EUROPEAN FORUM FOR GEOSTATISTICS

DISAGGREGATION METHODS FOR GEOREFERENCING INHABITANTS WITH UNKNOWN PLACE OF RESIDENCE : THE CASE STUDY OF POPULATION CENSUS 2011 IN THE CZECH REPUBLIC

Project	ESSnet project GEOSTAT 1B	
Agreement No.	50502.2009.004-2009.860	
WP	WP-1: Geostatistics	
Task	Task 1B.2: Case studies for the “Top down” approach	
Deliverable	Case study	
Date	2013-12-11	

Contributors	Czech Statistical Office	Jaroslav Kraus, Štěpán Moravec, Petr Klauda

ABSTRACT

When processing Population and Housing Census 2011 an overwhelming majority of population data was georeferenced into building points, what enables applying aggregation method for producing a population grid. However, there are about 93 thousands of people (i.e. about 0,9 % of the total census population), who can be linked down only to the level of statistical districts, but not to the exact place of usual residence (e.g. homeless people, people living in buildings without final approval or in emergency buildings or shelters). That means, distribution of these persons into the exact place (with x,y coordinates) or alternatively into grids must be conducted through some disaggregation method.

This case study shows different methods on how to disperse not assigned people either into fictive places of usual residence, or into grids. The methods are applied via ArcGIS software on spatial example of a small town Abertamy in the northern part of Czech Republic very close to Germany. The paper discusses these methods and indicates their advantages and disadvantages. The final solution for the Czech Republic will be adopted before publishing census results in grid format.

Prague Wednesday, December 11, 2013

Authors:

Jaroslav Kraus, CZSO, Na padesátém 81, 100 82, Praha 10, e-mail: jaroslav.kraus@czso.cz

Štěpán Moravec, CZSO, Na padesátém 81, 100 82, Praha 10, e-mail: stepan.moravec@czso.cz

Petr Klauda, CZSO, U Divadla 828, 530 02, Pardubice, e-mail: petr.klauda@czso.cz

TABLE OF CONTENTS

OBJECT OF THE CASE STUDY

STARTING SITUATION

According to the Population and Housing Census 2011 the total number of usually resident population in the Czech Republic amounts to 10 436 560 persons. During the process of census data processing an overwhelming majority of population data was georeferenced to building points due to high quality of field works. The final georeferenced results are stored in the Register of Census Districts and Buildings managed by the Czech Statistical Office. This register contains totally 1 790 122 georeferenced inhabited building points with 10 343 479 persons, which have a known exact place of usual residence.

DESCRIPTION OF THE PROBLEM

Thanks to high coverage of georeferenced data (more than 99 %) an aggregation method for producing population grid can be applied in the Czech Republic. Nevertheless, there are about 93 thousands of people (i.e. about 0,9 % of the total census population), who can be linked down only to the level of statistical districts (irregular territorial units in average bigger than grids), but not to the exact place of usual residence. That means, distribution of this 0,9 % of population into the exact place (with x,y coordinates) or alternatively into grids must be conducted through some disaggregation method. The problem with unknown place of residence in census was caused by missing, incomplete or incorrect address data in census forms and usually affected following groups of inhabitants:

- homeless people
- people living in emergency buildings or shelters
- people living in buildings without final approval

A STUDIED TERRITORY

For testing a case study was chosen a small town Abertamy in the northern part of the Czech Republic very close to Germany. This town had according to the Census 2011 totally 1 213 inhabitants, from whom 46 were not georeferenced. The territory of a municipality has been divided into 6 statistical districts, whereby every statistical district involved some persons without coordinates. The number of known inhabited buildings amounted to 214.

METHODOLOGY

METHOD 1: CREATING NEW RANDOM BUILDING POINTS

AIM

Following previous discussion, there is a necessity to create an appropriate number of random point features - in according to a number of unplaced population in this territory.

METHODOLOGICAL PROCEDURE

The Number of Points parameter can be specified as a number or as a numeric field in the constraining feature class containing values for how many random points to place within each feature. The field option is only valid for polygon or line constraining features. If the number of points is supplied as a number, each feature in the constraining feature class will have that number of random points generated inside or along it.

