Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR TECHNICAL REPORT 2015-005
NOVEMBER 2015

# Guidelines for Linking Contextual Factors and Survey Data: An Application with Data from the German Family Panel *(pairfam)*

Tom Hensel (hensel@demogr.mpg.de)
Michaela Kreyenfeld (kreyenfeld@demogr.mpg.de)
Rainer Walke (walke@demogr.mpg.de)

For additional material see www.demogr.mpg.de/tr/

This technical report has been approved for release by: Vladimir Shkolnikov (shkolnikov@demogr.mpg.de), Head of the Laboratory of Demographic Data.

# Guidelines for Linking Contextual Factors and Survey Data: An Application with Data from the German Family Panel (*pairfam*)

Tom Hensel, Michaela Kreyenfeld, Rainer Walke

**Abstract.** Geo-coded survey data are a prerequisite for any analysis that examines the impact of contextual factors on individual-level outcomes. The local unit may be districts, municipalities, or blocks of houses. However, regional identifiers are sensitive information and including them in a scientific-use-file of a survey may violate data protection regulations. This Technical Report demonstrates how analyses with geo-coded data from the German Family Panel (*pairfam*) may be conducted. The procedure that we suggest has been tailored for *pairfam*, but it is applicable to other data sets as well.

**Keywords** Data Processing, Data Privacy, Geo-Coded Data, Information Privacy

## 1 Introduction

In Germany, the Federal Statistics Act (*Bundesstatistikgesetz*) regulates the rights and obligations of data providers. This law stipulates that micro-level data may only be made available for scientific usage if the individual respondent cannot be identified, unless a "disproportional" effort is conducted (Rat für Sozial- und Wirtschaftsdaten 2012). This means that sensitive information, such as the respondents' name, date of birth, and detailed place of residence information, must be removed before the data is made available to the research community. Geographically detailed place of residence information is sensitive because it increases the risk that, with the help of a few additional attributes, an individual may be identified in a micro-level data set (Cavoukian and Emam 2011). However, regional information is necessary for certain types of research, such as investigations that seek to unravel the effects of regional contextual conditions on behaviour. Moreover, this includes investigations of the effect of the local population structure on demographic behaviour. For instance, within the context of social science, it is possible to conduct various analyses of how the sex ratio (the ratio of males to female in a population) of specific age groups influences family behavior (e.g. Stauder 2010).

Data providers resolve the problem of providing sensitive regional information in different ways. The German-Socio-Economic Panel (*GSOEP*), for example, allows users in so called "Safe-Centres" at the German Institute for Economic Research to conduct analysis with the regional identifiers of the *GSOEP* (Frick et al. 2010). Regional information of the German Mikrozensus is also accessible for onsite users (Christians 2006). Users of the German Family panel ("Panel Analysis of Intimate Relationships and Family Dynamics", known by the acronym *pairfam*) may travel to the Research Data Centres in Bremen, Chemnitz, Cologne, Munich, or Jena to conduct analysis with geo-coded data. Onsite data analysis is probably one of the most straightforward strategies to enable analysis with regional data in light of the given legal constraints. However, onsite data analysis is costly, both for data providers, who provide the office space, as well as for the guest researchers, who (usually) need to travel to the Research Data Centres.

For the analysis of the *pairfam* data we propose a strategy where no onsite visit is required for the researcher:

- The researcher prepares a macro-level data set that contains the contextual variable of interest (such as the sex ratio by region) and the regional identifier (such as an official municipality key) for all municipalities in Germany. The contextual variable needs to be a categorical variable[1.] This data set is sent to the data provider.
- The data provider (a person at the *pairfam* User Support) merges this information to the *pairfam* data set which contains the regional identifier.
- For data protection reasons, the regional identifier is then replaced by a "cluster variable". This cluster variable indicates if several respondents live in the same regional unit. Afterwards the data with the personal identifier, the categorical contextual variable, and the cluster variable is sent back to the user, who can then merge this information to *pairfam* (for which he or she holds a valid user agreement).

