 31 - The population raster can be downloaded after it is visible on the map.



The included layer will contain the selected Parameter [var], aggregated to the specific aggregation type [type] and at the desired resolution [res].

Figure 32 - .zip data with points data, cell count and raster data.

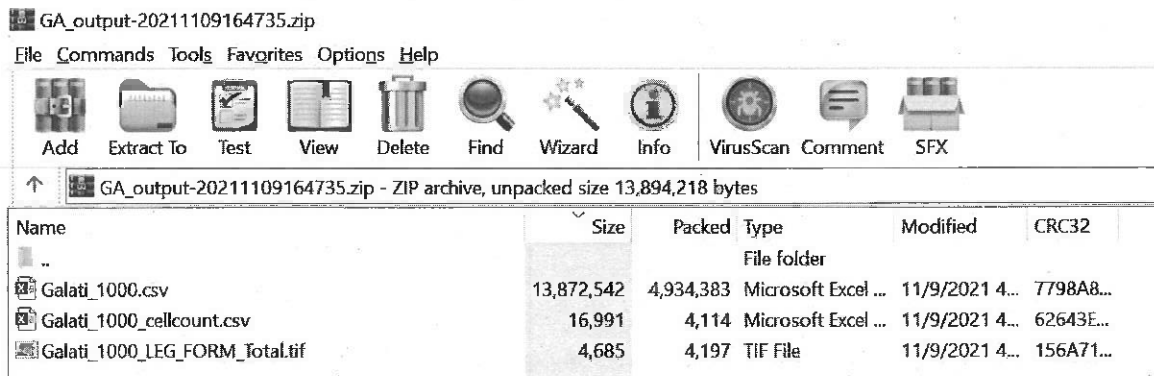
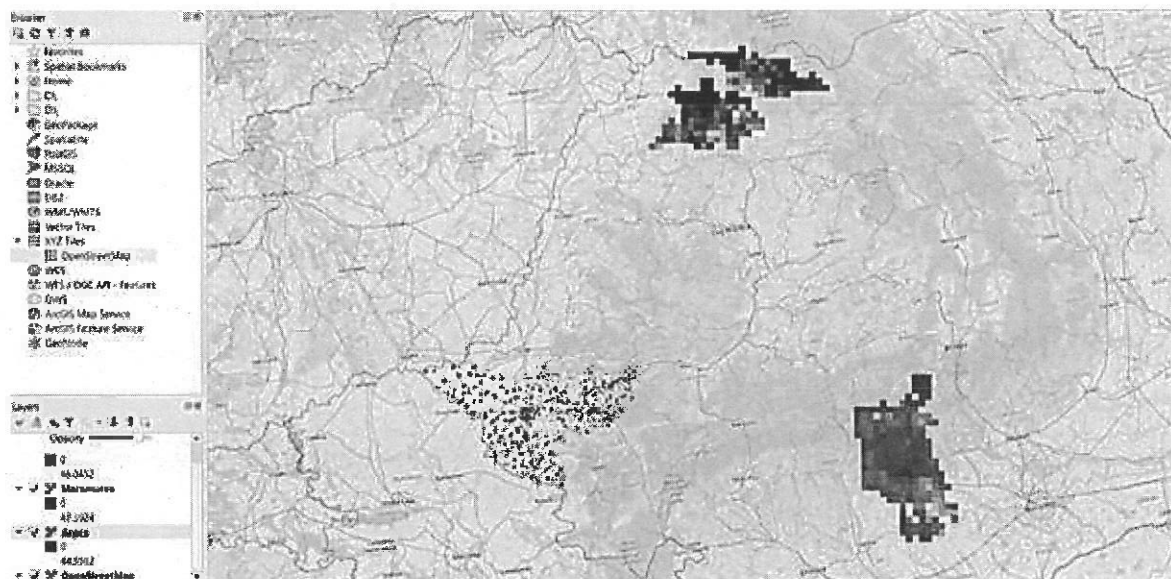


Figure 33:

After downloading the file and unpacking it you can then display the raster also in your favorite GIS software. For each county you can download the raster file with different resolutions, but all compatible with the INSPIRE coding structure.