If you are using a constraining feature class that has more than one feature, and you wish to specify the total number of random points to be generated (as opposed to the number of random points to be placed inside each feature), first use the Dissolve tool so that the constraining feature class only contains a single feature, then use that dissolved feature class as the constraining feature class.

This part of solution is based on ArcGIS 10.1 software.

During this step an appropriate number of randomly defined building points (in the Czech Republic called IDOBs) is created, which corresponds with a number of displaced persons. It is also necessary to define an interval of persons, who can be assigned to one random building point. As a result we gain a random number of population, which is assigned to all created building points within specific statistical district. For Abertamy the solution is following:

Number of rIDOBs: 47

Number of persons by rIDOBs: <1;4>

dim max, min

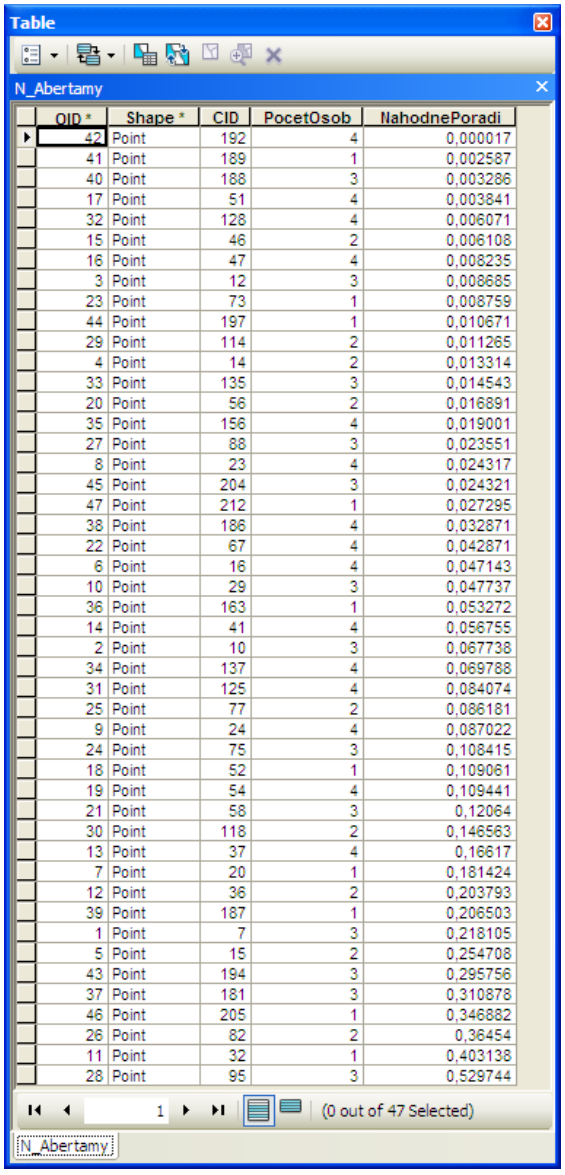
max = 4

min = 1

x = (Int((max-min+1)*Rnd+min))