While this procedure does reduce travel expenses, the downside is that the user can only use a categorical variable for his/her regional analysis. He/she has to decide beforehand on the classification of this particular variable. Furthermore, no information about the geographical location of a particular municipality is made available to the researcher. Thus, the researcher cannot conduct any detailed check for possible spacial autocorrelations.

In this Technical Report, we describe the abovementioned strategy in more detail. In order to demonstrate our procedure, we use data from the first wave of the German Family Panel, to which we merge municipality information of the regional sex ratio. The paper is structured as follows: The next section (section 2) describes the German Family Panel (*pairfam*) in more detail, as well as the structure of the regional identifier. Section 3 explains our three step procedure to conduct regional analysis with *pairfam*. Section 4 provides some exemplary investigations, and section 5 concludes by discussing the pros and cons of our procedure.

[1] The number and size of categories must ensure that no category contains only a single municipality.

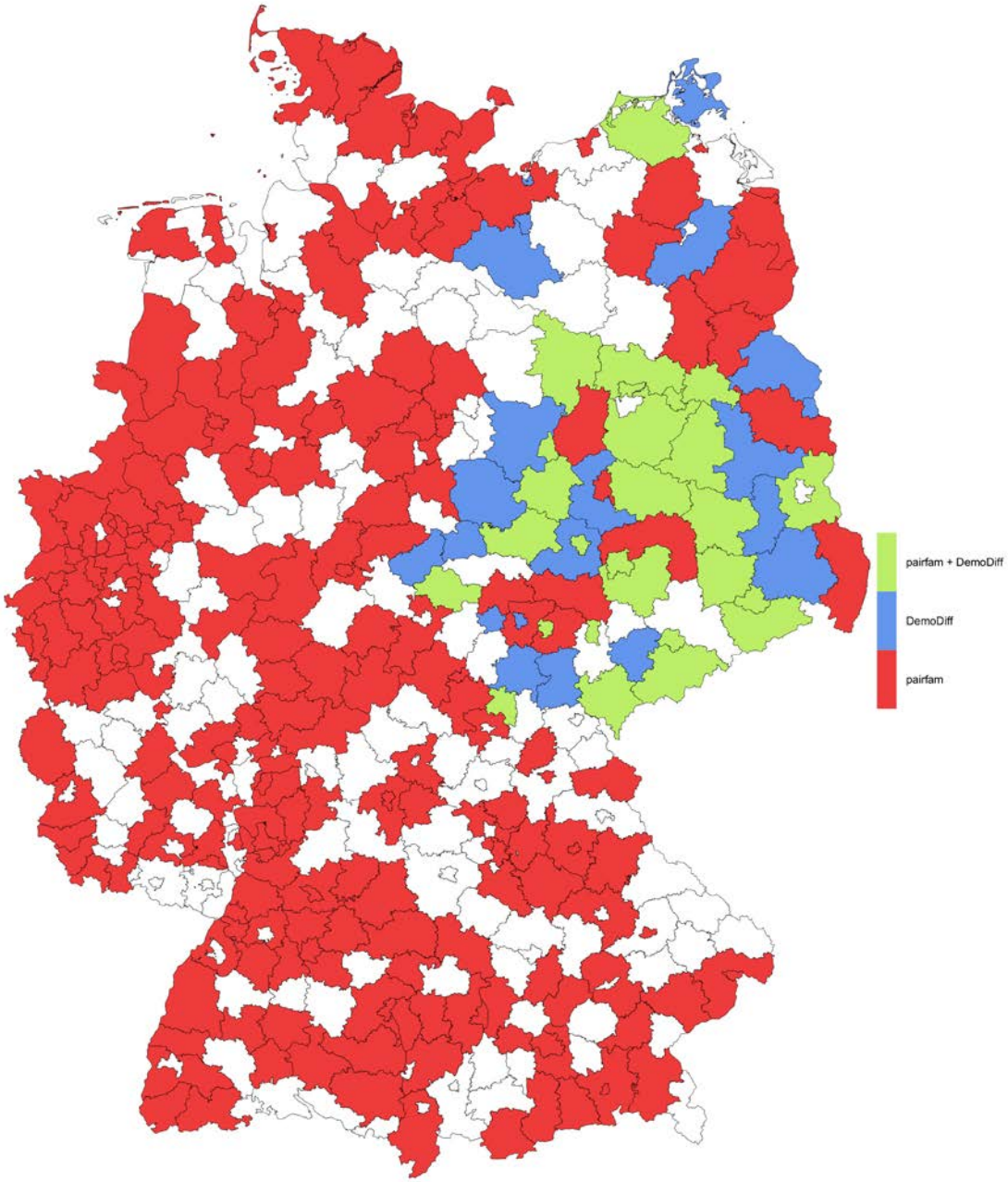
# 2 Contextual and Regional Information in pairfam and DemoDiff