Figure 33 - QGIS display of raster data for different counties and different resolutions all in a single map.



If it is required to download the data, and the codes at different resolutions, the user has to initiate a separate download each time he change any of the input parameters, like the resolution, the selected data (variable) or the selected aggregation type.

Annex 3: GAC microdata

IDnn files - contains the variables related to general information of the agricultural holding.

VARIABLE	COD	Expected Value or Range	Variable app	Type	Collected value / examples	Categories/ Nomenclature/ Classification structure	SuSo	No
INFORMAȚII GENERALE PRIVIND EXPLOATAȚIA AGRICOLĂ								
LOCALIZAREA EXPLOATAȚIEI AGRICOLE								
Cheia interviului/chestionarului	i_key	text	inter-view_key	c	94-95-87-87		GAC20_Roma-nia9.tab	
Numărul mapei	C1P1a	valoare numerica	HLD_ID	i	280098		GAC20_Roma-nia9.tab	
Cod identificare APIA	C1P1b	text sau blank	C01P01_1b	c	999		GAC20_Roma-nia9.tab	
Numărul chestionarului în cadrul mapei	C1P2a	valoare numerica	HLD_ID2	i	163		GAC20_Roma-nia9.tab	
GPS_Lat	C1P2b_1	coord.GPS sau blank	visitLocation_Latitude	d	47.78160932		visit_summary.tab	
GPS_Long	C1P2b_2	coord.GPS sau blank	visitLocation_Longitude	d	24.76437437		visit_summary.tab	
Județul	C1P3	text						
Comuna/ Orașul/ Municipiul	C1P4	text						
Localitatea componentă (satul)	C1P5a	text						
Cod SIRUTA judet	Sirjud	cod Siruta	C01P01_3	i	38/County Siruta Code	Options-in-question-1.3. County.txt	GAC20_Roma-nia9.tab	
Cod SIRUTA comuna	Sircom	cod Siruta	C01P01_4	i	14450/City Siruta Code	Options-in-question-1.4. Municipality_Commune_Town.txt	GAC20_Roma-nia9.tab	
Cod SIRUTA sat	Sirsat	cod Siruta	C01P01_5a	i	14478/Village Siruta Code	Options-in-question-1.5a. Component locality (village).txt	GAC20_Roma-nia9.tab	
Strada	C1P6a	text sau blank	C01P01_6a	c	Bercheza/999		GAC20_Roma-nia9.tab	
Nr.	C1P6b	text sau blank	C01P01_6b	c	43/999		GAC20_Roma-nia9.tab	

