RDC Working Paper

No. 53

The cell key method at the Research Data Centres of the Statistical Offices of the Federation and the Federal States

Part 1: Introduction of the new confidentiality procedure

Stefanie Setzer, Johannes Rohde, Volker Güttgemanns, Patrick Rothe

Published: January 2026

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 9449-8000
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2876
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2876
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Januar 2026

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

This publication is an English translation of an article originally published in the 03/2024 issue of Wista - Wirtschaft und Statistik. The original article can be found on www.destatis.de.

The Cell Key Method at the Research Data Centres of the Statistical Offices of the Federation and the Federal States

Part 1: Introduction of the new confidentiality procedure

Stefanie Setzer[1], Johannes Rohde[2], Volker Güttgemanns[3], Patrick Rothe[4]

Abstract:

The Research Data Centres of the Statistical Offices of the Federation and the Federal States introduce the cell key method for selected statistics as a new procedure for maintaining confidentiality. This method protects respondents from re-identification by creating uncertainty about the number of cases actually contributing to the result by perturbing the number of cases. The article presents how the cell key method works and provides both an easy-to-understand introduction to the topic and detailed methodological information.

Key Words: data protection, confidentiality, stochastic perturbation, disclosure control, post-tabular

[1] Research Data Centre of the Federal Statistical Office Germany, e-mail: stefanie.setzer@destatis.de
[2] "Mathematical-statistical methods and experimental statistics" at Information und Technik Nordrhein-Westfahlen, e-mail: Johannes.Rohde@it.nrw.de
[3] Research Data Centre of the Federal States, main office (formerly), e-mail: Volker.Guettgemanns@it.nrw.de
[4] "Fundamental questions of official statistics, Digitalisation, Research Data Centre, Centre of Competence Analses" at the Bayrisches Landesamt für Statistik, e-mail: Patrick.Rothe@statistik.bayern.de

Inhalt

I. Introduction 5
II. Statistical disclosure control in the Research Data Centres 6
1. Why do the Research Data Centres take confidentiality so seriously? 6
2. The current standard: cell blocking 6
3. The new confidentiality procedure: the cell key method 7
III. Functionality of the cell key method 8
1. Record Keys 8
2. Transition matrix 3
3. Perturbation tableau 4
4. Table creation 4
IV. Formal explanation of the cell key method 7
1. Requirements for the perturbation values 8
2. Parameters of the cell key method 8
3. Transition matrix 9
4. Further optional restrictions for the calibration of the transition matrix 10
5. Calculation of the transition matrix 10
6. Record keys and cell keys 11
7. Assignment of the perturbation values 12
V. Conclusion 12
VI. References 13

I. Introduction

The Research Data Centres (RDC) of the Statistical Offices of the Federation and the Federal States provide microdata for scientific use. To ensure data protection, there are two options: either anonymisation, in which data is modified in such a way that no confidentiality risks can arise during evaluation before it is made available to the scientific community, or confidentiality, in which the data is protected by modifying the results. The procedure is chosen depending on the way of data access: In the so-called off-site ways of access, where the data is transmitted to the users, the data is anonymised. This is always accompanied by a loss of information. When using on-site ways of data access, i.e. a safe centre or remote execution, the data remains in the secure premises of the statistical offices. There, the full information potential of the data can usually be retained. However, the generated results are subject to a confidentiality check.

Anyone who has already used one of the Research Data Centres on-site ways of access knows the effects of this confidentiality check: After the results have been made available, three large X often catch the viewers eye. The Research Data Centres usually use this blocking pattern if the publication of a result represents a confidentiality risk. But why are the Research Data Centres so concerned about confidentiality? And are there alternatives to the three big X's?

Chapter 2 answers the question of why confidentiality is so important to the Research Data Centres. The cell key method as an alternative to blocking with the three capital X is explained in Chapter 3. Chapter 4 then formally explains the methodology of the cell key method. A short conclusion with a reference to the second part of the article concludes the paper.

II. Statistical disclosure control in the Research Data Centres

1. Why do the Research Data Centres take confidentiality so seriously?

The answer to this question is simple: because they are legally obliged to do so. The obligation to maintain confidentiality is regulated in Section 16 of the Federal Statistics Act (Bundesstatistikgesetz – BStatG). According to this, "individual details about personal and factual circumstances that are provided for federal statistics [...] must be kept confidential"[5] (Section 16 (1) BStatG). However, this paragraph also regulates exceptions that allow statistical offices and the scientific community to publish data and results under certain conditions. Publication is possible, for example, if the individual data has been combined with the results of other respondents or if the individual data cannot be attributed to the persons concerned. These two exceptions allow data and results to be made available and at the same time constitute the obligation to maintain confidentiality. The publication of results is permitted as long as the results allow no conclusions about individuals.

The reason for this legal regulation and the therewith stated high value of confidentiality is easy to understand: Official statistics surveys often state the obligation for respondents to provide information. Respondents - be they individuals, companies, businesses or others - are therefore often unable to decide for themselves what specific information they wish to reveal. In order to compensate for this interference with informational self-determination, the legislator guarantees respondents that their information cannot be assigned to them. The same applies to surveys with voluntary participation.

The respondents' trust in the non-attributability of their information is the basis which ensures that questions are answered truthfully without fear of disclosure. It therefore contributes significantly to the high quality of the data. The protection of the entrusted data is therefore always the top priority for official statistics - and therefore also for the Research Data Centres.

