N° 108 ÉCONOMIE ET STATISTIQUES
October 2019 WORKING PAPERS DU STATEC

New approaches in georeferenced statistical data and related confidentiality issues

Author : Krisztina Dékány,
SOGETI Luxembourg SA
Krisztina.Dekany@sogeti.lu

Abstract

According to the recent EU Regulations, the National Statistical Institutes of the EU Member States have the obligation to provide some demographic variables of the 2021 EU Census at a 1 km² grid level. The publication of such geographically detailed information requires an appropriate data protection methodology in order to reduce the risk of disclosing information on individual statistical units.

Therefore, STATEC has established a new methodology which allows the provision of statistics at a 1 km² grid level while complying with the rules of statistical data confidentiality.

The aim of this paper is to outline some statistical disclosure control (SDC) methods using the dataset of the current *Registre des Bâtiments et des Logements* of the STATEC as an example.

2

Introduction

In 2018, the EU Member States adopted a proposal from the European Commission on a direct statistical action for the dissemination of selected topics of the upcoming European 2021 population and housing census[1]. According to this decision, STATEC has to supply selected census variables by a 1 km² grid. The spatial grid to be used is based on the INSPIRE (Infrastructure for Spatial Information in the European Community) Directive of 2007[2].

As this is a new situation for Luxembourg, STATEC has now established a methodology that allows the generation of highly disaggregated datasets while respecting the general rules of data confidentiality. This paper demonstrates this method by using variables from the Register of Buildings and Dwellings, reflecting the situation at the end of 2016 (Table 1), with the aim of dissemination by 1 km² grid.

The *Registre des Bâtiments et des Logements* contains the following basic characteristics:

Table 1: The main characteristics of the Register of Buildings and Dwellings (2016)

Number of buildings	144 950
Number of dwellings	239 308
Types of dwellings:	
Single-family houses	120 206
Apartments	84 878
Semi-residential houses	34 224

After applying an algorithm to assign each of the buildings and dwellings to the proper grid cell, STATEC had to examine confidentiality protection methods to secure the dataset. Indeed, territorial variables' datasets always require special attention regarding protection against disclosure. The general principle says that the smaller the territorial unit under examination and the smaller the number of cases, the more important the need for data protection.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1799&rid=1

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007L0002&from=EN

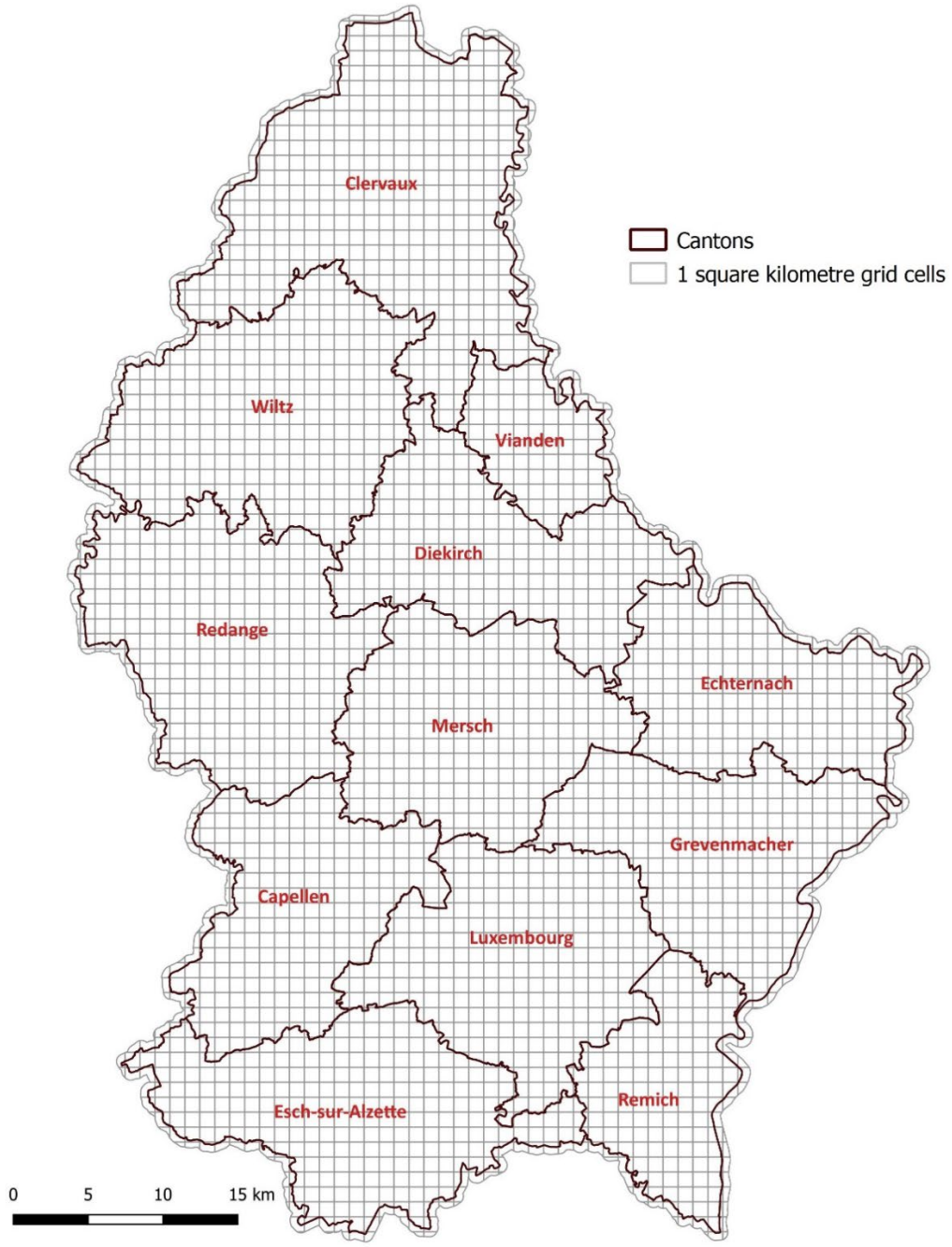
1. The European Reference Grid

According to the Commission Implementing Regulation 2018/1799 "on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km² grid", the objective is to disseminate one dataset per Member State, containing selected topics of the 2021 population and housing census geocoded to a 1 km² grid. Specific reference is made to the INSPIRE Directive of 2007. In particular, in order to achieve comparable harmonized outputs across the European Union, an EU-wide constant area grid consisting of 1 km² cells needs to be used (Figure 1). A grid structure is better for spatial statistical analysis, as grid cells of the same size make comparisons easier and remain fixed over time. Administrative entities can change and hence hamper comparability. The Grand-Duchy of Luxembourg counts 2937 grid cells of 1 km². Among these, 364 grid cells overlap with borders of adjacent countries. The Commission Regulation also provides for a single virtual grid cell per country to account for persons that cannot be geo-localised otherwise. This is for instance the case for homeless persons.

The 1 km² reference grid as referenced in the INSPIRE Directive specifies the following:

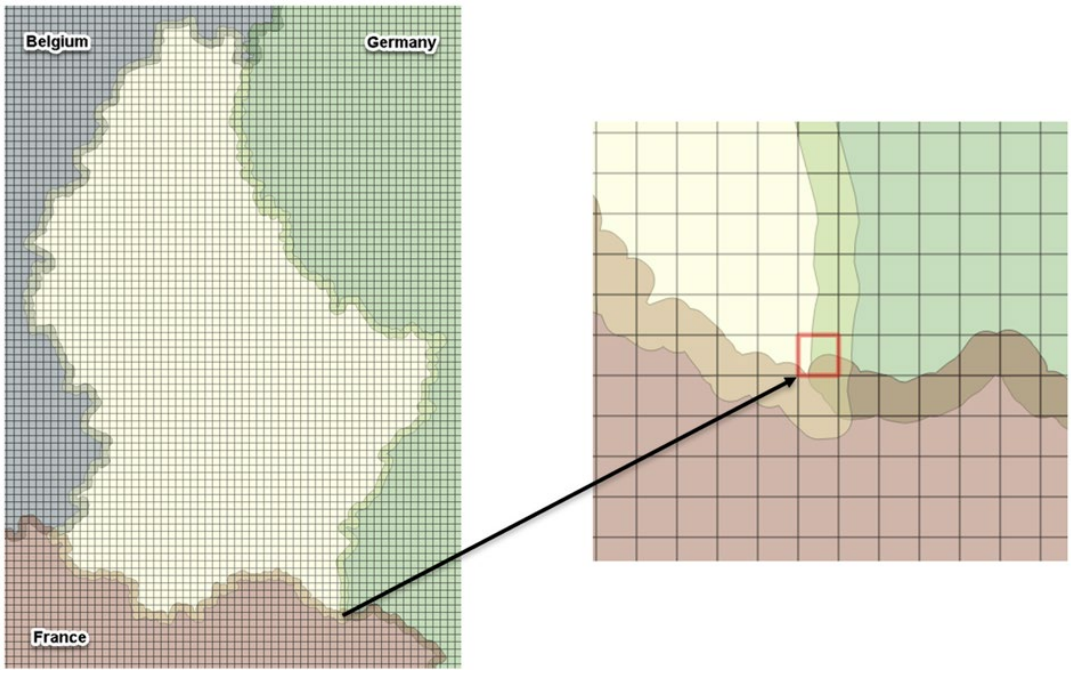
- The statistical 1 km² reference grid for pan-European usage is the Equal Area Grid 'Grid_ETRS89-LAEA1000'. The spatial extent of the reference grid in the coordinate system specified for this grid is limited to easting values between 900 000 and 7 400 000 metres and northing values between 900 000 and 5 500 000 metres.
- Each individual grid cell of the 1 km² reference grid is identified by a unique grid cell code, which is composed of the country code of the transmitting Member State followed by the character '_', is prepended to the cell code of each grid cell transmitted by that Member State. After this, the following characters are needed: 'CRS3035RES1000mN'. This is followed by the northing value in metres of the grid point in the lower-left corner of the grid cell, followed by the character 'E', followed by the easting value in metres of the grid point in the lower-left corner of the grid cell.

Figure 1: Cantons of the Grand-Duchy of Luxembourg with the grid superposed



Regarding the 364 overlapping grid cells along the borders (Figure 2), each country must supply data with the variables for these grid cells as well. EUROSTAT then aggregates the data from the various countries and the frequency counts from each individual country are no longer distinguishable.

Figure 2: Grid cells with overlapping border



2. The theory of confidentiality

Statistical disclosure control (SDC) denotes the process that seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities (Hundepool, A. et al., 2012, p. 1).