__esri_field_calculator_splitter__

Number of persons = X

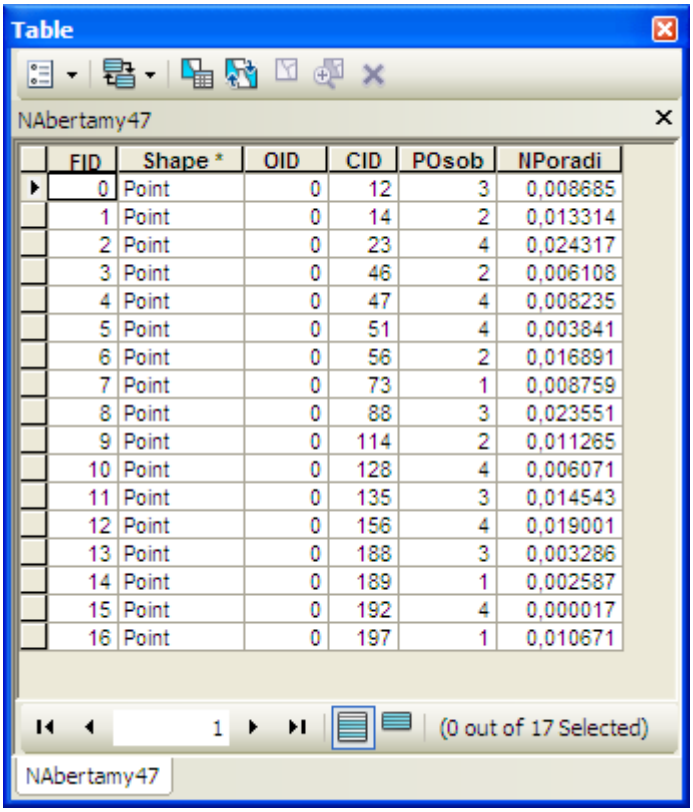


OID *	Shape *	CID	PocetOsob	NahodnePoradi
42	Point	192	4	0,000017
41	Point	189	1	0,002587
40	Point	188	3	0,003286
17	Point	51	4	0,003841
32	Point	128	4	0,006071
15	Point	46	2	0,006108
16	Point	47	4	0,008235
3	Point	12	3	0,008685
23	Point	73	1	0,008759
44	Point	197	1	0,010671
29	Point	114	2	0,011265
4	Point	14	2	0,013314
33	Point	135	3	0,014543
20	Point	56	2	0,016891
35	Point	156	4	0,019001
27	Point	88	3	0,023551
8	Point	23	4	0,024317
45	Point	204	3	0,024321
47	Point	212	1	0,027295
38	Point	186	4	0,032871
22	Point	67	4	0,042871
6	Point	16	4	0,047143
10	Point	29	3	0,047737
36	Point	163	1	0,053272
14	Point	41	4	0,056755
2	Point	10	3	0,067738
34	Point	137	4	0,069788
31	Point	125	4	0,084074
25	Point	77	2	0,086181
9	Point	24	4	0,087022
24	Point	75	3	0,108415
18	Point	52	1	0,109061
19	Point	54	4	0,109441
21	Point	58	3	0,12064
30	Point	118	2	0,146563
13	Point	37	4	0,16617
7	Point	20	1	0,181424
12	Point	36	2	0,203793
39	Point	187	1	0,206503
1	Point	7	3	0,218105
5	Point	15	2	0,254708
43	Point	194	3	0,295756
37	Point	181	3	0,310878
46	Point	205	1	0,346882
26	Point	82	2	0,36454
11	Point	32	1	0,403138
28	Point	95	3	0,529744

The next step is based on recalculation of usually living population with unknown place of residence for randomly defined building points in order to gain appropriate number of displaced persons for individual statistical districts. In fact, there is selected such a number and composition of random building points, which give a total number of persons to be distributed.