2. The current standard: cell blocking

Up to now, the statistical offices have generally ensured confidentiality by means of cell blocking[6]. With this form of confidentiality, all information that poses a confidentiality risk is replaced by a blocking pattern ("XXX"). This procedure has proven itself in official statistics, but has some serious disadvantages:

- Loss of information: Using cell blocking, not only the critical data itself is blocked (primary blocking). To prevent these values from being recalculated, they must be counter-blocked with non-critical data (secondary blocking). If blocked data can be recalculated via other tables,

[5] German original: "Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind [...] geheim zu halten."
[6] In addition, the method SAFE – Sichere Anonymisierung Für Einzeldaten (Secure Anonymisation of microdata) was used for the census 2011 (Höhne, 2015).

blocks must also be inserted there (cross-table blocking). This means that a single value that has to be blocked can quickly result in a large number of other blocks of values that are not critical themselves.

- High effort: Since the confidentiality check for cell blocking cannot usually be fully automated, the confidentiality check is very time-consuming and resource-intensive and users sometimes have to wait a long time for their results.
- Dissatisfaction: This method implies that not all values of interest can be published. In some cases, entire tables have to be blocked. Cell blocking therefore often leads to dissatisfied data users.

3. The new confidentiality procedure: the cell key method

Due to the disadvantages of cell blocking described above, the statistical offices have decided to introduce the cell key method (CKM) for some first statistics.[7] This decision has a direct impact on the provision of data in the Research Data Centres as they always ensure confidentiality in accordance with the confidentiality rules defined by the specialised departments. The cell key method is a data-modifying procedure to ensure the confidentiality of frequency count tables. The advantages of the method are:

- Using the cell key method there are no blocked cells.
- The results show a high data quality and cross-table consistency.
- The effort for the confidentiality check is notably less.
- Disclosure risks through mistakes in the cross-table confidentiality check can be ruled out.

However, these advantages are accompanied with some disadvantages:

- Using the cell key method, tables are no longer additive.
- Especially for frequency counts, the perturbation can be relatively high.
- Since the cell key method is originally a procedure to ensure the confidentiality of frequency count tables, additional effort may be required to check the results of multivariate analyses.
- The cell key method is less intuitive than other confidentiality methods and therefore requires extensive explanations.

The cell key method is presented below for the basic form of ensuring the confidentiality of frequency count tables.

[7] The Cell-Key-Method was originally developed by the Australian Bureau of Statistics (ABS).

III. Functionality of the cell key method

The cell key method is a post-tabular data-modifying confidentiality procedure. This means that the procedure starts with the generation of results and confidentiality is achieved by modifying frequency counts. The protective effect is achieved by creating uncertainty with regard to the original frequency count by imposing an error term on each table value. The procedure ensures that the perturbation is consistent, i.e. that logically identical case numbers remain identical across all tables. Thus, marginal totals of a cross-tabulation, for example "federal state x age", always state the same values that are also stated in the corresponding univariate frequency count tables of the two variables. This applies regardless of whether the value appears as an internal or marginal table cell.

The following sections provide a step-by-step explanation of the functionality of the cell key method. Please note that a great advantage of the cell key method is that the realisation of confidentiality itself is mostly automated.

1. Record Keys

To apply the cell key method, an additional variable which contains the so-called record key is first added to the initial data set. It consists of a random number between 0 and 1 which is permanently assigned to each unit of observation.

To illustrate this, figure 1 shows data for a fictitious municipality in which the age and income of all adult residents are recorded in classified form. For clarity reasons, a record key with only two decimal places is assigned; in real use cases, the number of decimal places is usually higher.

ID	Age	Income	Record Key
1	young	medium	0,54
2	young	high	0,68
3	old	low	0,14
4	old	medium	0,93
5	young	medium	0,51
6	old	medium	0,37
7	old	low	0,84
8	old	high	0,19
9	old	medium	0,26
10	young	high	0,43
11	old	medium	0,99
12	young	medium	0,74
13	young	medium	0,65
14	old	low	0,79
15	old	medium	0,25

Original dataset | Assigned record key

Figure 1: Assigning the record keys to the original data set

2. Transition matrix

The transition matrix determines the value with which a frequency count is perturbed (Kleber/Gießing, 2018). For each original frequency count, the probability with which it will be modified to a certain other value is decided. In Table 1, for example, the original frequency count 10 would remain a 10 with a probability of 0.7, would become a 9 or 11 with a probability of 0.1 and would become an 8 or 12 with a probability of 0.05. That way, various framework conditions can be defined, for example the maximum deviation, the probability of retaining an original frequency count, the exclusion of 1 and 2 in the perturbed table or a higher probability of retention for higher frequency counts.

Original frequency	Perturbed frequency count																	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0,7	0	0	0,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0,3	0	0	0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0,5	0,35	0,15	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0,25	0,5	0,20	0,05	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05

Zeroes remain unmodified

No 1 or 2 in the perturbed table

Maximum deviation +/-2

Higher probalility of unmodified values for higher frequency counts

Figure 1: Fictional example of a transition matrix

Please note that this is a fictitious example of a transition matrix in order to simplify its possible design. It would also be possible, for example, to specify that all zeros in the perturbed table are real zeros, that no value remains unchanged or that large perturbations are more likely than small ones. As the transition matrix thus controls how (dis)similar the original and the perturbed table are, it represents the core of the cell key method and is designed with great effort. The actual transition matrices as well as the defined framework conditions are subject to strict confidentiality. Both are specifically adapted to the respective requirements of each statistic.

3. Perturbation tableau

A perturbation tableau is created on the basis of the transition matrix (Kleber/Gießing, 2018). To illustrate this, figure 3 shows the perturbation tableau based on the transition matrix from outline 1 as a grid. For its creation, the probabilities of each row of the transition matrix are cumulated. In the so-called lookup step, it leads to the perturbation value for the respective original number of cases based on the associated cumulative transition probability.