The three most common practices limiting disclosure in microdata are:

- eliminating information that directly identifies individuals;
- suppressing data that may indirectly identify individuals;
- introducing uncertainty into the reported data.

The most common disclosure risk scenarios are:

- **identity disclosure**, when grid data with small counts means that the respective individual unit is at risk of being identified from the table;
- **attribute disclosure** can occur if an attribute of one or more individual(s)/household(s) can be learnt from the grid data, for example a small group of persons with a specific age/sex combination in a small municipality where the data exhibit that one/some of them fall into a category of the place-of-birth variable considered sensitive;
- **disclosure by differencing** might happen if someone takes the difference of two tables and the resulting table is disclosive, for example different geographical variables, such as grids and the NUTS classification, potentially increase the risk of disclosure by differencing.

According to the relevant EU Directive, Member States must replace a numerical cell value by the flag 'confidential' if the numerical cell value must not be disclosed for reasons of statistical confidentiality. Additionally, there are some methods which can help to secure the dataset in other ways.

In order to choose the appropriate process, the identifying variables need to be selected. For the purpose of the SDC process, the variables can be classified:

- **Identifying variables**: these contain information that can lead to the identification of respondents and can be further categorised as:
  - **Direct identifiers**: reveal directly and unambiguously the identity of the respondent. Examples are names, passport numbers, social identity numbers and addresses. Direct identifiers should be removed from the dataset prior to release. Removal of direct identifiers is a straightforward process and always the first step in producing a safe microdata set for release.
  - **Quasi-identifiers** (or key variables): contain information that, when combined with other quasi-identifiers in the dataset, can lead to the re-identification of respondents. This is especially the case when they can be used to match the information with other external information or data. Examples of quasi-identifiers are race, birth date, sex and postal codes, which might be easily combined or linked to publicly available external information and make identification possible.
- **Non-identifying variables** are variables that cannot be used for the re-identification of respondents. This could be because these variables are not contained in any other data files or other external sources and are not observable to an intruder. According to the OECD glossary, an intruder is a "data user who attempts to link a respondent to a microdata record or make attributions about particular population units from aggregate data."

3. Categorisation of SDC methods

The key objective of any of the confidentiality methods should be to ensure minimum information loss and maximum data utility. Utility refers to the quality of the output after SDC application and can be assessed by analysing the impact of SDC methods on statistical analysis. The confidentiality methods can be grouped in several ways.

3.1. Perturbative and non-perturbative methods

Perturbative methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. It therefore deliberately changes the data slightly. Conversely, **non-perturbative methods** reduce the amount of information released by means of suppression or aggregation of data, but not by alteration.

3.2. Pre-tabular and post-tabular methods

Tabular data is aggregate information on entities presented in tables. According to this definition, the **pre-tabular method** is applied to the microdata, and grid data are generated from the protected microdata. On the contrary, the **post-tabular method** is applied directly to the grid data and they must be applied again to every new table.

Pre-tabular methods are generally more flexible, with parameters that can be varied to achieve a balance between disclosure risk and utility. However, they can cause statistical damage to the resulting tables that is difficult for the users of these tables to measure. Hence, the parameters of these methods are often set in a way to minimise this damaging effect. Another advantage is that some pre-tabular methods can be unbiased, whereas most post-tabular methods involve suppressing cells and then introducing bias in the estimation of parameters or making some parameters not estimable.

The fact that pre-tabular SDC methods only need to be applied once is an important advantage compared to post-tabular methods, when the aim is to produce tables for multiple geographies. To take spatial features into account, pre-tabular methods seem to be more appropriate because they allow the use of geographical information to directly target the riskiest records for the perturbation.

4. The most common SDC methods

There is a large variety of methods to protect confidential information (Table 2). Many of them have already been applied by different countries in the past.

Table 2: SDC methods summary with important characteristics
(based on Antal, L. - Enderle, T. – Giessing, S., 2017)

	Non-perturbative methods		**Perturbative methods**	
	Cell suppression	**Global recoding**	**Record swapping**	**Random noise**
Tabular method	pre-tabular	pre-tabular	pre-tabular	post-tabular
Information loss	often high	often high	can be set to small	can be set to small
Main process of the method	replacing values with missing value (NA or x)	combine categories or constructs intervals	exchanges variables between records	changes counts in frequency cells
Consistency	hard to carry out consistently when higher dimensional and linked tables are considered	can be consistent and additive	tables generated from the same protected microdata set are consistent and additive	either consistency or additivity can be preserved
Additivity	most parts of the table remain additive, but some additivity is lost			

5. Measuring information loss in different SDC methods

A practical consideration in releasing a protected table would be to balance data confidentiality and data quality (Sukasih, A. – Jang, D. – Czajka, J., 2012, p. 6). The SDC area always faces such a trade-off, where overprotection could lead to greater loss of information. On the other hand, the use of less suppression to avoid too much loss of information may widen the room for disclosure risks. In evaluating the quality of the published table, the table producer can approach this task from the point of view of estimation; i.e. by evaluating aggregate suppressed data (magnitude and/or the frequencies) relative to the population or original (unprotected) table. In addition, the evaluation can be carried out by comparing the loss of information resulting from the use of different SDC methods.

Information loss measures can help to select between variants (e.g. different parameter settings) of the same protection method, or also to choose between different protection methods. For perturbative methods, not only the usual descriptive statistics (max, median, mean) are useful, but it is preferred to have more punctual and sophisticated measures, such as:

- absolute differences (AD);
- relative differences (RAD) between original and altered counts in a table (or grid);
- (squared) differences of the square roots between original and altered counts.

5.1. Measuring information loss in microdata

The protected microdata set should be analytically valid. "Analytically valid" means that it approximately preserves the following characteristics with respect to the original categorical data (Hundepool, A. et al., 2012, pp. 100-101):

- means and covariances on a small set of sub-domains (subset of records and/or variables);
- marginal values for a few tabulations of the data;
- at least one distributional characteristic;
- compare raw records in the original and the protected data sets (pairing);
- compare some statistics computed on the original and the protected data sets.

5.2. Measuring information loss in tabular data

In order to choose the best post-tabular method for the census, the following information loss measures can help in a comparison (Hundepool, A. et al., 2012, pp. 200-201):

- binomial hypothesis test in the case of random rounding (realisation of the random stochastic perturbation scheme follows the expected probabilities);
- non-parametric signed rank test: the location of the empirical distribution after the application of the SDC method has changed;
- distance metrics between the original and perturbed cells;
- examination of the original and the perturbed tables' variance by the average cell size of the rows, columns and the entire table;
- analysis of variance (ANOVA – 'between' variance);
- test of independence between categorical variables that span the table (Pearson Chi-Squared Statistics and Cramer's V statistic);
- Spearman's Rank Correlation: measures the direction and strength of the relationship between two variables (based on ranking of the cell counts).

6. Recommendations from Eurostat

In 2017, the Eurostat project *Harmonized Protection of Census Data in the ESS* aimed at harmonising disclosure control techniques concerning census' in European countries (Table 3), for hypercubes on the one hand and for grid data on the other hand. For this project, two complementary methods have been chosen, as they seemed to offer a good compromise between confidentiality and utility loss.

The suggested method is a combination of the targeted record swapping and the random noise (cell key) method. Some countries indicated they intended to use one method only: either data swapping or the cell key method or some other method. Countries that do not use a combination of pre- and post-tabular SDC methods are advised to use the cell key method.

If many EU Member States use the same method (though perhaps with different parameters) this will help to prepare European-level data in a more straightforward way. In order to maintain consistency between European and national data releases, Member States are encouraged to apply the same SDC method to all kinds of data releases.

Given the different statistical confidentiality rules in European countries, it is advisable to recommend not just a single method. Recommending a selection of two methods which may or may not be used in combination and can be controlled with parameters and options offers flexibility to the countries. Using both or only one of the two recommended methods, even with different parameter values, the output by grid will be similar enough to allow for comparison of the statistics between countries.

Table 3: Methods according to the EU
(based on Antal, L. - Enderle, T. – Giessing, S., 2017)

Advantages and **disadvantages**

Criterion	Cell suppression / Global recoding	Random noise / Cell key	Targeted record swapping
Risks of identification	no common understanding which small counts have to be protected	all frequencies possibly changed → protection provided	
Risks of attribute disclosure	no common understanding to which extent protection against risks of attribute disclosure ("all 40-50 year old males married ") is needed	all "0" frequencies may result from perturbation → protection provided	
Differencing risks	inconsistencies in suppressions across grids may cause disclosure risks	perturbation protects also against differencing disclosure risks	
Information loss	information loss due to suppressed cells might be high	information loss low, controlled to a large extent by parameters selected by the NSI	information loss depends on parameters selected by NSI cumulative effects possible

6.1. The record swapping method

The basic idea consists of transforming the original microdata-table by exchanging values of confidential variables among individual records. Some pairs of records are selected in the microdata set. The paired individuals/households match on some variables in order to maintain the analytical properties and to minimize the bias of the perturbed microdata set as much as possible. Record swapping exchanges some of the non-equal variable-values between paired individuals/households (see example in Figure 3). Since this exchange introduces uncertainty to the microdata, an intruder's assumption about a certain individual/household might not be correct.

Figure 3: Record swapping example (Moore, R. A., pp. 3)

Example. A microdata file contains the age and income for 6 respondents. In order to protect the anonymity of the respondents, income values are randomly swapped among the records. Incomes on the first and sixth records, those on the second and third, and those on the fourth and fifth are pairwise swapped.

Original Responses

#	Age	Income
1	21	20,000
2	24	30,000
3	35	30,000
4	36	25,000
5	45	55,000
6	50	15,000

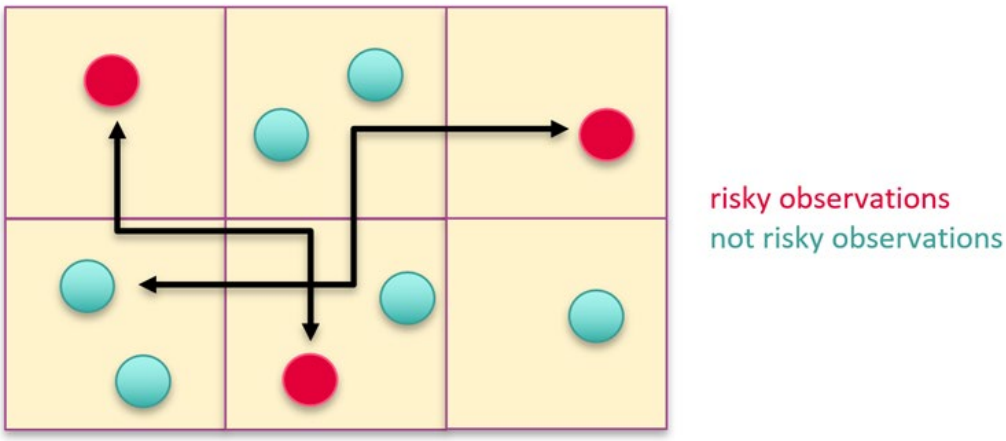
Responses After Swap #1

#	Age	Income
1	21	15,000
2	24	30,000
3	35	30,000
4	36	55,000
5	45	25,000
6	50	20,000

Records 2 and 3 appear unchanged. Record 2 was swapped with a record having the same income. For these respondents, the swap has provided no masking. The probability that a swap has masked a particular record is inversely proportional to the frequency of its value appearing in the file. For large data files, this is acceptable. An income which appears frequently in a microdata file does not as easily identify the respondent as one which appears very rarely.