Bloc	C1P6c	text sau blank	C01P01_6c	c	A3/999		GAC20_Roma-nia9.tab	
Scara	C1P6d	text sau blank	C01P01_6d	c	B/999		GAC20_Roma-nia9.tab	
Etaj	C1P6e	text sau blank	C01P01_6e	c	2/999		GAC20_Roma-nia9.tab	
Apart.	C1P6f	text sau blank	C01P01_6f	c	14/999		GAC20_Roma-nia9.tab	
Sector	C1P6g	text sau blank	C01P01_6g	c	1/999		GAC20_Roma-nia9.tab	
Telefon	C1P6h	text sau blank	C01P01_6h	c	0211234125/999		GAC20_Roma-nia9.tab	
Fax	C1P6i	text sau blank	C01P01_6j	c	0211234125/999		GAC20_Roma-nia9.tab	
Adresa e-mail	C1P6j	text sau blank	C01P01_6i	c	sample@site.com/999		GAC20_Roma-nia9.tab	
STATUT JURIDIC AL EXPLOATAȚIEI AGRICOLE								
Exploatații agricole fără personalitate juridică								
Exploatație agricolă individuală	C1P2s	valoare numerica	LEG_FOR M	i	1		GAC20_Roma-nia9.tab	1
Persoană fizică autorizată, întreprindere individuală	C1P2s	valoare numerica	LEG_FOR M	i	2		GAC20_Roma-nia9.tab	2
Întreprindere familială	C1P2s	valoare numerica	LEG_FOR M	i	3		GAC20_Roma-nia9.tab	3
Exploatații agricole cu personalitate juridică								
Regie autonomă	C1P2s	valoare numerica	LEG_FOR M	i	4		GAC20_Roma-nia9.tab	4
Societate/ asociație agricolă (Legea nr. 36/ 1991)	C1P2s	valoare numerica	LEG_FOR M	i	5		GAC20_Roma-nia9.tab	5
Societate comercială cu capital majoritar privat (Legea nr. 31/ 1990)	C1P2s	valoare numerica	LEG_FOR M	i	6		GAC20_Roma-nia9.tab	6
Societate comercială cu capital majoritar de stat (Legea nr. 31/ 1990)	C1P2s	valoare numerica	LEG_FOR M	i	7		GAC20_Roma-nia9.tab	7
Institut/ stațiune de cercetare, unitate școlară cu profil agricol	C1P2s	valoare numerica	LEG_FOR M	i	8		GAC20_Roma-nia9.tab	8
Primărie	C1P2s	valoare numerica	LEG_FOR M	i	9		GAC20_Roma-nia9.tab	9

Alte instituții publice	C1P2s	valoare numerică	LEG_FORM	i	10		GAC20_Roma-nia9.tab	10
Unitate cooperativă	C1P2s	valoare numerică	LEG_FORM	i	11		GAC20_Roma-nia9.tab	11
Alte tipuri (fundatie, asezamant religios, școală etc.)	C1P2s	valoare numerică	LEG_FORM	i	12		GAC20_Roma-nia9.tab	12
Exploatația agricolă cu personalitate juridică face parte dintr-un grup de exploatații agricole, altul decât cele menționate	C1P23	1 sau 2	C01P03b	i	1(Da) / 2 (Nu)		GAC20_Roma-nia9.tab	
DATE DE IDENTIFICARE								
Exploatații agricole fără personalitate juridică								
Numele, inițiala tatălui și prenumele capului exploatației agricole (utilizatorul)	C1P31a	text	C01P04_1a	c	text		GAC20_Roma-nia9.tab	
Cod numeric personal (CNP) ISU	C1P311	text sau blank	C01P04_1d	c	13 digits / 999		GAC20_Roma-nia9.tab	
Cod unic de identificare (CUI)/ Cod fiscal	C1P312	text sau blank	C01P04_1e	c	n digits / 999		GAC20_Roma-nia9.tab	
Exploatații agricole cu personalitate juridică								
Denumirea exploatației agricole cu personalitate juridică	C1P32	text sau blank	C01P04_2a	c	text		GAC20_Roma-nia9.tab	
Cod unic de identificare (CUI)/ Cod fiscal	C1P321	text sau blank	C01P04_2b	c	text		GAC20_Roma-nia9.tab	
Adresa capului exploatației agricole fără personalitate juridică (utilizatorul) sau a sediului social al exploatației agricole cu personalitate juridică								
Localitatea componentă (satul) pentru capul exploatației agricole fără personalitate juridică	C1P33a	text sau blank	C01P04_3a	c	text		GAC20_Roma-nia9.tab	
Strada	C1P33c	text sau blank	C01P04_3b	c	text		GAC20_Roma-nia9.tab	
Nr.	C1P33d	text sau blank	C01P04_3c	c	text		GAC20_Roma-nia9.tab	
Bloc	C1P33e	text sau blank	C01P04_3d	c	text		GAC20_Roma-nia9.tab	
Scara	C1P33f	text sau blank	C01P04_3e	c	text		GAC20_Roma-nia9.tab	