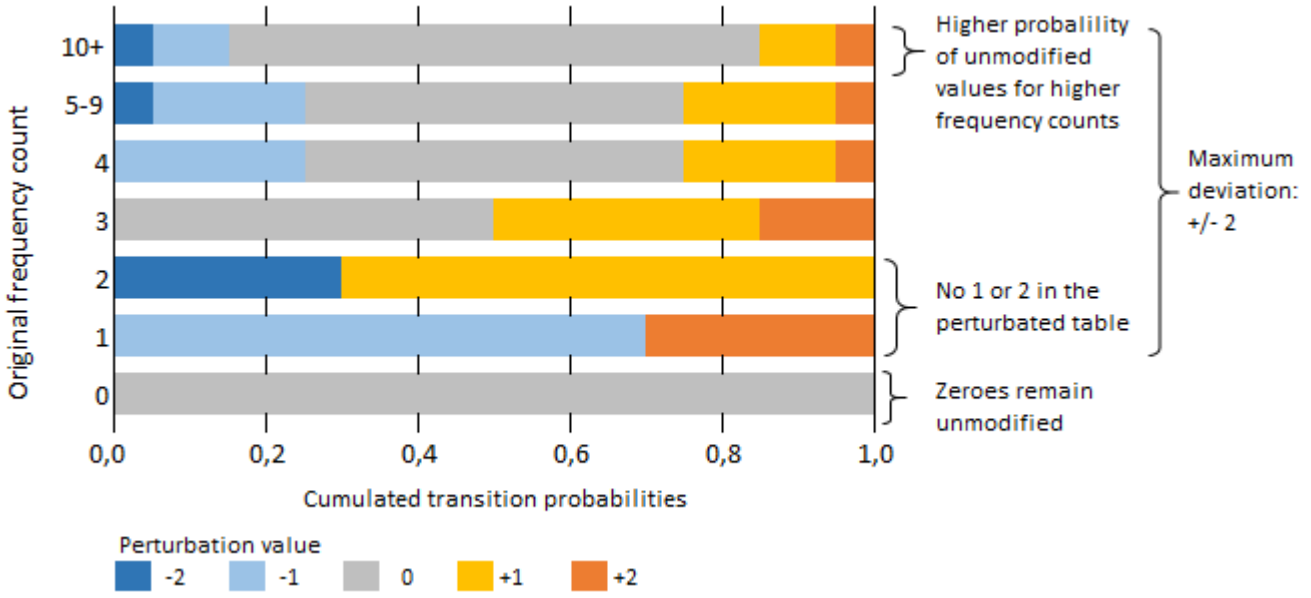

Figure 2: Fictional example of a perturbation tableau

For a better understanding, figure 3 shows an example of the modifications in the original frequency count "4".

Modification to	Equals perturbation with	Probability	Cumulated probability
3	-1	0,25	0,25
4	0	0,5	0,75
5	1	0,2	0,95
6	2	0,05	1

Figure 3: Cumulative transition probabilities for the original frequency count "4".

4. Table creation

When the perturbated tables are created, the original frequency count, the cumulative transition probability for this original frequency count and the record key generated at the beginning come together to determine the perturbation value:

When creating frequency count tables using the cell key method, not only the frequency counts are determined. For each table cell, the record keys of all observation units that contribute to the corresponding table cell are also aggregated. However, only the decimal places of the sum of the record keys are relevant, so the value before the decimal point is set to 0. This results in a value between 0 and less than 1, the so-called cell key.

Figure 5 illustrates this mechanism for the combination of the characteristics "young respondents with a medium income" and for the total of people with a low income.

ID	Age	Income	Record Key
1	young	medium	0,54
2	young	high	0,68
3	old	low	0,14
4	old	medium	0,93
5	young	medium	0,51
6	old	medium	0,37
7	old	low	0,84
8	old	high	0,19
9	old	medium	0,26
10	young	high	0,43
11	old	medium	0,99
12	young	medium	0,74
13	young	medium	0,65
14	old	low	0,79
15	old	medium	0,25

		Age		
		young	old	Sum
Income	low	0	3	3 ∑RK = 0,14+0,84+ 0,79 = 1,77 →Cell Key = 0,77
	medium	4 ∑RK = 0,54+0,51+ 0,74+0,65 = 2,44 →Cell Key = 0,44	5	9
	high	2	1	3
	Sum	6	9	15

Figure 4: Creating the cell keys for two examples by adding the corresponding record keys (RK) and setting the number before the decimal point to zero.

The calculated cell keys for all original frequency counts are now fed back to the perturbation tableau. Here, an alignment is made of the grid in which the cumulative probability corresponds to the determined cell key. The perturbation value for the respective original frequency count is then read from this column. The coloured marking of the relevant grid indicates the value with which the original frequency count is perturbed.

In the example, the cell key of 0.44 calculated for the four young people with a medium income results in a perturbation value of 0. This frequency count therefore remains unmodified. For the three people with a low income, the cell key is 0.77, which corresponds to a perturbation of +1 according to the perturbation tableau. This frequency count is therefore modified from 3 to 4.