## 2.1 Sampling Procedure

The German Family Panel (*pairfam*) was launched in 2008/09. In the first wave of this annual panel study, 12,402 persons of three birth cohorts (1971-73, 1981-83 and 1991-93) were interviewed. In 2009, The German Family Panel was supplemented by the subsample *DemoDiff* (Kreyenfeld et al. 2012), which included 1,489 respondents of the cohorts 1971-73 and 1981-83 in the first round. It was drawn to oversample the former eastern German states and, thus, included respondents who were living in East Germany (excluding West Berlin). The survey agency TNS Infratest conducted the survey and delivered the data to the *pairfam* team, who generated Scientific-Use Files (SUF) from the original data. The SUF files are available via the GESIS-Webportal for users who had signed the user agreement (for details, see www.pairfam.de/en).

The German Family Panel is based on a sample drawn from municipality registers (*Einwohnermeldestichprobe*). This is a common sampling approach in Germany, because there is no national central population register (*zentales Einwohnermelderegister*). Due to the lack of a central register, a two stage procedure is applied. First, all German municipalities (*Gemeinden*) are assigned to a certain strata depending on their population structure in relation to the total population of Germany. Then each selected municipality draws a sample from its registers and delivers the addresses to the survey agency TNS Infratest. The municipalities' samples were selected separately for the first round of the *pairfam* sample that was drawn in 2008/2009 and for the *DemoDiff* sample that was collected one year later in 2009/2010. Figure 1 shows the areas from which the samples were drawn. For data protection reasons, we do not show the sampling points at the municipality level, but only at the district level areas (*Kreise*) of the municipalities. Red denotes that there is as least one (original) *pairfam* sampling community. Blue indicates that at least one *DemoDiff* sampling point exists in this district. Green areas contain both, at least one *DemoDiff* and at least one *pairfam* sampling point. White indicates that no sampling was done.

**Fig. 1** Sample points of the *pairfam* and *DemoDiff* surveys, districts (2008)




Source: Data provided by Brüderl et al. 2015, Shapefile available at BKG (2015), authors' illustration.

## 2.2 Regional Identifiers in the German Family Panel

The scientific use files of the German Family Panel include information on the federal state (*Bundesland*), but no further fine-grained regional level information. However, in principle more detailed regional information is available in the original data set provided by the survey agency (for details, see Schmiedeberg 2015). Among other regional units, a municipality level identifier (*Gemeindeschlüssel*) is available. Hence, the place of residence of a respondent can be determined down to the municipality

level. For the municipality level, the German Statistical Office routinely provides indicators across a broad spectrum, such as population size (by gender), percentage share of employees covered by compulsory social insurance[2], individuals involved in road accidents, child care coverage, etc[3].

The municipality identifier (*Amtlicher Gemeindeschlüssel* or *AGS*) is a number sequence (Arbeitsgruppe Regionale Standards, 2013). It consists of eight digits generated as follows: The first two digits indicate one of the sixteen individual German federal states (*Bundesland*: e.g., 01 stands for Schleswig-Holstein, 16 for Thuringia). The third digit denotes the government district (*Regierungsbezirk*) with the number one (in federal states without such districts, a zero is used instead). The following two numbers, digits four and five, designate the urban area (in a district-free city only) or the concrete district (*Landkreis*). The sixth, seventh, and eighth digits identify the municipality (*Gemeinde*). Table 1 shows an exemplary list of the two first and two last consecutively numbered municipalities in Germany in 2008 as well as the structure of their Official Municipality Key.