In general, record swapping can be random or targeted. In the case of random record swapping, the individuals/households to be swapped are selected with equal probability (Figure 4), while in the case of targeted record swapping, records of high disclosure risk are determined and a pairing for each of these records are selected. This may be a good choice as it affects fewer records. Although in the case of targeted swapping, all records (including the non-risky ones) still have a chance of being swapped if there is no acceptable pair for the swapping at risk, so there is some uncertainty in all small counts. Please note that record swapping is applied to the microdata. Therefore, at least one of the variables of each hypercube needs to be swapped in order to obtain a perturbed hypercube that is actually different from the original one.

Figure 4: example of record swapping between grid cells



According to the example in Shlomo, N. - Tudor, C. – Groom, P. (2010), for the selected household to be swapped, first the level of geographical disclosure risk needs to be checked and a paired household at the appropriate geography, having the same control variables, should be chosen. For example, if the level of geographical disclosure risk is flagged at grid cell, then the household must be swapped with a similar household having the same control variables in a different grid cell but within the locality. The advantage of 'localised' data swapping is that it minimises the distance between household pairs. At higher aggregations of geography, this results in less distortion.
In a census context, geography variables are often swapped between households for the following reasons (Shlomo, N., 2007):

- given household characteristics, other census variables are likely to be independent of geography, therefore, it is assumed that less bias will occur;
- at a higher geographical level and within control strata, the marginal distribution is preserved;
- the level of protection increases by swapping variables that are highly "matchable", such as geography.

The data swapping procedure has the following advantages and disadvantages (Table 4):

Table 4: Advantages and disadvantages of record swapping for census outputs (based on Shlomo, N., 2007)

Advantages	Disadvantages
Consistent tables	High proportion of high-risk (unique) records left unperturbed
Preserves marginal distributions at higher aggregated levels	Errors (bias) in data, joint distributions distorted
Some protection against disclosure by differencing nested tables	Effects of perturbation hidden and cannot be accounted for in the analysis of the data
Fewer edit failures when swapping geographies	Method not transparent to users (perception of disclosure risk)

According to Shlomo, N. (2007) record swapping for census tables results in a high possibility that small cells in tables are true values and can be identified. Targeted record swapping lowers the disclosure risk but there is more distortion to distributions with respect to distance metrics. Higher swapping rates raise the level of protection but also cause severe distortion to the data. This is the only swapping method that doesn't provide confidentiality in cells with small values.

6.2. The cell key method

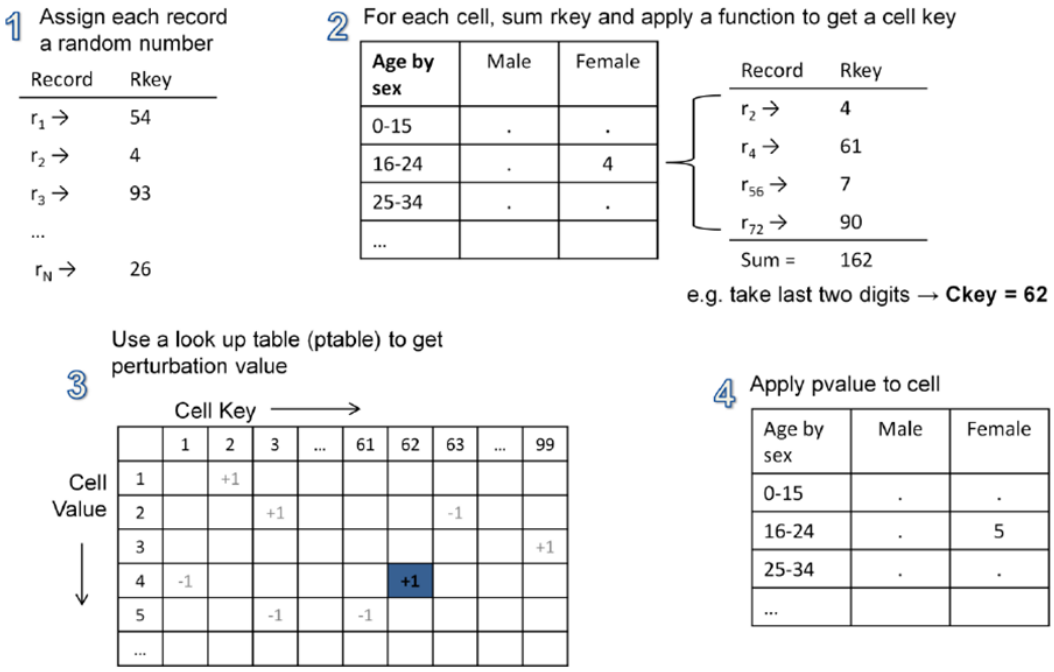
This is a post-tabular method, so it can be used only for frequency tables. The basic idea is to add random noise to the original counts. An essential part of the cell key method is based on an algorithm which applies a pre-defined level of perturbation to cells in each table. The same perturbation is applied to every instance of that cell independently. Therefore, the perturbed grid data will generally not add up exactly.

An implementation of an additive random noise as outlined in the following may involve three steps (Antal, L. – Enderle, T. – Giessing, S., 2017, pp. 8-9):

I. Cell key module (see example in Figure 5): should be drawn from a discrete uniform distribution, defined on some random integer values (for example integers between 1 and 100). The process that defines the cell keys has to be consistent, i.e. it must guarantee that the same cell always gets the same key in any grid cell or tabulation. For each cell, its cell key and its frequency are used to determine the noise applied to the cell. This step is actually

deterministic and can be implemented in such a way that the distributions of the noise match almost exactly the pre-defined distributions to be specified as parameters of the method.

Figure 5: Example of the cell key method
(source: Spicer, K. – Dove, I., pp. 9)



The randomness of the process lies entirely in the part that leads to the cell keys:

- Assign a random number to each record in the microdata (so called "record keys"). Record keys should be evenly distributed and defined on some given interval, for instance between 1 and 100.
- When computing the hypercube data, i.e. counting the number of records having the particular variable combinations of the hypercube cell, do the "same" with the record keys, i.e. take the sum of the record keys for those microdata records having the particular variable combinations of the hypercube cell. Take Modulo 100 (remainder when divided by 100) of the sum of these record keys. The result, referred to as "cell key" obviously lies then also between 1 and 100. Less obvious, but established mathematically, the cell keys will then also be evenly distributed on this interval.

II. Module to determine noise based on cell key and noise distribution parameter matrix: the performance of a random noise method can easily be controlled in a flexible way via parameter settings defining the probability distributions. Even a random rounding approach can be considered as random noise with specific noise distributions.

III. Module to restore additivity: the random noise added to a specific cell. However, according to this concept the perturbation is applied to each cell independently. Therefore, the perturbed grid data will generally not add up exactly. However, Eurostat accepts this non-additivity attribution if this method is used.

The advantages and disadvantages of this method can be summarised as follows (Table 5):

Table 5: SWOT analysis of the cell key for census outputs

Strengths: – adding noise methods are much easier to write, modify, run, and understand – does not change the mean of the variable for large datasets	**Weaknesses:** – the perturbed grid data will generally not add up exactly
Opportunities: – sensitive cells would in general contain a lot of noise – in contrast: non-sensitive cells would end up with little noise	**Threats:** – may introduce more variance – it could change some non-zero cells into zero cells (so-called false zero cells)

7. The generated method at STATEC

Regarding the obligation to disseminate variables by a 1 km² grid, STATEC tested the cell key method on data from the Register of Buildings and Dwellings, as outlined in the following sections. The description includes examples of the *type of dwellings* variable.

7.1. Measuring the disclosure risk in the original dataset

After creating the input table for the cell key method and before any perturbation is performed, the risk measurement had to be done using a function in the R-package[3]. According to this analysis, the global risk was 1.46 %, which represents the mean of the individual risks. It is likely that for the 2021 population census, the global risk will be higher, as the number of demographic variables will increase.

Figure 6 shows the k-anonymity measures, which are based on the principle that, in a safe dataset, the number of individuals sharing the same combination of values of categorical quasi-identifiers should be higher than a specified threshold (k) (Benschop, T. – Machingauta, C. – Welch, M., 2018, pp. 28-29). An individual violates k-anonymity if the sample frequency count (fk) for the key k is smaller than the specified threshold k. For example, if an individual has the same combination of quasi-identifiers as two other individuals in the sample, these individuals satisfy 3-anonymity but violate 4-anonymity.

[3] R is a programming language and free software environment for statistical computing.

Figure 6: K-anonymity of the original dataset of Register of Buildings and Dwellings

k-anonymity	Original data
2-anonymity	375 (0.161%)
3-anonymity	1027 (0.442%)
5-anonymity	2208 (0.950%)

Regarding the individual risks, Figure 7 shows their descriptive statistics. There are 3 386 records with the maximum risk of 1, which means that the combination is unique in the dataset.

Figure 7: Individual risk of the original dataset and the frequencies (fk)

```
      risk                fk
Min.   :0.0002953   Min.   :   1
1st Qu.:0.0016807   1st Qu.: 112
Median :0.0036101   Median : 277
Mean   :0.0146282   Mean   : 506
3rd Qu.:0.0089286   3rd Qu.: 595
Max.   :1.0000000   Max.   :3386
```

7.2. Setting the parameters for the perturbation

The minimum set of parameters that had to be specified were the following:

- ***D***: the maximum noise/perturbation (in other words, the maximum change of the original frequency);
- ***V***: the noise or perturbation variance (the expectation of the squared deviation of a random variable from its mean).