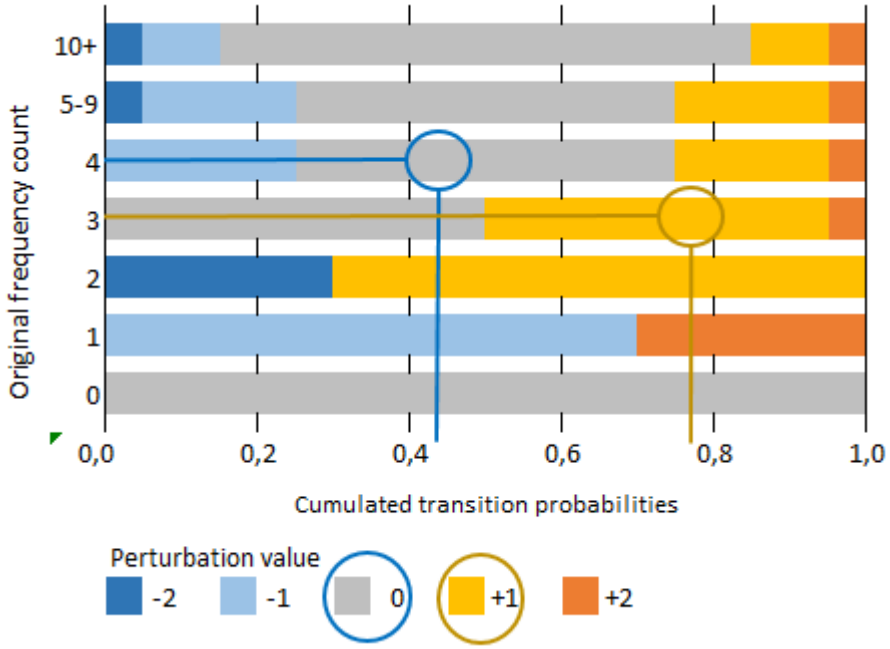

Figure 5: Lookup of the perturbation values from the perturbation tableau for the 4 young people with a medium income and a cell key of 0.44 (blue) and the 3 people with a low income and a cell key of 0.77 (yellow).

If the procedure is applied to all cells of the table, the original frequency count table is modified using the mechanism shown in figure 7.

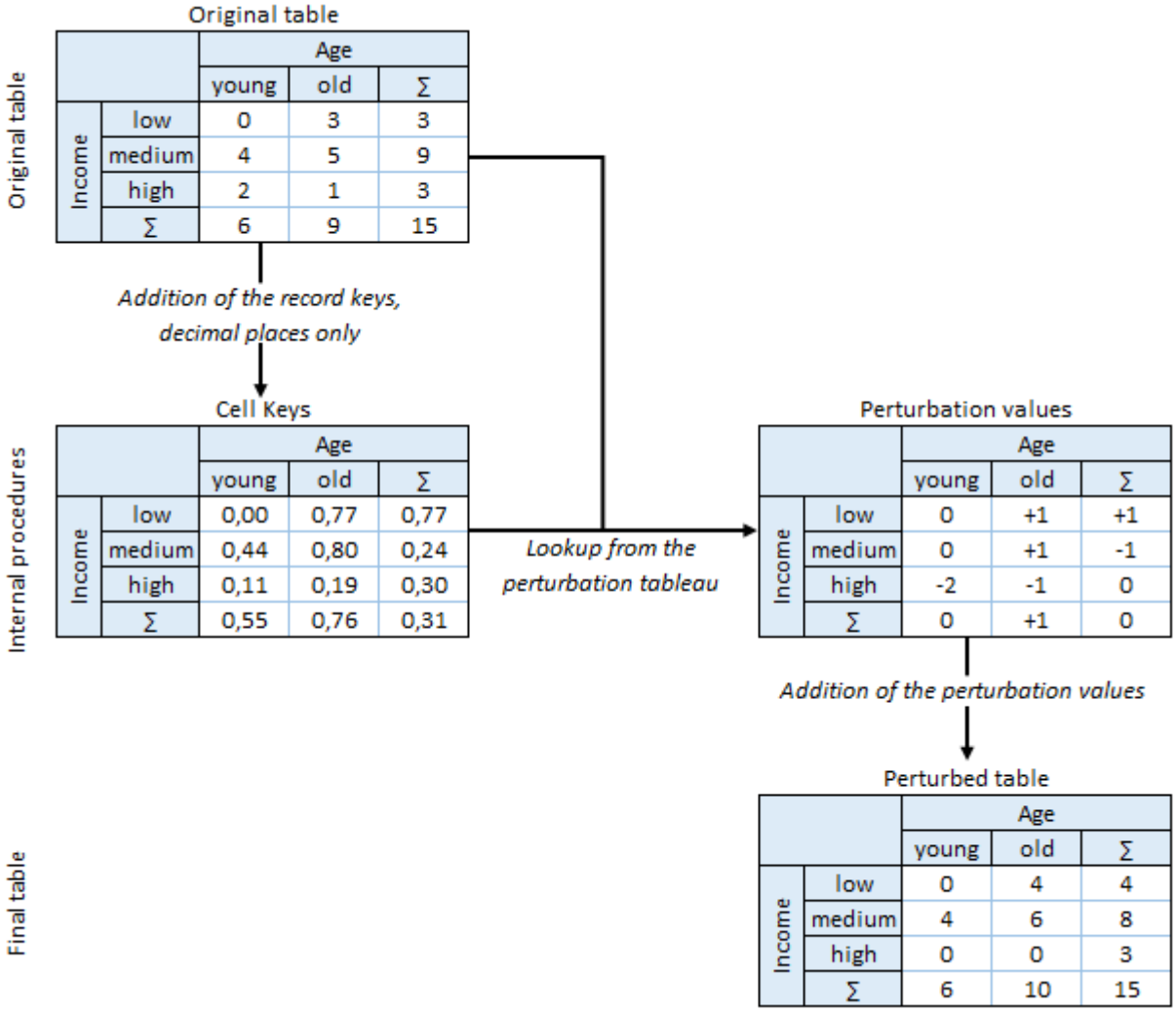

Figure 6: Depiction of the work steps of the cell key method

The perturbed table does not immediately show that it does not contain the original frequency count. At second glance, however, it quickly becomes clear that the inner cells do not generally add up to the marginal totals. This and other effects of the cell key method are presented in another article that appeared as RDC working paper no. 54 (Rothe et al, 2024).