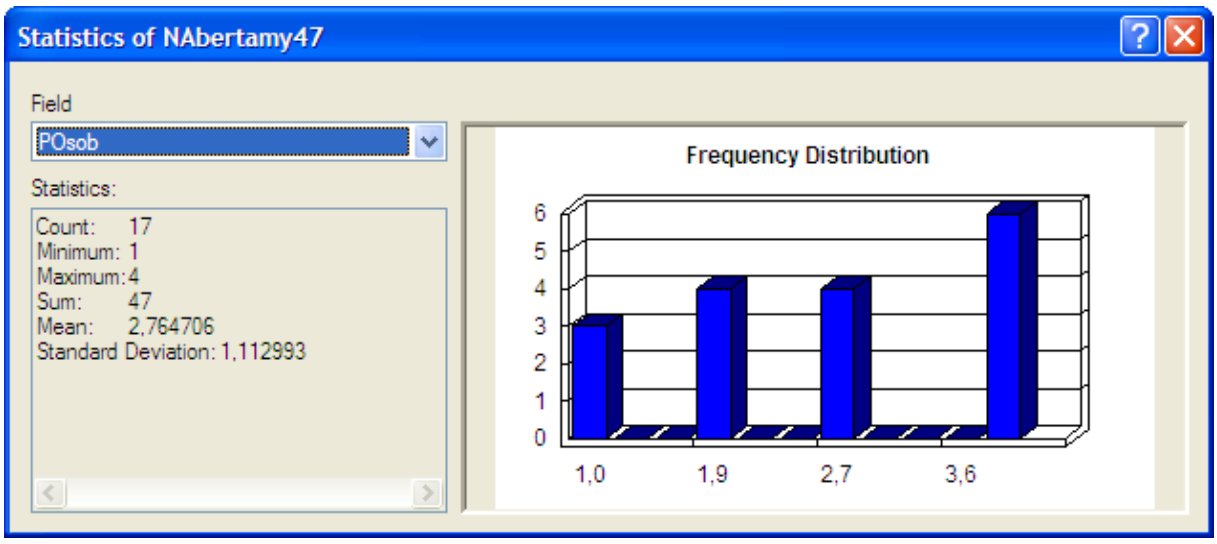
Random number <0;1> of selected persons:

```
dim max, min
max = 1
min = 0
x = ((max-min+1)*Rnd+min)
__esri_field_calculator_splitter__
RandomOrdering = x
```



FID	Shape *	OID	CID	POsob	NPoradi
0	Point	0	12	3	0,008685
1	Point	0	14	2	0,013314
2	Point	0	23	4	0,024317
3	Point	0	46	2	0,006108
4	Point	0	47	4	0,008235
5	Point	0	51	4	0,003841
6	Point	0	56	2	0,016891
7	Point	0	73	1	0,008759
8	Point	0	88	3	0,023551
9	Point	0	114	2	0,011265
10	Point	0	128	4	0,006071
11	Point	0	135	3	0,014543
12	Point	0	156	4	0,019001
13	Point	0	188	3	0,003286
14	Point	0	189	1	0,002587
15	Point	0	192	4	0,000017
16	Point	0	197	1	0,010671

Statistics of results:



METHODOLOGICAL CONSTRAINTS

The Constraining Extent parameter can be entered as a set of minimum and maximum x- and y-coordinates or as equal to the extent of a feature layer or feature class.

If both a constraining feature class and constraining extent are specified, the constraining feature class value will be used and the constraining extent value will be ignored.

When unable to place anymore random points within a constraining area without breaking the minimum allowed distance specified, the number of random points in the constraining area will be reduced to the maximum possible under the minimum allowed distance.

The Minimum Allowed Distance parameter can be specified as a linear unit or a field from the constraining features containing numeric values. This value will determine the minimum allowed distance between random points within each input feature. The field option is only valid for polygon or line constraining features. Random points may be within the minimum allowed distance if they were generated inside or along different constraining feature parts.

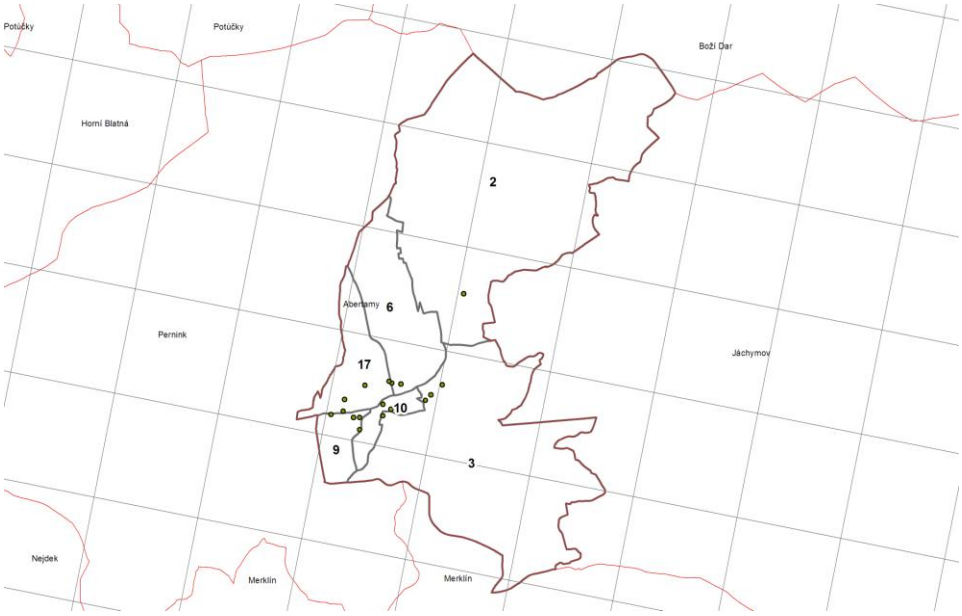
Using point features as the constraining feature class creates a random subset of the constraining point features. No new point locations are generated.