And from the remaining parameters:

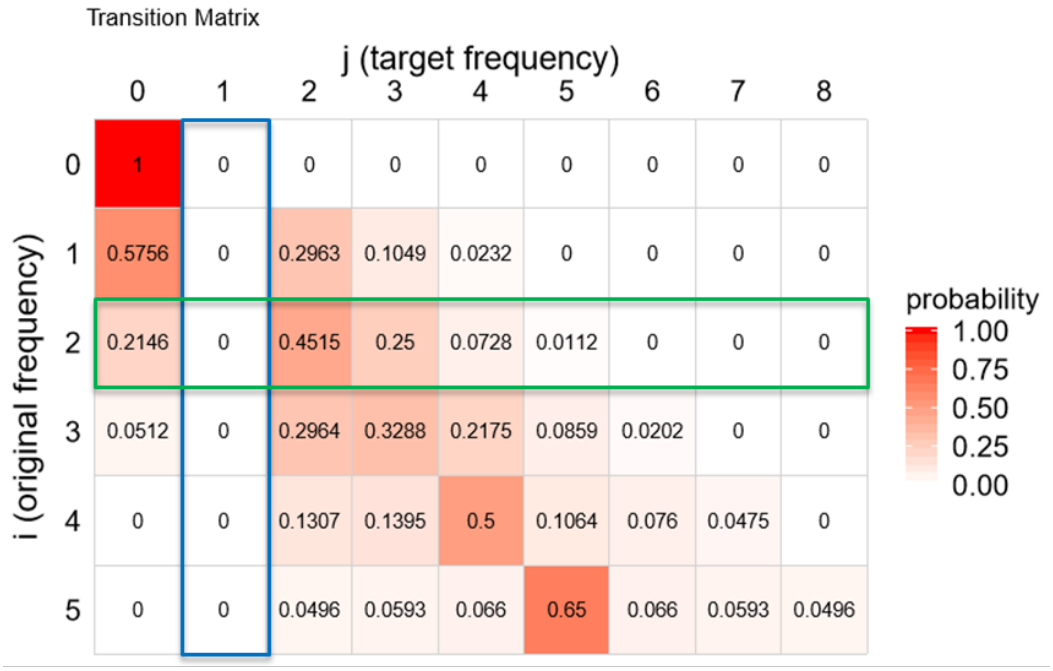
- ***js***: the perturbations shall not produce target frequencies equal to or below this threshold value (default is zero);
- ***pstay***: the probability of an original frequency to remain unperturbed (default is no preset probability (NA); this default produces the maximum entropy solution).

Based on the testing of different parameters, STATEC determined the required parameters as follows:

D=3, V=1.5, js=1, pstay = c(0.3,0.45,0.3,0.5,0.65)

With this setting, Figure 8 represents the transition matrix, which is a *heatmap* indicating the probability of the perturbation of the original value (the darker the background of a transition probability, the higher the probability).

Figure 8: Transition matrix with the adjusted parameters



Since the threshold value was set to *1*, ones were not allowed, hence the corresponding column in the transition matrix only consists of zero probabilities. For instance, when the original value is *2*, the probability that it becomes zero is 0.2146, to stay as *2* is 0.4515, etc., and because *D = 3*, it will never become *6* or higher. The setting of the *pstay* parameter was important, since for larger original values (which are not in the transition matrix) the distribution of the noise would remain the same.

7.3. Results from the cell key method

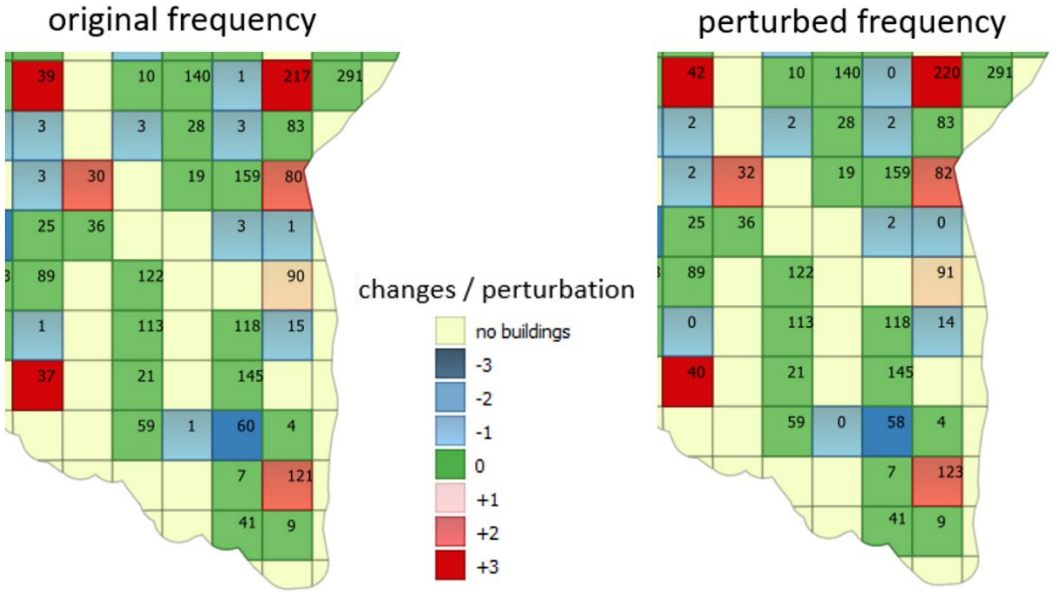
Examples of the cell key method's results are shown in Table 6, where the variables with *UWC* are the original–, while with *pUWC* are the perturbed frequencies; *build1* is in relation to one-family houses, *build3* shows the apartment counts and *build4* is linked to the number of semi-residential houses.

Table 6: Results from the cell key method (examples)

Dwelling_ID	*UWC_Total*	*pUWC_Total*	*change*	*UWC_build1*	*pUWC_build1*	*change_build1*	*UWC_build3*	*pUWC_build3*	*change_build3*	*UWC_build4*	*pUWC_build4*	*change_build4*
1	1873	1873	0	104	104	0	531	528	-3	1238	1238	0
2	2781	2780	-1	470	470	0	1512	1512	0	799	798	-1
3	3243	3245	2	588	587	-1	2041	2041	0	614	614	0
4	1805	1805	0	286	286	0	1265	1266	1	254	251	-3
5	6891	6893	2	283	283	0	3386	3386	0	3222	3222	0
6	2468	2466	-2	670	670	0	1591	1590	-1	207	204	-3
7	950	952	2	289	289	0	603	601	-2	58	58	0

Figure 9 below shows the results of the perturbation in the *build1* variable as two maps: a grid map with the original counts and another map with the perturbed frequency counts. Both illustrations concern the one-family houses-variable and display the Southeastern part of Luxembourg.

Figure 9: Example of the perturbation between the original and perturbed counts



7.4. Information loss measurement

In order to evaluate the confidentiality method, some measurements have been performed. Table 7 shows how each variable was perturbed. For instance, for the *build1* variable, the mean (-0.074) shows that it had a more negative perturbation, the maximum change is ±3 and the quartiles show that there were more +1 changes.

Table 7: Basic statistics of the loss measurements

	Min	*Q10*	*Q20*	*Q30*	*Q40*	*Mean*	*Median*	*Q60*	*Q70*	*Q80*	*Q90*	*Max*
Total	-3	-2	-1	0	0	-0.065	0	0	0	1	1	3
build1	-3	-2	-1	0	0	-0.074	0	0	0	1	1	3
build3	-3	-2	-1	0	0	0.09	0	0	1	1	1	3
build4	-3	-2	-1	-1	0	-0.093	0	0	1	1	1	3

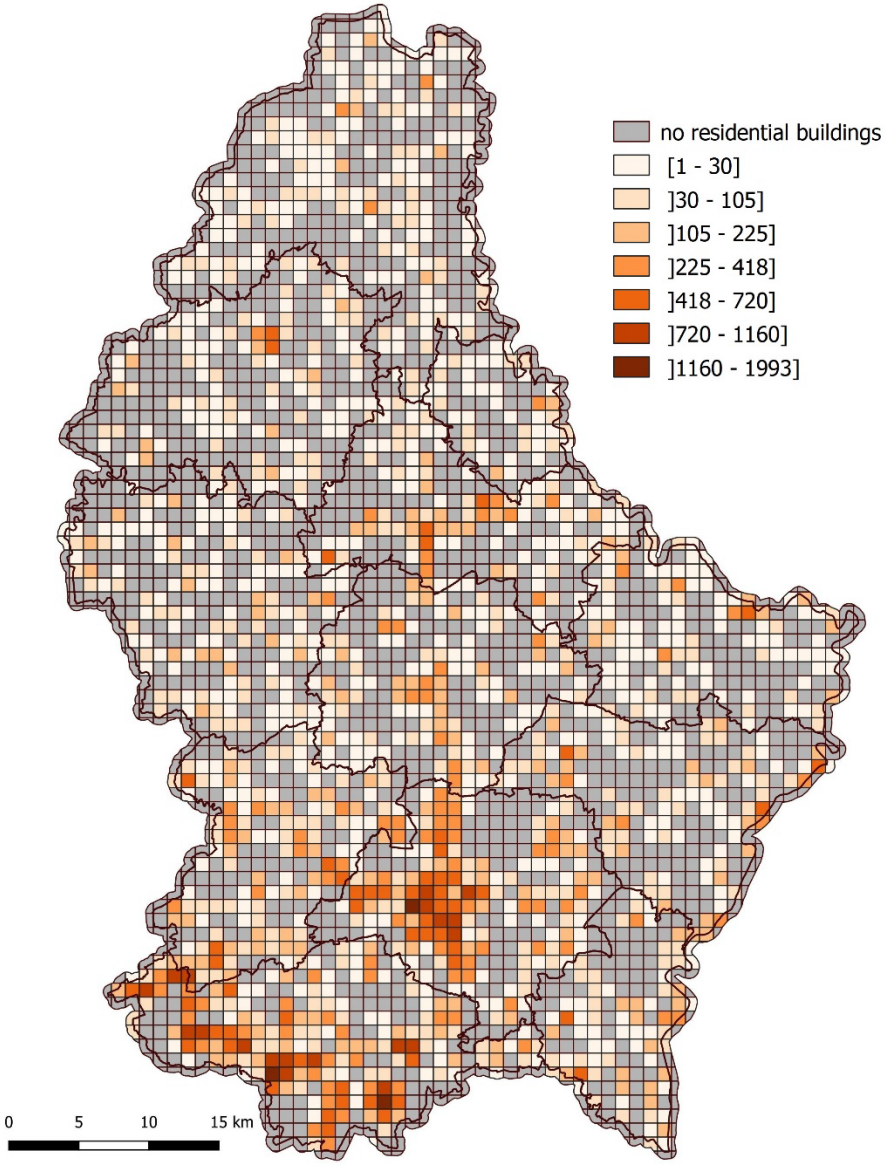
Another significant point is the number of grid units which were left unchanged. According to this test, 0 is the absolute distance between the original and perturbed value of the *build1* variable in 724 out of the total 1 565 grid units (46,3 %). Furthermore, there are 139 grid units which were non-zeros and perturbed to zero (called false zero). The cell key method doesn't allow false positives in the results; it is therefore impossible to have cells that were initially zero but have been perturbed to a number different from zero.