IV. Formal explanation of the cell key method

In its basic form, the cell key method is suitable for ensuring the confidentiality of frequency count tables. An extension of the cell key method for value tables is now also available (Gießing/Tent, 2019), although this has so far only been used in official German statistics for some value variables of the census 2022. Therefore, the formal explanation of the methodology of the cell key method below is limited to its basic form for frequency count tables based on the original publication by Fraser/Wooton (2005).

The basic idea of the cell key method is to create uncertainty among users as to whether a published frequency count corresponds to the original frequency count or if it has been slightly modified. For this purpose, all original frequency counts are overlaid with a perturbation value. If $d_i \in \mathbb{Z}$ is used to denote the perturbation value assigned to the i-th table field with original case number $j_i \in \mathbb{N}_0$, this leads to the modified frequency count k_i according to

$$k_i = j_i + d_i.$$

As with all data-altering procedures, one of the key features of the cell key method is that all original frequency counts can be altered regardless of their criticality or risk of detection.

1. Requirements for the perturbation values

The following properties apply to the perturbation values d_i for all i:

1.1. Unbiasedness: $E[d_i] = 0$, i.e. no systematic distortion of the overall result results from the perturbation of the table fields.

1.2. Non-negativity: $d_i \geq -j_i$, i.e. it is ensured for each table cell that the perturbation does not result in a negative frequency count $\forall i: k_i \geq 0$.

1.3. Integral values: $d_i \in \mathbb{Z}$, i.e. the perturbation values must be integers so that integerness is also guaranteed for the changed frequency counts. Conditions (b) and (c) therefore result in $k_i \in \mathbb{N}_0$.

2. Parameters of the cell key method

The application of the cell key method requires the definition of parameters and conditions that determine the stochastic model used to assign the perturbation values. A distinction is made between mandatory parameters and additional optional restrictions. The specific design of the parameterisation is always carried out by the specialist department responsible for the statistics in question.

Specifically, two parameters must be defined:

- Maximum deviation $D \in \mathbb{N}$ between the original and modified frequency counts, so that $\forall i: |d_i| \leq D$.
- Variance V of all deviations from the original frequency counts: The variance equals $V := Var[d_i] = E[d_i^2] = \sum_{i=-D}^{D}(p_i \cdot d_i^2)$.
  By the determination of D the value range of V implicitly results in $(0; D^2]$.

The definition of the maximum deviation between the original and modified frequency counts serves in particular to control the loss of information induced by the data modification and thus the quality of the output that is kept confidential. The variance of the deviations from the original frequency counts is used

to calibrate the uncertainty that the data modifications create for the users. A low variance implies that a high proportion of the original frequency counts are changed only slightly or not at all (i.e. $d_i = 0$). The larger V is selected, the greater the probability that (considering the specified maximum deviation D) high deviations will occur between the original and modified frequency counts.

3. Transition matrix

The maximum deviation from the original frequency count and the variance of the perturbation values largely determine the probabilities with which a specific perturbation value is assigned to an original frequency count. Below, $p_{jk} \in [0;1]$ denotes the probability that the original frequency count j is modified to the frequency count k (or receives the perturbation value $d = |j-k|$). As this probability only depends on the value of the original frequency count but not on a specific table cell i, the index i can here be omitted. The (conditional) transition probabilities are collected for all combinations of possible original frequency counts and modified frequency counts in the so-called transition matrix $\mathbf{T}$:

$$\mathbf{T} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0k} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1k} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots & p_{2k} & \cdots \\ \vdots & \vdots & \vdots & \ddots & & \\ p_{j0} & p_{j1} & p_{j2} & & p_{jk} & \\ \vdots & \vdots & \vdots & & & \ddots \end{pmatrix}$$

$\mathbf{T}$ is a square matrix that has the following properties:

- $\mathbf{T}$ contains the so-called probabilities p_{jj} with $j = k$ on the main diagonal, i.e. the probabilities that an original frequency count j remains unchanged.
- Row-wise, T contains the conditional probability distributions for the perturbations of the original frequency count j, that is $\forall j: \sum_k p_{jk} = 1$.
- Due to the mandatory determination of a maximum deviation DD from the respective original case numbers, $p_{jk} = 0$ for all $k \notin \{j-D, \dots, j+D\}$.
- For small frequency counts $j < D$, the interval actually defined by D for the possible frequency counts after data modification $[j-D, \dots, j+D]$ cannot be fully exploited in order to meet the condition of non-negativity of the modified frequency count. The actual interval for the possible frequency count after changing the original frequency count j is therefore $\Pi = [\max\{j-D;0\}, \dots, j+D]$. Negative perturbation values are therefore only possible to a limited extent when the original frequency count is small (while at the same time ensuring that the modified frequency count is undistorted) (Höhne/Höninger, 2018).