**Table 1** Structure of the Official Municipality Key (AGS)

| | **Official Municipality Key (digits)** | | | |
|---|---|---|---|---|
| | **1-2** | **3** | **4-5** | **6-8** |
| | Bundesland | Regierungsbezirk | Landkreis | Gemeinde |
| Flensburg, City of | 01 | 0 | 01 | 000 |
| Kiel (state capital) | 01 | 0 | 02 | 000 |
| … | … | … | … | … |
| Ziegelheim | 16 | 0 | 77 | 055 |
| Saara | 16 | 0 | 77 | 056 |

Source: Statistische Ämter des Bundes und der Länder (2015c).

[2] Statistische Ämter des Bundes und der Länder (2015a).
[3] Statistische Ämter des Bundes und der Länder (2015b).

## 2.3 Three Step Procedure to Conduct Regional Analyses with the German Family Panel

In the following section, we describe how a researcher may conduct a regional analysis with the German Family Panel. The researcher sends a file to the User Support of the German Family Panel that contains the regional code (AGS) and the respective categorized regional information (such as the sex ratio on the municipality level). The User Support of the German Family Panel matches this regional information to the German Family Panel and then sends the researcher a file which contains the categorized regional information with the personal identifier of *pairfam*. The researcher also receives an anonymized cluster variable.

In order to demonstrate this procedure, the following example describes *Alice* the researcher, and *Bob* as a person at the User Support of the German Family Panel. *Alice* wants to investigate how the local sex ratio influences partnership behaviour.

*First Step: Alice Assembles Regional Information*

In a first step, *Alice* needs to obtain information that maps the sex ratio on the regional level. As mentioned above, regional information on the municipality level is freely available from the Regional Database Germany (*Regionaldatenbank*) of the Federal Statistical Office and from the statistical Offices of the Länder.[4] *Alice* uses this source to generate the sex ratio on the municipality level as of December 31, 2008. Table 2 shows the data that Alice has generated: The variable "name" denotes the name of a municipality, the variable "ags" indicates the regional identifier. The variable "sexratio" is a metric variable that Alice generated herself, based on the number of females and males in the municipality. *Alice* needs to group the sex ratio into broader categories. The broad classification assures that an individual cannot be traced. In our example, *Alice* has grouped the sex ratio into 10 categories. The categorical variable is called "sexratio_c". In line with the concept of "De-facto Anonymisierung" (Art. 16 Para. 6 BStaG[5]), *Bob* would allow as many categories as needed so that *Alice* would not be able to trace the individual.

[4] Statistische Ämter des Bundes und der Länder (2015c).
[5] BMJV (2015).

**Table 2** Regional data prepared by the user (*Alice*)

| name | ags | sexratio[6] | sexratio_c |
|---|---|---|---|
| Flensburg, City of | 01001000 | 1.08 | 6 |
| Kiel (state capital) | 01002000 | 0.99 | 3 |
| … | … | … | … |
| Ziegelheim | 16077055 | 1.30 | 9 |
| Saara | 16077056 | 1.11 | 7 |

Source: Statistische Ämter des Bundes und der Länder (2015c).
Note: sexratio_c is the classified sexratio.

*Second Step: Bob Merges Regional Information to pairfam*

*Alice* sends the data (ideally in STATA-format) as depicted in Table 2 to *Bob*. *Bob* merges this data to the first wave of the German Family Panel via the regional identifier. *Bob* keeps the personal identifier ("case"), the regional code ("ags") and the classified sex ratio (see Table 3).

**Table 3** Geo-coded data prepared by the data provider (*Bob*)

| case | ags | sexratio_c |
|---|---|---|
| 1 | 01001000 | 6 |
| 2 | 01001000 | 3 |
| … | … | … |
| 13890 | 16075132 | 9 |
| 13891 | 16075132 | 7 |

Source: Statistische Ämter des Bundes und der Länder (2015c), authors' calculations.