Etaj	C1P3 3g	text sau blank	C01P04_3f	c	text		GAC20_R oma- nia9.tab	
Apart.	C1P3 3h	text sau blank	C01P04_3 g	c	text		GAC20_R oma- nia9.tab	
Sector	C1P3 3i	text sau blank	C01P04_3 h	c	text		GAC20_R oma- nia9.tab	
Telefon	C1P3 3j	text sau blank	C01P04_3i	c	text		GAC20_R oma- nia9.tab	
Fax	C1P3 3k	text sau blank	C01P04_3 k	c	text		GAC20_R oma- nia9.tab	
Adresa e-mail	C1P3 3l	text sau blank	C01P04_3j	c	text		GAC20_R oma- nia9.tab	
COD DE COMPLETITU- DINE								
Interviu complet	C13P 45	valoare nu- merica sau zero	C13P45	i	1-complet / 2-partial		visit_sum- mary.tab	
Exploatație agricolă desfi- ințată	C01P 02b	valoare nu- merica sau zero	C01P02b	i	2		visit_sum- mary.tab	1
Exploatație agricolă tem- porar fără activitate	C01P 02b	valoare nu- merica sau zero	C01P02b	i	3		visit_sum- mary.tab	2
Interviu refuzat	C01P 02b	valoare nu- merica sau zero	C01P02b	i	1		visit_sum- mary.tab	3
Alte situații (exploatație agricolă neidentificată, necontactată etc.)	C01P 02b	valoare nu- merica sau zero	C01P02b	i	4		visit_sum- mary.tab	4

Annex 4: Variables for geo-references the agricultural holdings

Microdata (IDnn files)

The names of variables *visitLocation_Latitude* and *visitLocation_Longitude*, from Survey Solutions questionnaire, in GAC microdata are *C1P2B_1* for GPS_Latitude and *C1P2B_2* for GPS_Longitude.

The variables names are: "I_KEY" "C1P1B" "C1P311" "C1P312" "C1P321" "C1P3" "C1P4" "C1P5A" "SIRJUD" "SIRCOM" "SIRSAT" "C1P6A" "C1P6B" "C1P6C" "C1P6D" "C1P6E" "C1P6F" "C1P6G" "C1P2S" "C1P2B_1" "C1P2B_2"

I_KEY	C1P1B	C1P311
00-00-14-41: 1 RO :	42	afc97ea131fd7e2695a98ef34013608f97f34e1d: 16336
00-00-22-72: 1 R0123456789:	31	b9d0f181ffb5bd226d5690dc753aa2d0fbd88174: 18
00-00-26-73: 1 ROSIOARA :	14	dbb8e594cd7045fae7b28e6a595293675765e450: 12
00-00-29-23: 1 R0539322636:	13	da39a3ee5e6b4b0d3255bfe95601890afd80709: 8
00-00-38-70: 1 R0010690458:	11	83a62f7c422b23a8676758af971278862f0ad7b9: 8
00-00-51-93: 1 (Other) :	797216	(Other) :2960549
(Other): :3225902 NA's :	2428581	NA's : 248977

C1P312	C1P321	C1P3	C1P4
0 2918207	1590120 : 38	JUDETUL ARGES : 152487	POPESTI : 6788
- : 1618	14818116: 15	JUDETUL SUCEAVA: 151165	CALINESTI : 6134
99 : 563	999 : 15	JUDETUL DOLJ : 127020	VANATORI : 5979
990 : 348	41228341: 10	JUDETUL PRAHOVA: 124649	ORAS BORSA : 5713
09 : 335	6064801 : 10	JUDETUL BACAU : 122322	MUNICIPIUL ORADEA: 5573
(Other): 55856	(Other) : 31457	(Other) :2545053	(Other) 3184282
NA's : 248981	NA's :3194363	NA's : 3212	NA's : 11439