8. Examples of grid-based data display, using the 2019 *Registre des Bâtiments et Logements* database

The following section offers selected examples of how data stemming from the Register of Buildings and Dwellings have been confidentialised using the cell key method and how they can be displayed. The data are reflecting the situation at the end of 2016 (snapshot of the register: 01/03/2019).

8.1. Number of residential buildings

Map 1: Grand-Duchy of Luxembourg:
Number of residential buildings* per km², 2016



Source: STATEC, Registre des Bâtiments et Logements, Administration du Cadastre et de la Topographie (ACT)
* "Residential buildings" means the number of constructions for residential purposes (**not** residential units). These can be free-standing houses, a row of single-family houses or a block with apartments.

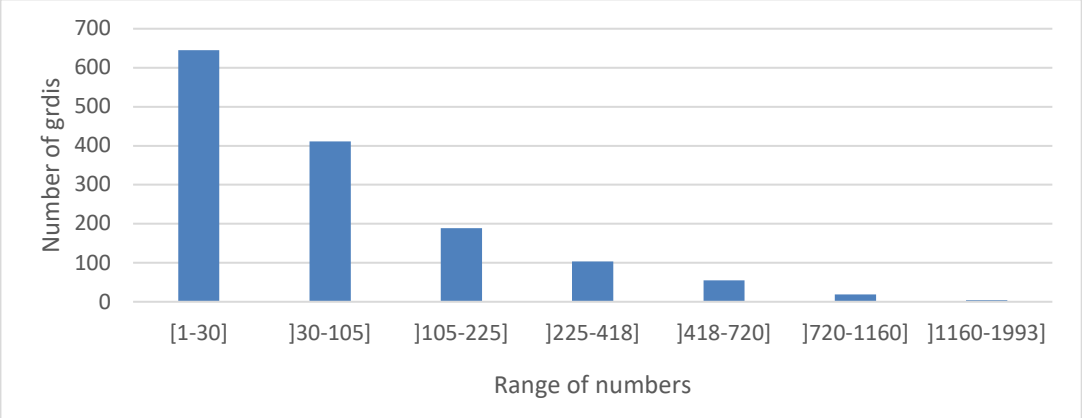
For the purpose of this paper, the categories shown in the legends of the maps have been determined by the Jenks natural breaks classification method. In the Jenks algorithm, classes are based on natural groupings inherent in the data. Class breaks are identified that best group similar values and that maximise the differences between classes. The features are divided into classes whose boundaries are set where there are relatively big differences in the data values. Of course, the magnitude classes can be adapted according to the needs.

Map 1 represents the perturbed, i.e. publishable counts of the residential and semi-residential buildings. The highest densities are obviously found in the city of Luxembourg (especially in the areas of Belair, Limpertsberg, Gare Central and Bonnevoie-Sud) and in the major municipalities of the canton of Esch-sur-Alzette, in particularly Differdange, Dudelange, Schifflange, Pétange and, of course, Esch-sur-Alzette.

The map clearly shows a higher density of residential buildings along the valley of the Alzette from Luxembourg-City towards the north, along the A7 motorway and the rail line. Higher densities are also noted for Ettelbrück, Diekirch and Wiltz.

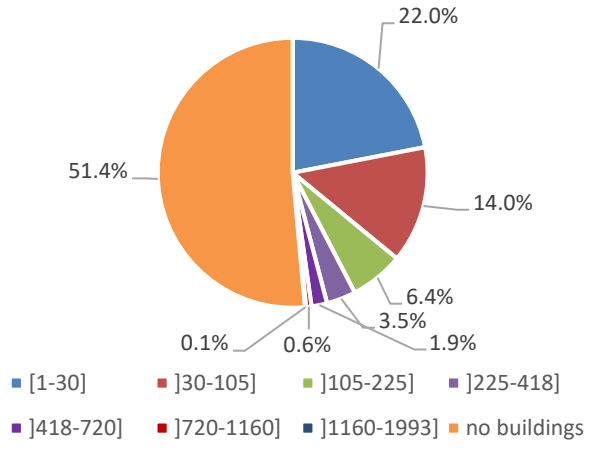
Among the total of 2 937 grid cells that cover the country, 1 511 show no residential buildings, which represents a share of 51.4 %.

Chart 1: Grand Duchy of Luxembourg: number of residential buildings, 2016



Source: STATEC, Registre des Bâtiments et Logements, ACT

Chart 2: Grand Duchy of Luxembourg: residential buildings – share of frequency classes, 2016

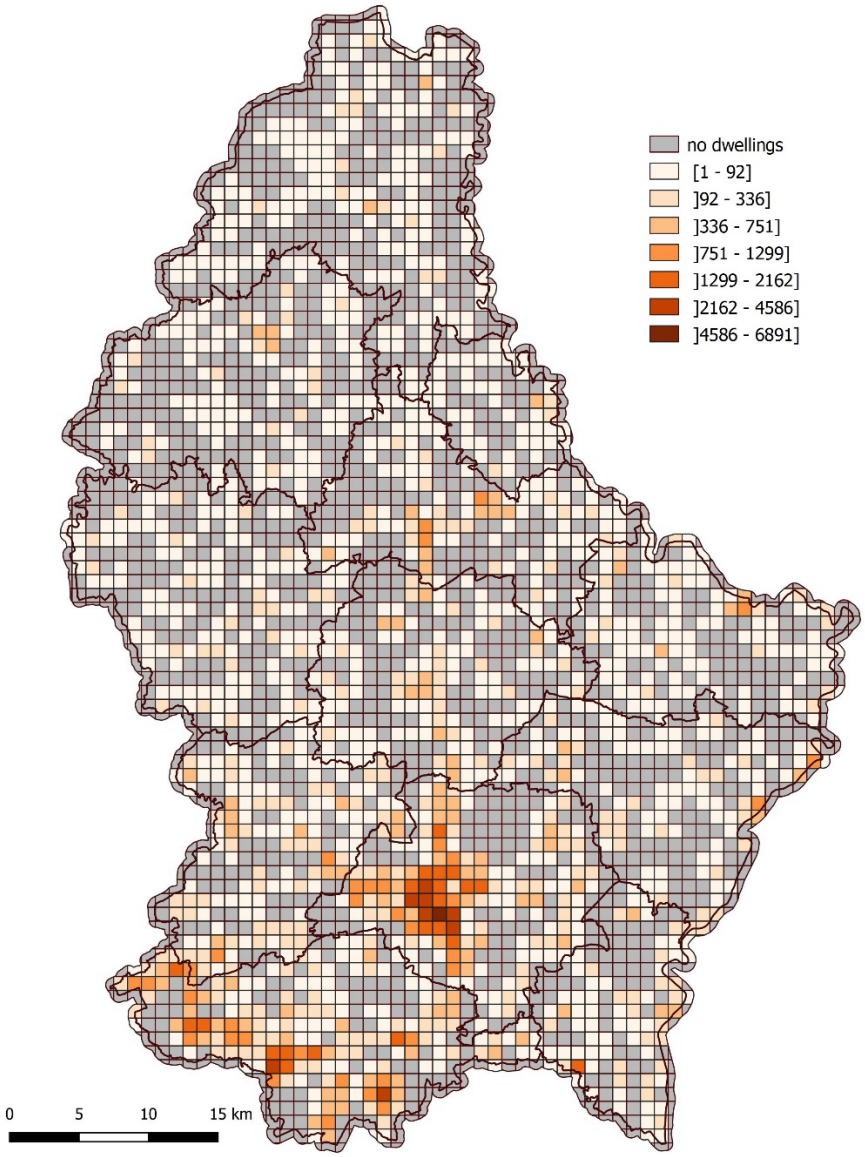


Source: STATEC, Registre des Bâtiments et Logements, ACT

8.2. Number of residential dwellings

Map 2 shows the absolute number of dwellings, i.e. the individual residential units that may be single family houses (detached or semi-detached), apartments in apartment buildings or apartments in semi-residential buildings (where a part of the building, typically the ground floor, is used for commercial or other non-residential purposes). The map shows the same basic pattern as the previous one, but the numbers are obviously higher, especially in urban areas. The same areas with high counts can be detected (the capital, Käerjeng, Dudelange, Esch-sur-Alzette, Differdange, Bettembourg) and, to a lesser degree, Ettelbrück and Diekirch. Comparatively high values are also noted for the municipalities of Echternach, Mertert and Mondorf. Unsurprisingly, the highest numbers are registered in Luxembourg-City, and here primarily in the areas of Hollerich and Bonnevoie.