Noninteger (whole) positive values for the Number of Points and Minimum Allowed Distance parameters will be rounded to the nearest whole number. Nonnumeric and negative values are set to 0.

This part of solution is based on ArcGIS 10 software

RESULTS :

Number of placed persons (47) for randomly defined building points according to statistical districts – a case of Abertamy



METHOD 2: CREATING OF POPULATION CENTERS OF GRAVITY

AIM

The aim is very similar to the previous method: to place persons with unknown place of residence on spatially weighted gravity centre. The weights are persons with known place of residence (IDOB) by statistical districts.

METHODOLOGICAL PROCEDURE

The mean center is a point constructed from the average x and y values for the input feature centroids.

Use projected data with this tool to accurately measure distances.

The x and y values for the mean center point features are attributes in the Output Feature Class. The values are stored in the fields XCOORD and YCOORD.

The Case Field is used to group features for separate mean center computations. When a Case Field is specified, the input features are first grouped according to case field values, and then a mean center is calculated for each group. The case field can be of integer, date, or string type. Records with NULL values for the Case Field will be excluded from analysis.

The Dimension Field is any numeric field in the input feature class. The Mean Center tool will compute the average for all values in that field and include the result in the output feature class.

For line and polygon features, feature centroids are used in distance computations. For multipoints, polylines, or polygons with multiple parts, the centroid is computed using the weighted mean center of all feature parts. The weighting for point features is 1, for line features is length, and for polygon features is area.

Map layers can be used to define the Input Feature Class. When using a layer with a selection, only the selected features are included in the analysis.

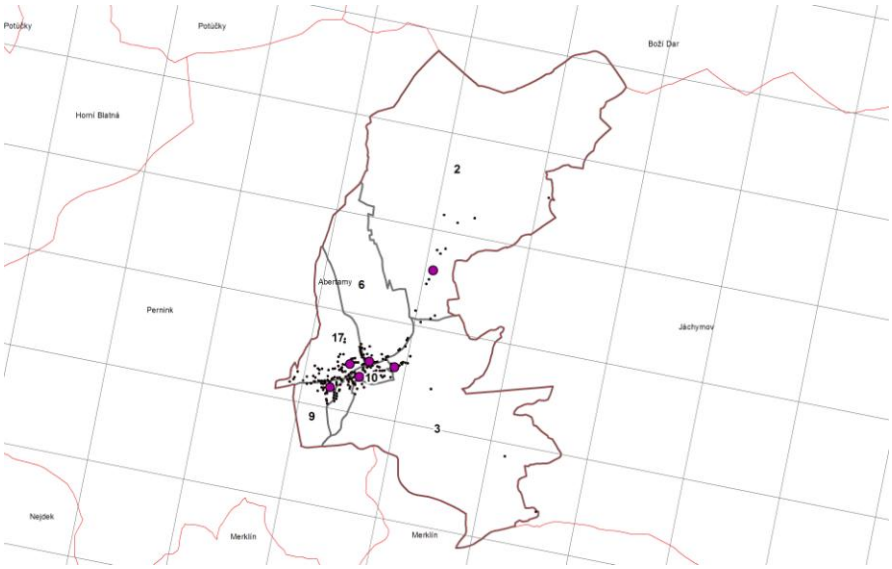
This part of solution is based on ArcGIS 10 software.