4. Further optional restrictions for the calibration of the transition matrix

In addition to the above-mentioned mandatory parameters D and V, further additional conditions can optionally be formulated which influence the shape of the transition matrix. Possible restrictions are:

- Determination of a probability of retention: It can be determined that a certain proportion P_V of all original frequency counts shall remain unmodified. In this case, $P_V \cdot 100\%$ of all frequency counts in the confidential tables correspond to their respective original values.
- Original zeros shall remain unmodified: $p_{00} = 1$ respectively $\forall k > 0: p_{0k} = 0$. Characteristic combinations that do not exist in reality remain excluded even after using the cell key method.
- Exclusion of small frequency counts after data modification: The output of very small modified frequency counts can optionally be excluded up to and including a threshold value $m > 0$. $\forall j: p_{jk} = 0$ then applies to the corresponding modified frequency counts k = 1, ..., m. The corresponding column vectors for the excluded modified frequency counts in $\mathbf{T}$ then correspond to the zero vector. Specialised departments often exclude 1 as a modified frequency count ($\forall j: p_{j1} = 0$).
- Ensuring a symmetrical distribution: $\forall d \in \{-D, \dots, D\}: p_{j(j-d)} = p_{j(j+d)}$, i.e. modifications of an original frequency count by the perturbation values $-d$ and d have the identical probability. This results in a symmetrical distribution of the transition probabilities for each frequency count j around the corresponding probability of retention p_{jj}. Please note that the symmetry property of the perturbation distribution can only be guaranteed for those original frequency counts j for which $j \geq D$ applies. If small modified frequency counts up to and including the threshold value m are also excluded, this condition expands to $j > D + m$.

5. Calculation of the transition matrix

On the basis of the mandatory parameters and considering the further optional restrictions on the design of the transition matrix, the specific (conditional) distribution of the perturbed values must be calculated for each original frequency count. For this, Marley/Leaver (2011) propose to maximise the entropy of the (conditional) perturbation distributions. Entropy represents a measure of dispersion of a probability distribution, the maximisation of which can be interpreted in this context as minimizing the risk of detection (Marley/Leaver, 2011). If Π denotes the set of possible modified frequency counts for a given original frequency count, then the maximisation problem for calculating the perturbation distribution for the original frequency count j is

$$\max_{\mathbf{p}_{jk}} \left\{ -\sum_{k \in \Pi} \left(p_{jk} \cdot \log_2 p_{jk} \right) \right\}.$$

Including all formulated secondary conditions, this results in a non-linear system of equations. Gießing (2016) presents an approach to solve the optimisation problem using a Lagrangian approach, which is used in the R package ptable (Enderle, 2023) to create CKM-transition matrices. Further information on maximising entropy can be found in Enderle/Vollmar (2019).

For all original frequency count $j > m + D$, the perturbation distribution is structurally identical, i.e. $\forall l \geq 0: p_{jk} = p_{(j+l)(k+l)}$. The resulting perturbation distribution for $j = m + D + 1$ – shifted to the right by $a \in \mathbb{N}$ columns in $\mathbf{T}$ – also applies to all larger original frequency counts $j + a$.

6. Record keys and cell keys

Once the transition matrix $\mathbf{T}$ has been specified, concrete perturbation values are assigned to the original frequency counts. This is done based on the available data.

For this, a fixed random number $r_s \sim U[0; 1]$ is first assigned to each statistical unit $s = 1, \ldots, S$ in the microdata set. The random numbers are drawn from a continuous uniform distribution and are called record keys. The record key assigned to a statistical unit remains identical for all evaluations that are created for the affected statistics, at least for the current reporting period.

At the table cell level, the record keys are used to create a "key number" for each individual table cell, which is used to assign the perturbation value for the relevant table cell. This so-called cell key c_i is calculated for table cell i according to

$$c_i = \sum_{s \in I} r_s - \left\lfloor \sum_{s \in I} r_s \right\rfloor \sim U[0; 1)\,,$$

where the set I contains all feature carriers s that contribute to table cell i. To calculate the cell key for table field ii, the sum of the record keys of all feature carriers contributing to ii is reduced by their next smaller integer amount (back term with lower Gaussian bracket). This calculation operation is necessary because uniformly distributed random variables lose their distribution properties when summed (front term) and the correction restores the property of a continuous uniform distribution for the cell keys. As a result, after this step, each table field to be overlaid has a specific cell key with a value $c_i \in [0; 1)$, which depends on the specific feature carriers or their record keys that contribute to the table field. The procedure presented here ensures cross-table consistency of the changed case numbers, since logically identical table fields (i.e. with identical contributing feature carriers) always receive the same cell key.

7. Assignment of the perturbation values

In a final step, the generated cell keys are used to decide which specific overlay value is assigned to a table field based on the specified transition matrix.

For this purpose, the transition matrix is converted into a so-called overlay tableau by gradually aggregating the overlay distribution for each original case number j. For this purpose, the distribution function F_j is defined according to

$$F_j(d) \coloneqq \sum_{b \leq d;\, d \in \Lambda} p_{b|j}\,,$$

where the set $\Lambda_j = \{d_1, d_2, \dots, d_{n-1}, d_n\}$ contains all perturbation values that are suitable for j and $p_{d|j}$ denotes the probability of a perturbation of the original frequency count j with the perturbation value d. The value of the individual transition probabilities is mapped over the width of the intervals $[0; F_j(d_1)]$, $(F_j(d_1); F_j(d_2)]$, ..., $(F_j(d_{n-1}); F_j(d_n)]$, where; $F_j(d_n) = 1$. Thus, the union of all intervals covers the interval $[0; 1]$ completely and without overlaps.