Map 2: Grand-Duchy of Luxembourg: Number of dwellings*, per km², 2016



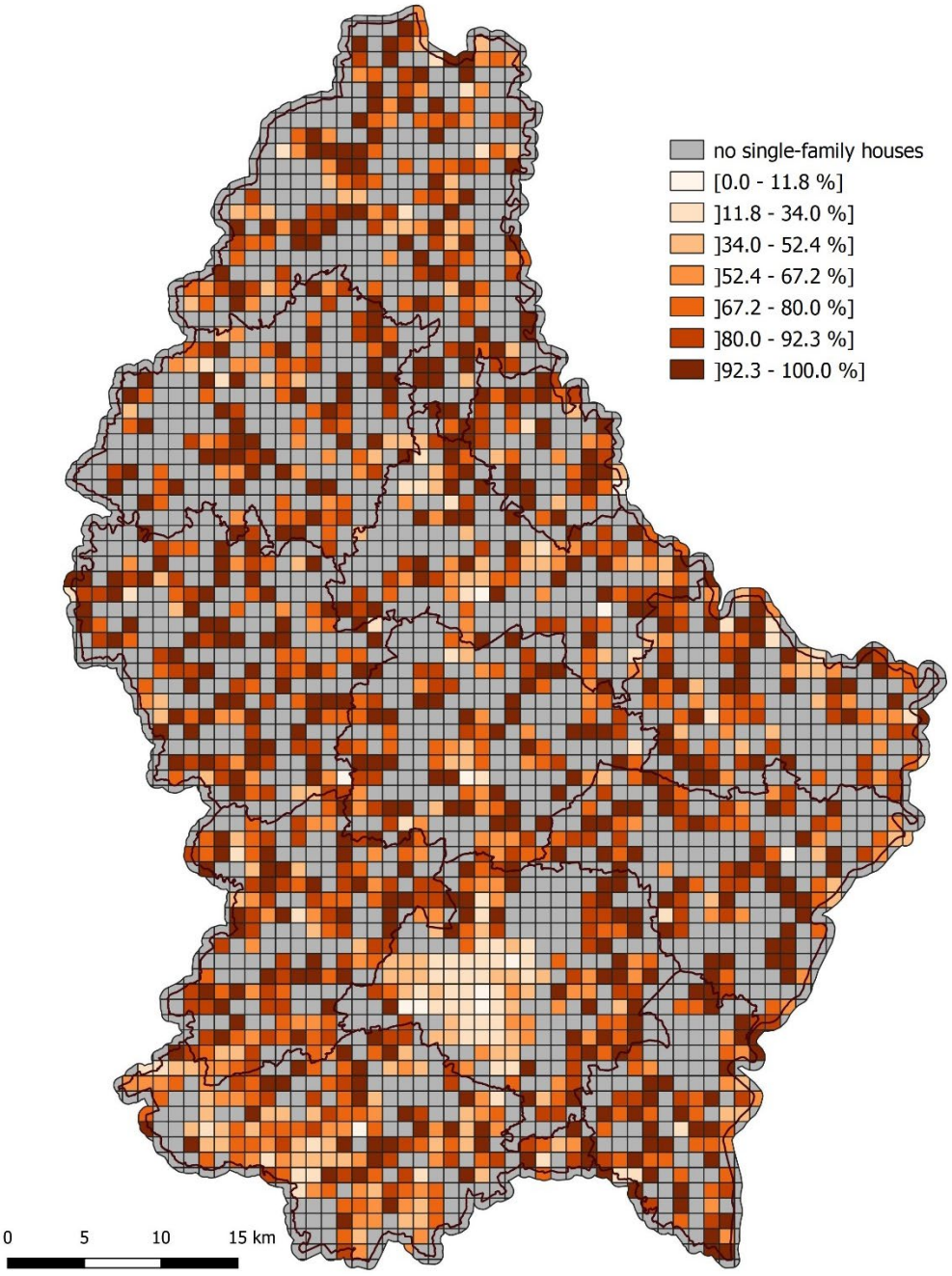
Source: STATEC, Registre des Bâtiments et Logements, ACT

* Dwellings refer to residential units that can either be single-family houses (4-,3-or 2-facades), individual apartments in residential buildings, or individual apartments in semi-residential buildings.

8.3. Share of single-family houses

At national level, 51.3% of all dwellings are single-family houses (free-standing, semi-detached, or in a row). Map 3 shows that most single-family houses are in the more rural areas of the country, with the highest shares registered in the grid cells covering the municipalities of Stadtbredimus, Frisange, Niederanven, Consdorf, Beaufort, Beckerich, Putscheid, Boulaide and Kiischpelt. Conversely, certain grid cells of Luxembourg-City and its immediate surroundings show the lowest shares.

Map 3: Grand-Duchy of Luxembourg: share of single-family houses* in the total number of dwellings, per km², 2016

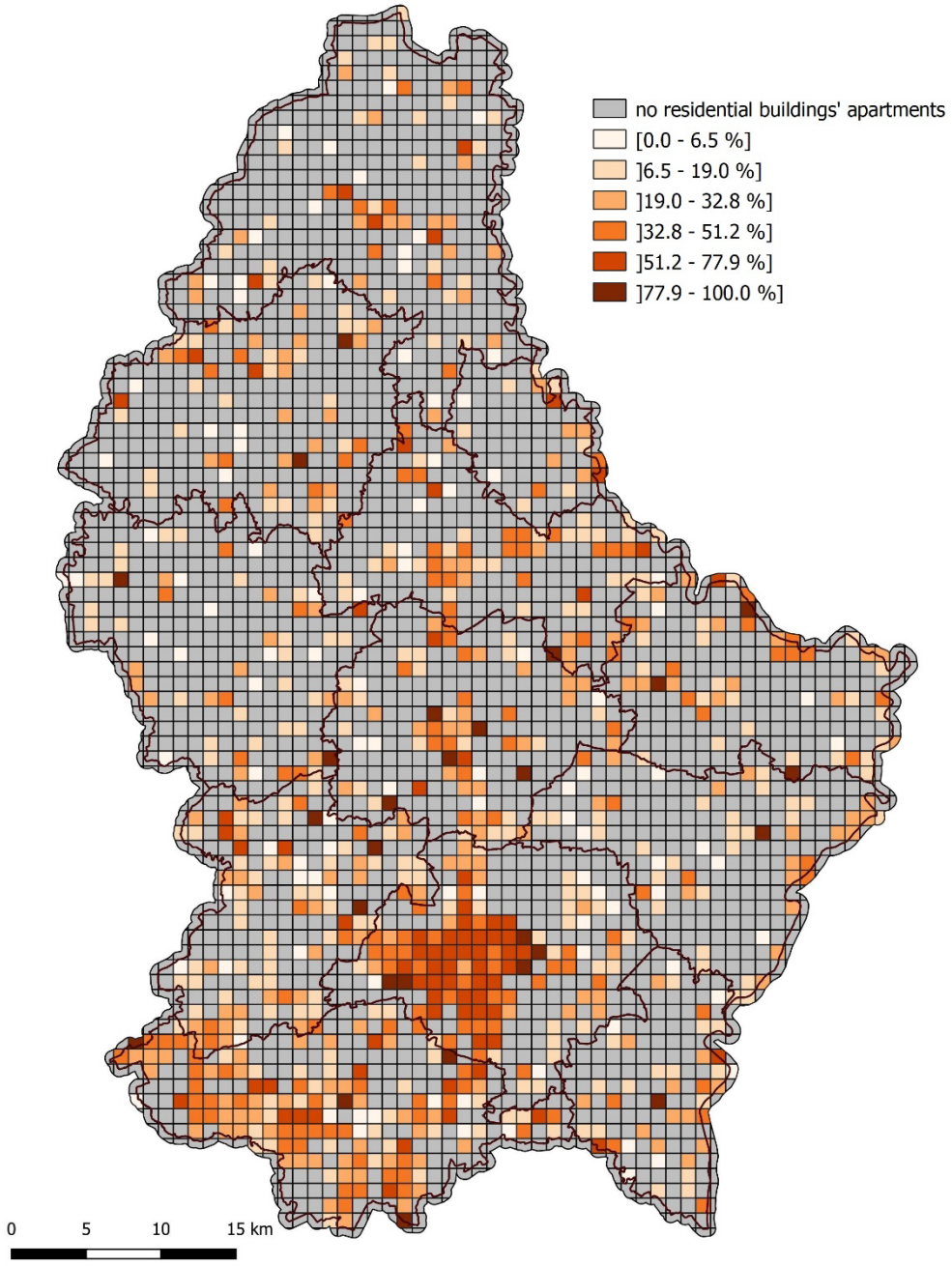


Source: STATEC, Registre des Bâtiments et Logements, ACT
* Single-family houses : detached (4-facades) or semi-detached (3- or 2- facades) residential units.

8.4. Share of residential buildings' apartments

Map 4 shows the opposite situation with the shares of residential building apartments in the residential units' total. For the country as a whole, 35 out of every 100 residential units are apartments in residential buildings. Obviously, the percentage is considerably higher in the densely populated municipalities such as Luxembourg-City (and then especially in the grid cells covering the areas of Hamm, Weimerskirch, Merl and Kirchberg) but also in the grid cells belonging to Mersch or Wiltz.

Map 4: Grand-Duchy of Luxembourg: share of residential buildings' apartments in in the total number of dwellings, per km², 2016

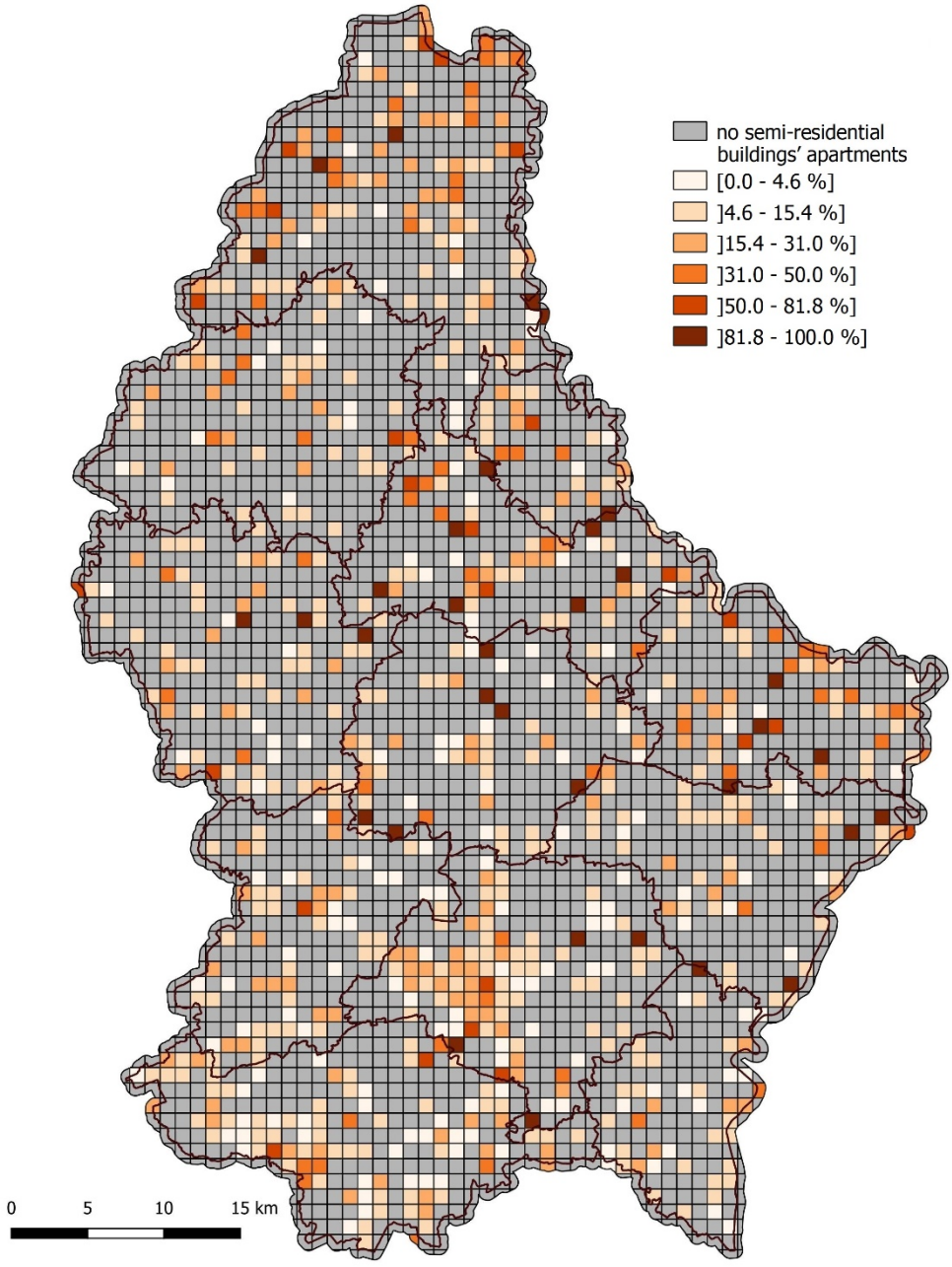


Source: STATEC, Registre des Bâtiments et Logements, ACT

8.5. Share of semi-residential buildings' apartments

Finally, Map 5 shows the share that apartments in semi-residential buildings take in the total number of residential units. With an average share of 13.7% at national level, the geographical pattern is less sharp as it shows a limited number of grid cells with high percentages. Moreover, these are spread across the entire country. Noteworthy are the 20 grid cells of the highest category (81.8%-100%) registered in the Diekirch canton, and 14 grid cells in both the Grevenmacher and Echternach canton. Conversely, the cantons of Remich and Capellen only feature 2 grid cells of the highest category.