METHODOLOGICAL CONSTRAINTS

Geoprocessing considerations for shapefile output – see ArcGIS 10 software

RESULTS

Spatially weighted gravity centre - Abertamy



METHOD 3: CALCULATION OF POPULATION WEIGHTS OF GRIDS

AIM

Third method differs from the previous 2 approaches in the 2 aspects. Firstly, it aims to distribute not georeferenced population just into grids, i.e. not into particular building points (with x,y coordinates). Secondly, it aims to respect a known spatial distribution of population, which is of course based on georeferenced population only.

METHODOLOGICAL PROCEDURE

In the first step we calculate a population weight of each inhabited grid segment within an affected statistical district, which contains some persons with unknown place of residence. A grid segment is defined as a whole grid cell or just its part bordered by a border of statistical district. The formula is:

$$Wgi = \frac{Pgi}{Psj}, where$$

$Wgi \ldots Population\ weight\ of\ grid\ segment\ i;$

$Pgi \ldots Georeferenced\ population\ number\ of\ grid\ segment\ i;$

$Psj \ldots Total\ georeferenced\ population\ of\ statistical\ district\ j, where\ a\ grid\ segment\ i\ belongs\ to$

In the second step we calculate a population number distributed to each inhabited grid segment within affected statistical district according to a formula:

$$Dgi = Wgi \times Nsj, where$$

$Dgi \ldots Population\ number\ distributed\ to\ grid\ segment\ i;$

$Wgi \ldots Population\ weight\ of\ grid\ segment\ i;$

$Nsj \ldots Total\ number\ of\ not\ georeferenced\ persons\ within\ statistical\ district\ j$

In the previous step we multiply a decimal number (*Wgi*) with an integer number (*Nsj*), but the target value is integer. Hence, in the next step is necessary to round the population number distributed into each inhabited grid segment (*Dg*) to an integer value. The sum of rounded population numbers distributed into all grid segments, which belong to one statistical district, represents a total population number distributed within this statistical district. This figure should be the same as an initial number of not georeferenced persons within affected statistical district.

Finally, the number of distributed not georeferenced persons is added to the initial number of georeferenced persons for each grid segment:

$$Tgi = Pgi + Dgi, where$$

$Tgi \ldots Total\ population\ number\ of\ grid\ segment\ i$

METHODOLOGICAL CONSTRAINTS

By testing this method on the whole territory of the Czech Republic it came out different types of irregularities and deviations. Most frequently appeared a problem with rounding of decimal numbers of population distributed to inhabited grid segments to integer values, which resulted in an increase or a decrease of the distributed population number within statistical district in comparison to the initial number of not georeferenced persons. This problem affected almost 12 % of all statistical districts with not georeferenced population. The difference for individual statistical districts was between -2 up to +4 persons.

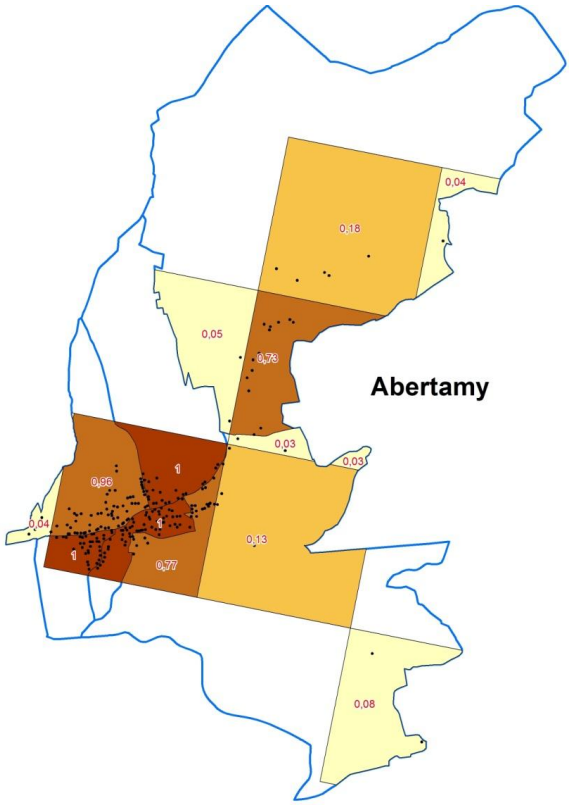
One specific subgroup under the previous deviation (about 6 % of cases) was related to statistical districts without any inhabited buildings. In this case the whole population of such a statistical district was not georeferenced, i.e. the place of residence of all persons and their territorial distribution was unknown.