Since the cell keys are equally distributed on the interval $[0; 1)$, the perturbation value can be assigned based on a simple comparison between the cell key of a table cell and the distribution function of the perturbation values. The perturbation value d_i is assigned to the table cell i with cell key c_i using the following mechanism:

$$d_i = d_r | d_r \in \Lambda : c_i \in (F_j(d_{r-1}); F_j(d_r)]$$

The perturbation value d_r is therefore selected as the perturbation value for the table cell i if the cell key of the table cell falls within the subinterval which is spanned by the distribution function values of d_{r-1} and d_r (and whose width corresponds to the probability of an overlay with d_r). This assignment mechanism is applied to all table cell that are to be perturbed. The uniform distribution property of the cell keys ensures that across all tables the proportion of table cells perturbed with a specific perturbation value corresponds to the respective transition probability $p_{d|j} \cdot 100\%$.

V. Conclusion

With the cell key method, a new secrecy procedure is finding its way into official statistics and thus also into Research Data Centres. The cell key method is a method based on post-tabular stochastic perturbation. The fixed assignment of a record key to each statistical unit ensures that modifications of the original frequency count occur consistently and can be replicated across different program runs. However, this comes at the expense of the additivity of the result tables. The RDC working paper 54

“The cell key method at the Research Data Centres of the Statistical Offices of the Federation and the Federal States. Part 2: Effects of the new confidentiality procedure” describes in detail what effects the application of the cell key method has on frequency tables and key figures (Rothe and others, 2024).

VI. References

- Enderle, Tobias/Vollmar, Meike. Geheimhaltung in der Hochschulstatistik. In: WISTA Wirtschaft und Statistik. Issue 6/2019, page 87 ff.
- Enderle, Tobias. ptable: Generation of Perturbation Tables for the Cell-Key Method. R package version 1.0.0. 2023. [Access on 30.04.2024]. Available at: CRAN.R-project.org
- Fraser, Bruce/Wooton, Janice. A proposed method for confidentialising tabular output to protect against differencing. Work session on statistical data confidentiality. Supporting paper. Genf 2005. [Access on 30.04.2024]. Available at: unece.org
- Gesetz über die Statistik für Bundeszwecke (Federal Statistics Act – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 14 des Gesetzes vom 8. Mai 2024 (BGBl. I Nr. 152) geändert worden ist. [Access on 30.04.2024]. Available at www.gesetze-im-internet.de
- Giessing, Sarah/Tent, Reinhard. Concepts for generalising tools implementing the Cell-Key-Method to the case of continuous variables. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Den Haag 2019. [Access on 30.04.2024]. Available at: unece.org
- Giessing, Sarah. Computational Issues in the Design of Transition Probabilities and Disclosure Risk Estimation for Additive Noise. In: Domingo-Ferrer, Josep/ Peji-Bach, Mirjana (Herausgeber). Privacy in Statistical Databases. LNCS (Lecture Notes in Computer Science). 2016. Issue 9867, page 237 ff. DOI: 10.1007/978-3-319-45381-1_18
- Höhne, Jörg. Das Geheimhaltungsverfahren SAFE. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Issue 2/2015, page 16 ff. [Access on 30.04.2024]. Available at: www.statistischebibliothek.de
- Höhne, Jörg/Höninger, Julia. Die Cell-Key-Methode – ein Geheimhaltungsverfahren. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Issue 3+4/2018, page 14 ff. [Access on 30.04.2024]. Available at: www.statistischebibliothek.de
- Kleber, Birgit/Gießing, Sarah. Geheimhaltung beim Zensus 2021. In: Methoden – Verfahren – Entwicklungen. Nachrichten aus dem Statistischen Bundesamt. Issue 2/2018, page 3 ff. [Access on 30.04.2024]. Available at: www.destatis.de

- Marley, Jennifer K./Leaver, Victoria L. A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis. In: Proceedings of 58th World Statistical Congress. 2011. [Access on 30.04.2024]. Available at: 2011.isiproceedings.org
- Rohde, Johannes/Seifert, Christiane/Gießing, Sarah/Setzer, Stefanie (in cooperation with Breitenfeld, Jörg/Brings, Stefan/Höhne, Jörg/Höninger, Julia/Rothe, Patrick/Schedding-Kleis, Ulrike). Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrens. Version 1.1 vom 23.04.2021. Internal document of the Federal Statistical Office and the Statistical Offices of the Federal States.Rothe, Patrick/Güttgemanns, Volker/Rohde, Johannes/Setzer, Stefanie. Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens. In: WISTA Wirtschaft und Statistik. Issue 3/2024, page 45 ff. English translation available as RDC working paper no. 54 at www.forschungsdatenzentrum.de

Up to now, the following RDC working papers were published:

- *Arbeitspapier Nr. 52:* Möglichkeiten zur Verknüpfung mehrerer Mikrozensus-Jahre zu einem Mikrozensus-Panel (2012–2015), S. Brocker, H. Mühlenfeld, November 2020
- *Arbeitspapier Nr. 51:* Record-Linkage bei fehlenden Betriebsnummern im Produzierenden Gewerbe, L. Möll, Juli 2020
- *Arbeitspapier Nr. 50:* Statistische Geheimhaltung - Der Schutz vertraulicher Daten in der amtlichen Statistik, P. Rothe, Februar 2019
- *Arbeitspapier Nr. 49:* Formal, faktisch oder absolut nachgefragt? Die Auswirkungen der Entgeltumstellung auf die Entwicklung der Nachfrage in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, R. Voshage, U. Drafz, H. Habla, M. Klumpe, C. Meisdrock, K. Nowak, S. Raab, A. Richter, M. Rößner, K. Schmidtke, November 2015
- *Arbeitspapier Nr. 48:* Bereitstellung der Daten des Zensus 2011 über die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, S. Raab, C. Meisdrock, Oktober 2015
- *Arbeitspapier Nr. 47:* FiRe – Ein Schritt zur Teilautomatisierung der Geheimhaltungsprüfung, J. Pohlisch, J. Höninger, R., Voshage, Dezember 2014
- *Arbeitspapier Nr. 46*: Entwicklung des Mikrosimulationsmodells EITDsim, G. Struch, April 2013
- *Arbeitspapier Nr. 45*: Ein Mikrosimulationsmodell zur Berechnung der einkommensteuerlichen Bemessungsgrundlage und der Steuerzahlung auf Basis des Taxpayer-Panels, T. P., Schmidt, April 2013
- *Arbeitspapier Nr. 44*: Das SAS-Makro newvar. Entwicklung und Anwendung eines Hilfsinstruments zur effizienten Erstellung neuer Variablen in der DRG-Statistik, T. Hochgürtel, T. Lösch, März 2012
- *Arbeitspapier Nr. 43*: Average wage, qualification of the workforce and export performance in German enterprises: Evidence from KombiFiD data, J. Wagner, Januar 2012
- *Arbeitspapier Nr. 42*: The Quality of the KombiFiD-Sample of Enterprises from Manufacturing Industries: Evidence from a Replication Study, J. Wagner, Dezember 2011
- *Arbeitspapier Nr. 41*: How to define an enterprise and assign trade declarations to the right one: Exploration of German traders' micro transaction data, C. Stirböck, August 2011
- *Arbeitspapier Nr. 40*: Definition von nutzerseitigen Kriterien für Datenstrukturfiles, J. Höninger/M. Rosemann/R. Voshage, Juli 2011
- *Arbeitspapier Nr. 39:* Improvement of data access – The long way to remote data access in Germany, M. Brandt/M. Zwick, Juni 2011
- *Arbeitspapier Nr. 38:* Decentralised Access to Confidential Microdata in Europe, M. Brandt/P. Eilsberger/M. Zwick, Juni 2011
- *Arbeitspapier Nr. 37:* Masking Micro Data with Stochastic Noise, J. Höhne/J. Höninger, Mai 2011
- *Arbeitspapier Nr. 36:* Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, A. Vogel, April 2011
- *Arbeitspapier Nr. 35:* Temporary agency work and firm performance, S. Nielen/A. Schiersch, April 2011
- *Arbeitspapier Nr. 34:* Harmonisation of statistical confidentiality in the Federal Republic of Germany, M. Brandt/A. Crößmann/C. Gürke, März 2011
- *Arbeitspapier Nr. 33:* Remote Access. Eine Welt ohne Mikrodaten ??, G. Ronning/P. Bleninger/ J. Drechsler/C. Gürke, Februar 2011

- *Arbeitspapier Nr. 32:* Compiling a Harmonized Database from Germany´s 1978 to 2003 Sample Surveys of Income and Expenditure. T. Bönke/C. Schröder/C. Werdt, Mai 2010
- *Arbeitspapier Nr. 31:* The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009
- *Arbeitspapier Nr. 30:* Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten – Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/M. Zwick, Januar 2008
- *Arbeitspapier Nr. 29:* Reichtum in Niedersachsen und anderen Bundesländern –Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008
- *Arbeitspapier Nr. 28:* Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009
- *Arbeitspapier Nr. 27:* Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008
- *Arbeitspapier Nr. 26:* Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/K. Kohn/A. C. Lembcke, August 2008
- *Arbeitspapier Nr. 25:* German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008
- *Arbeitspapier Nr. 24:* Zum Einkommensreichtum Älterer in Deutschland – Neue Reichtumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/J. Merz, Februar 2008
- *Arbeitspapier Nr. 23:* Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/D. Oberschachtsiek/R. Pohl, November 2007
- *Arbeitspapier Nr. 22:* Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung – Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007
- *Arbeitspapier Nr. 21:* KombiFid – Kombinierte Firmendaten für Deutschland, S. Bender/J. Wagner/ M. Zwick, November 2007
- *Arbeitspapier Nr. 20:* Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007
- *Arbeitspapier Nr. 18:* Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007
- *Arbeitspapier Nr. 17:* Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/R. Lenz, Dezember 2006
- *Arbeitspapier Nr. 16:* Die europäische Union – Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006
- *Arbeitspapier Nr. 15:* Keeping in Touch – A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/L. Osberg, November 2006
- *Arbeitspapier Nr. 14:* Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/C. Gräb/S. Kriete-Dodds, November 2006
- *Arbeitspapier Nr. 13:* Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006
- *Arbeitspapier Nr. 12:* Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006
- *Arbeitspapier Nr. 11:* Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

- *Arbeitspapier Nr. 10:* Konstruktion und Bewertung eines ökonomischen Einkommens aus der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/F. Neher/C. Schröder, August 2006
- *Arbeitspapier Nr. 9:* Anonymising business micro data – results of a German project, R. Lenz/M. Rosemann/D. Vorgrimler/R. Sturm, Juni 2006
- *Arbeitspapier Nr. 8:* Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/H.-P. Hafner/D. Schmidt, Juni 2006
- *Arbeitspapier Nr. 7:* A standard for the release of microdata, R. Lenz/ D. Vorgrimler/M. Scheffler, Juni 2006
- *Arbeitspapier Nr. 6:* Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006
- *Arbeitspapier Nr. 5:* De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/M. Zwick, Oktober 2005
- *Arbeitspapier Nr. 4:* Matching German turnover tax statistics, R. Lenz/D. Vorgrimler, Juni 2005
- *Arbeitspapier Nr. 3:* The research data centres of the Federal Statistical Office and the statistical offices of the Länder,*S*. Zühlke/M. Zwick/S. Scharnhorst/T. Wende, März 2005
- *Arbeitspapier Nr. 2:* Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/M. Zwick, April 2005
- *Arbeitspapier Nr. 1:* Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005