Map 5: Grand-Duchy of Luxembourg: share of semi-residential buildings' apartments in the total number of dwellings, per km², 2016



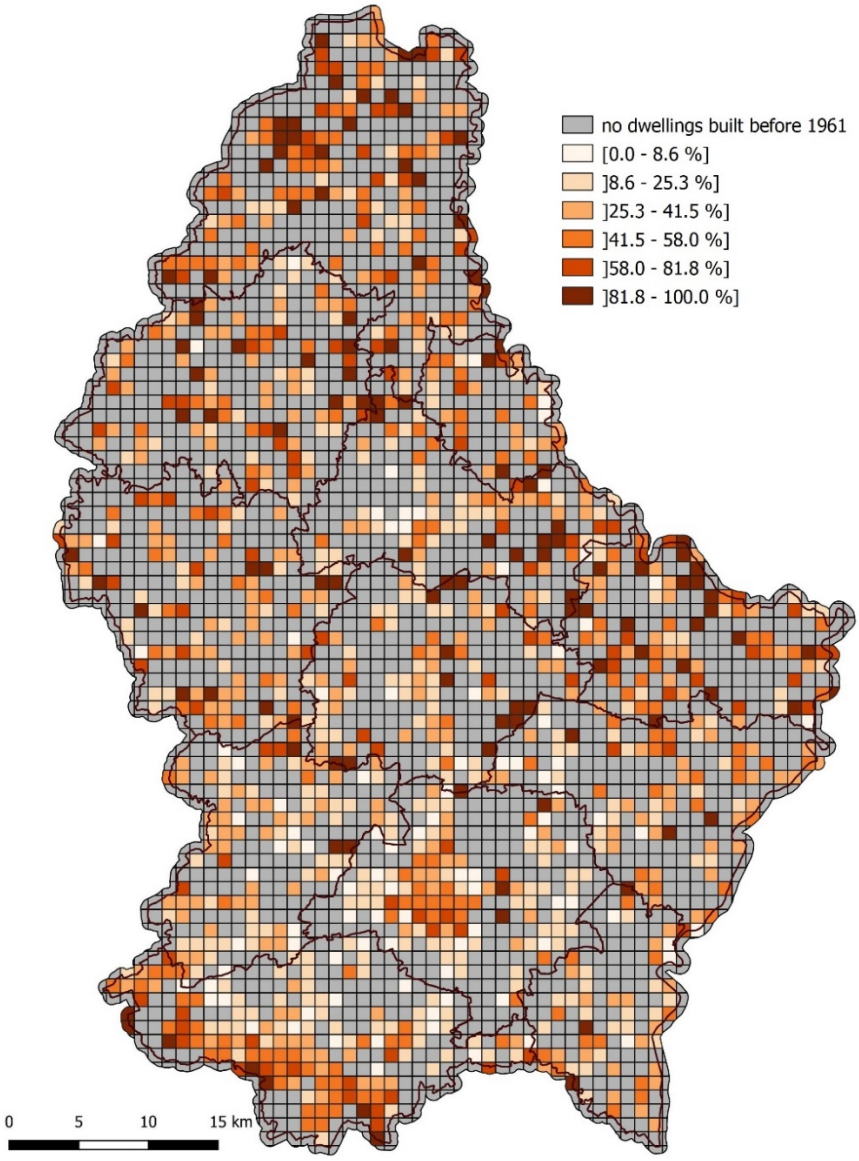
Source: STATEC, Registre des Bâtiments et Logements, ACT

8.6. Dwellings by periods of construction

The next series of maps (Maps 6-8) focus on the age of the residential units and their spatial distribution in the country according to three age categories: residential units built before 1961, between 1961 and 2000, and after 2000.

Map 6 shows the share of residential units (irrespective of the type) built before 1961. At the level of the country 34.6% of all dwellings were built before 1961. The grid cells with the highest percentages are mainly in the northern part of the country (for instance those covering municipalities such as Wincrange, Weiswampach, Troisvierges), in the north-east (Berdorf, Consdorf, Waldbillig and Echternach), and in the south-west of the Esch-sur-Alzette canton (in particular the municipalities of Differdange, Esch-sur-Alzette and Dudelange).

Map 6: Grand-Duchy of Luxembourg: share of dwellings built before 1961 in the total number of dwellings, per km², 2016

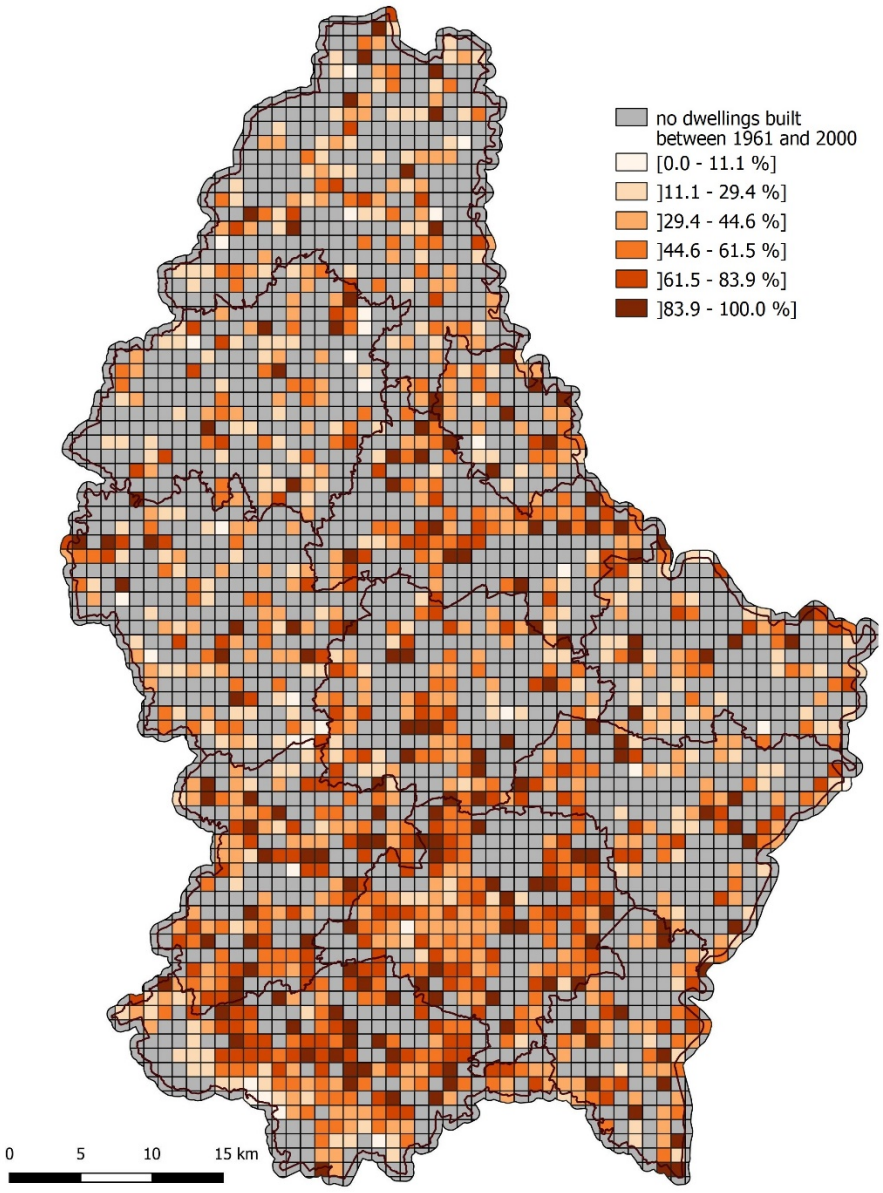


Source: STATEC, Registre des Bâtiments et Logements, ACT

From 1961 to 2000, the southern part of the Grand-Duchy has experienced a very substantial activity in terms of residential construction (see Map 7 below): 43.6% of all dwellings on the territory of the Grand Duchy were built in this period. Much higher proportions can be found in grid cells in the south-west part of the country in the municipalities such as Contern, Bettembourg, Mondercange, Sanem and Roeser. Grid cells along the Moselle between Schengen and Wormeldange also show relatively high shares.

Residential building activities along the Alzette-valley between Luxembourg-City and Mersch also become obvious (Walferdange, Steinsel, Lorentzweiler, Lintgen). Further north, the eastern part of the Diekirch canton shows high shares, especially the grid cells of Bourscheid municipality (for instance those covering the Lipperscheid and Bourscheid localities) and in municipalities such as Bettendorf and Reisdorf.