Last, but not least problem represented grid segments of one statistical district, which had the same population weight. A typical example is a statistical district containing only 1 person without x,y coordinates on one side, but consisted of just 2 inhabited grid segments with the same population number on the other side. The same population number of both grid segments results in the same population weight (0,5), which after rounding amounts to 1. Nevertheless, 2 grid segments with such a number of distributed population are in contradiction with only one not georeferenced person.

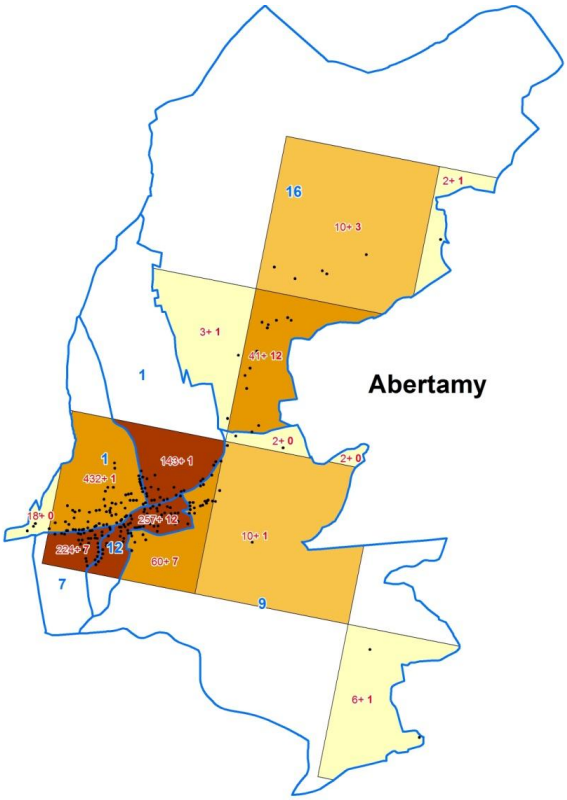
All aforementioned problems require definition of additional assumptions and consequent manual corrections, what is quite demanding.

RESULTS

1) Layer of population grids with relative population weight – Abertamy



2) Layer of population grids with number of additionally distributed persons – Abertamy



DISCUSSION OF RESULTS

For presented 3 methods were ascertained some obvious differences, especially noticable by comparison of method 3 with methods 1 and 2, which both had due to similar methodological solution relatively common features. The most important difference was based on the fact, that within first two methods were distributed persons with unknown place of residence into individual building points (although fictive), whereas the third method enabled distribution of these persons only to the level of grids. Another crucial difference lay in a character of distribution of not georeferenced population, which was carried out according to a spatial distribution of inhabited buildings within methods 1 and 2, but according to a spatial distribution of usually living population within method 3.

As a common feature for all methods could be seen a fact, that there was needed relatively enough handwork and manual corrections in order to finalize the methods, thus, implementation of some automatization of processes seems to be indispensable.

Elaboration and testing of described methods provided the Czech Statistical Office with a different ways, how to include persons with unknown place of residence into a total set of spatially localized population. From the analyzed methods will be chosen for a final solution such a combination of approaches, which will enable maximal automatization of the whole process, which will respect the census results in terms of territorial structure (e.g. according to statistical districts) and which will lead to a trustworthy presentation of census results in grids of different sizes.