C1P5A	SIRJUD	SIRCOM	SIRSAT	C1P6A
POIANA : 8923	38 : 152487	106746 : 5713	26573 : 5571	0 1619354
BORSA : 6434	332 : 151165	26564 : 5573	106755 : 5208	Principala: 100882
SLOBOZIA: 6421	163 : 127020	15108 : 4702	76709 : 3847	PRINCIPALA: 75538
POPESTI : 6168	298 : 124649	32394 : 4592	72016 : 3576	principala: 18835
LUNCA : 6096	47 : 122322	13668 : 4420	54984 : 3404	Bisericii : 8695
(Other) :3181360	154 : 117401	(Other):3197696	(Other):3201057	(Other) 1399223
NA's : 10506	(Other):2430864	NA's : 3212	NA's : 3245	NA's : 3381

C1P6B	C1P6C	C1P6D	C1P6E	C1P6F
0 : 778668	0 :3192412	0 :3199352	0 :3201717	0 :3189811
1 : 45913	- : 1974	A : 5384	1 : 4238	- : 1975
2 : 43669	1 : 1194	B : 3278	2 : 3836	2 : 1694
3 : 40263	99 : 957	1 : 2362	3 : 2960	1 : 1670
4 : 39412	2 : 934	- : 2017	- : 2043	4 : 1573
(Other):2274574	(Other): 24999	(Other): 10074	(Other): 7668	(Other): 25740
NA's : 3409	NA's : 3438	NA's : 3441	NA's : 3446	NA's : 3445

C1P6G	C1P2S	C1P2B_1	C1P2B_2
0 :3207968	Min. : 0.000	Min. :-29.80	Min. :-115.26
- : 2028	1st Qu.: 1.000	1st Qu.: 44.90	1st Qu.: 23.66
1 : 796	Median : 1.000	Median : 45.86	Median : 25.13
990 : 698	Mean : 1.014	Mean : 45.60	Mean : 24.91
99 : 691	3rd Qu.: 1.000	3rd Qu.: 46.91	3rd Qu.: 26.57
(Other): 10279	Max. :12.000	Max. : 57.01	Max. : 127.41
NA's : 3448			

APIA shape files

The data comes from extraction from APIA file, named "ipa_2020_changed_centroids.zip". The variables names are: "gid" "farm_id" "sirsup_cod" "bloc_nr" "parcel_nr" "crop_nr" "cat_use" "crop_code" "area_decla" "agro_env" "crop_code_" "area_dec_1" "m_13_anc" "long" "lat"

APIA areas

The file "APIA_2020_suprafete_sept_2021.xlsx", has used for identification of APIA Code for the holding with no APIA code collected, in GAC collected data, but they have ISU code *CIP311* (C01P04_1d) or fiscal code *CIP321* (C01P04_2b). The variables names are: "IDF" "NUME" "CNP_CUI" "TARA" "TIP_F" "ADREXPL" "JUD_DOM" "LOCAL_DOM" "SECTOR" "SAT" "COD_POSTAL" "STRADA" "NR_STRADA" "NR_CLADIRE" "NR_SCARA" "NR_APART" "SIRUTA" "LOCAL" "SUP_UTILIZATA" "SUP_SOL"

The final file generated based on matching the data and corrections, which will be used in geo-reference application as data collected from GAC and corrected based on APIA data, as the coordinates must correspond to the location of the holding and not to the place where the interview was done, has the following structure:

```
interview_id
interview_key
C01P01_1b
GEO_LCT
C01P01_3
C01P01_4
C01P01_5a
C01P01_6a
C01P01_6b
C01P01_6c
C01P01_6d
C01P01_6e
C01P01_6f
C01P01_6g
LEG_FORM
visitLocation_Latitude
visitLocation_Longitude
visitLocation_Accuracy
visitLocation_Altitude
visitLocation_Timestamp
Latitude_APIA
Longitude_APIA
Flag
```



Competence makes a difference!

Project selected under the Administrative Capacity Operational Program, co-financed by
European Union from the European Social Fund