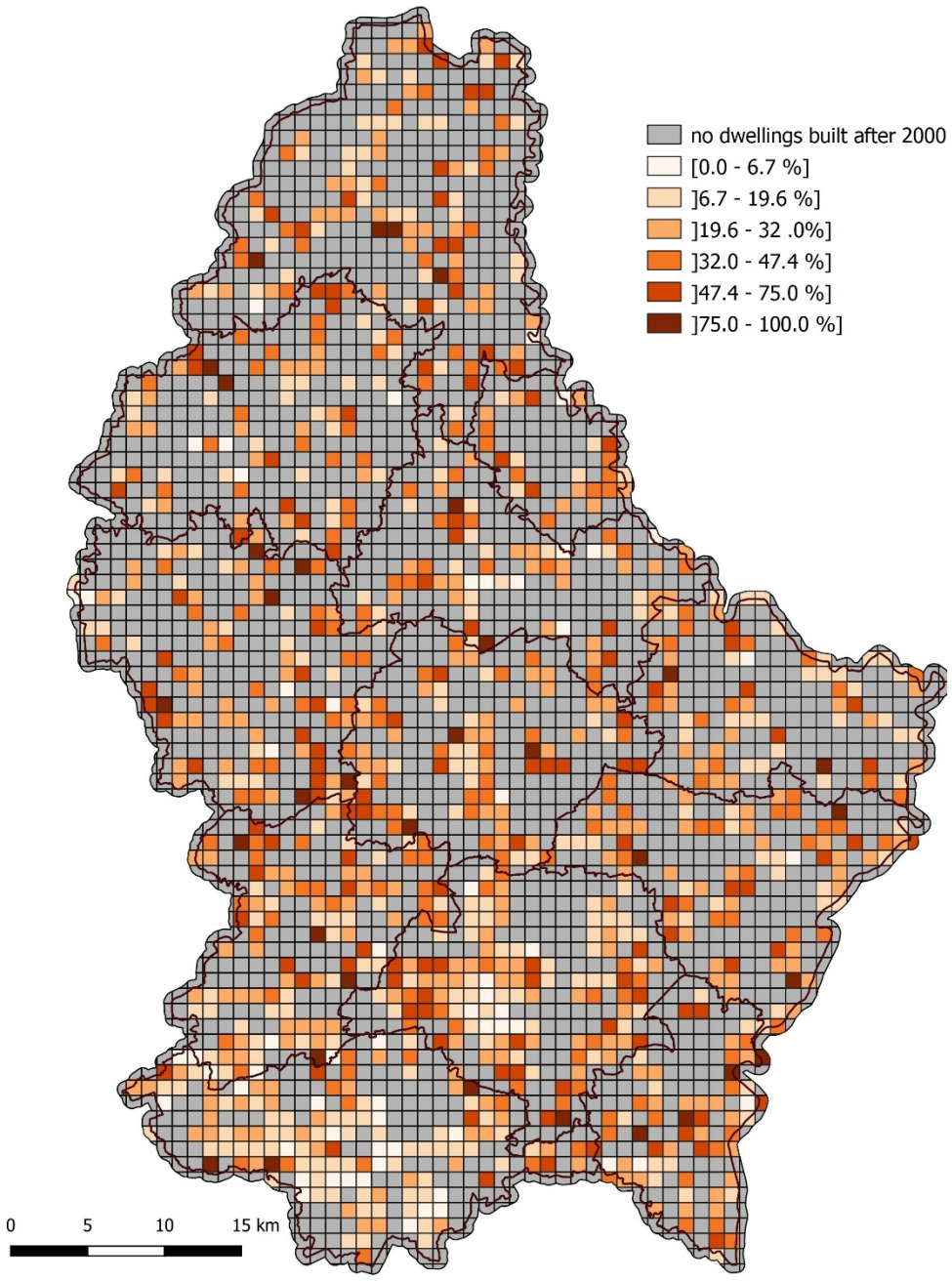
Map 7: Grand-Duchy of Luxembourg: share of dwellings built between 1961 and 2000 in the total number of dwellings, per km², 2016



Source: STATEC, Registre des Bâtiments et Logements, ACT

Map 8 displays the situation of residential building after the year 2000. Compared to the previous map, covering a timespan roughly twice as long, it is only normal that the pattern appears "lighter" overall (share at country level (21.8%). Nevertheless, some grid cells show a high proportion of relatively recent residential constructions, especially those located in the area where the cantons of Capellen, Mersch and Redange meet (municipalities such as Saeul and Helperknapp), but also in the eastern part of the Manternach municipality, in Weiler-la-Tour and Roeser as well as in the south of Stadtbredimus, in the east of Waldbillig and in the north of Winseler, to name but a few.

Map 8: Grand-Duchy of Luxembourg: share of dwellings built after 2000 in the total number of dwellings, per km², 2016

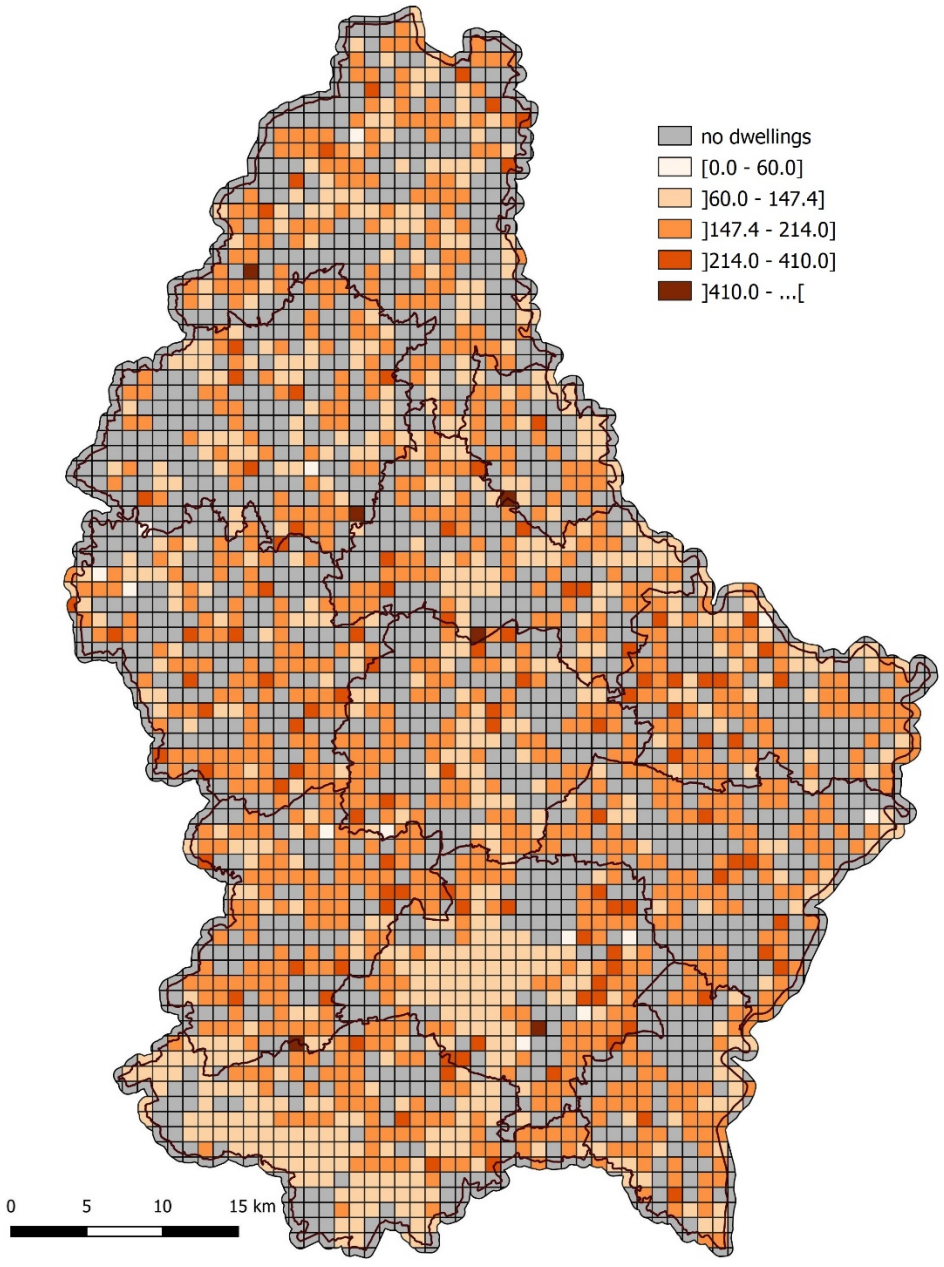


Source: STATEC, Registre des Bâtiments et Logements, ACT

8.7. Dwellings' average surface

The *Registre des Bâtiments et des Logements* also features information on the average size (in m²) of the dwellings. This is presented on a 1 km² grid in Map 9, in which all types of dwellings are considered. Like in Map 1, the grey grid cells are the areas where no dwellings exist. The urban areas of the country (the capital, Esch-sur-Alzette, Dudelange, Differdange, Schifflange, Bettembourg, Pétange, Ettelbruck and Diekirch) show dwellings with the lowest average surface, while there are a few grid cells with a very high average, essentially influenced by a few very large residential units. On average, the Luxembourg dwelling has a surface of 131.35 m².

Map 9: Grand-Duchy of Luxembourg: dwellings' average surface (m²), per km², 2016



Source: STATEC, Registre des Bâtiments et Logements, ACT

9. Summary

Along the lines of the European legislation, STATEC must publish selected items from the 2021 population and housing census by a 1 km^2 grid. As this is a new and far more detailed geographical level for the publication of survey results, there is a need to work on data confidentiality.
EUROSTAT suggests using the targeted record swapping method with the cell key method to deal with confidentiality issues. STATEC has investigated the application of the cell key method by using a dataset from the *Registre des Bâtiments et Logements* and generated test results respecting all principles of data protection.
The methodology selected and tested is now ready to be adopted for the upcoming 2021 census. Furthermore, STATEC has already started to investigate the use of the targeted record swapping method, the other method recommended for censuses by Eurostat.

10. Bibliography

European Commission (EU) Regulation 2018/1799: the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km^2 grid, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1799&rid=1

Antal, L. – Enderle, T. – Giessing, S. (2017): Harmonised protection of census data in the ESS, Development and testing of recommendations; identification of best practices, Statistical disclosure control methods for harmonised protection of census data

Benschop, T. – Machingauta, C. – Welch, M. (2018): Statistical Disclosure Control: A Practice Guide

EUROSTAT (2019): EU legislation on the 2021 population and housing censuses – Explanatory notes

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., De Wolf, P.-P. (2012): Statistical Disclosure Control. Chichester, Wiley Series in survey methodology

Moore, R. A.: Controlled data-swapping techniques for masking public use microdata sets, U.S. Bureau, https://www.census.gov/srd/CDAR/rr96-04_Controlled_DataSwapping.pdf

Shlomo, N. (2007): Statistical Disclosure Control Methods for Census Frequency Tables, International Statistical Review / Revue Internationale de Statistique, Vol. 75, No. 2(August)

Shlomo, N. - Tudor, C. – Groom, P. (2010): Data Swapping For Protecting Census Tables, https://www.researchgate.net/publication/221144044

Spicer, K. – Dove, I.: Progress towards a table builder with in-built disclosure control for 2021 Census, Office for National Statistics (UK).

Sukasih, A. – Jang, D. – Czajka, J. (2012): Implementing Multiple Evaluation Techniques in Statistical Disclosure Control for Tabular Data, Mathematica Policy Research