*Third Step: Bob generates a Cluster Variable and sends file back to Alice*

To ensure data protection, the regional units ("ags") cannot be delivered to the user. For this reason, they are replaced by randomized cluster variables, a "cluster_id" that indicates if a person lives in the same regional unit. This data (see Table 4) is sent

[6] Alice used the following 10 categories: [< 0.91]; [0.91, 0.97]; [0.97, 1.00]; [1.00, 1.03]; [1.03, 1.06]; [1.06, 1.09]; [1.09, 1.14]; [1.14, 2.21]; [1.21, 1.32]; [>1.32] (derived using all municipalities sex ratio deciles).

back to the user (*Alice*) who can merge the data to her *pairfam* data set via the person's identification number ("case").

**Table 4** Anonymized data prepared by the data provider (*Bob*)

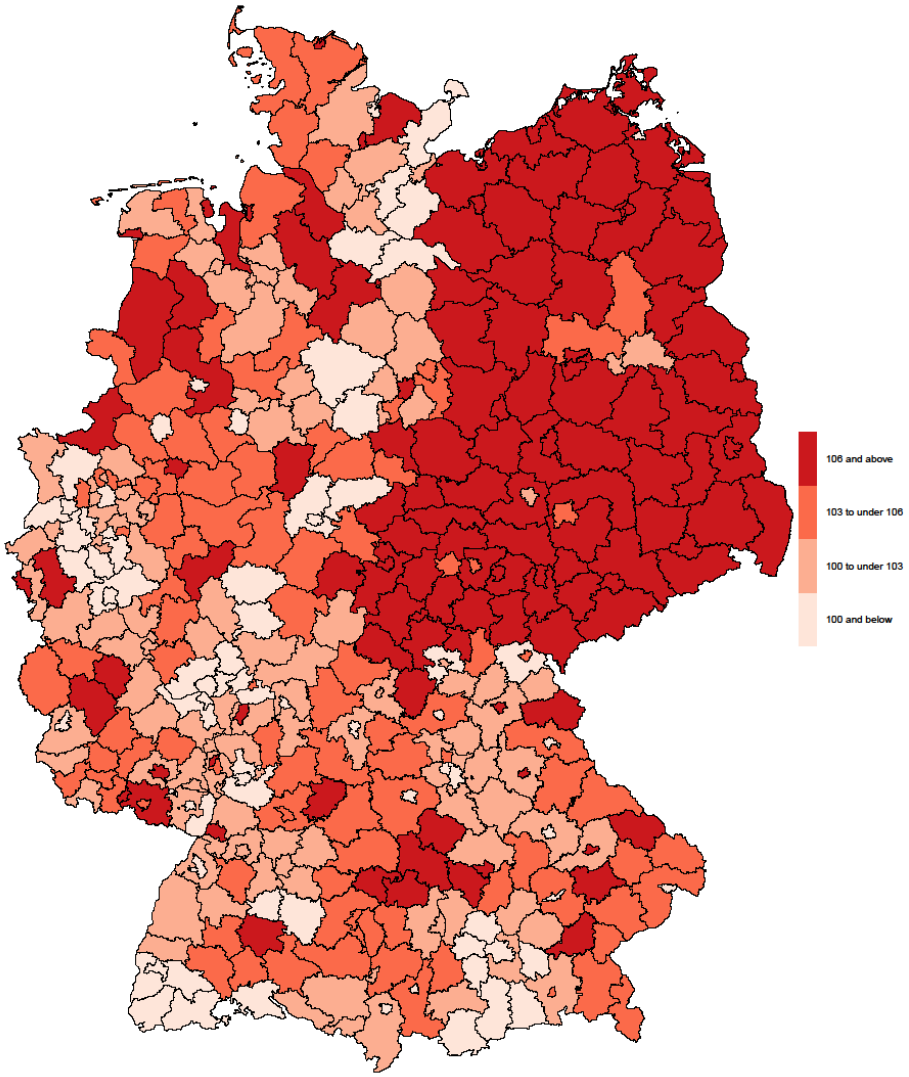
| case | cluster_id | sexratio_c |
|---|---|---|
| 111000 | 335 | 6 |
| 174000 | 284 | 3 |
| … | … | … |
| 920174000 | 148 | 9 |
| 920303000 | 148 | 7 |

Source: Brüderl et al. (2015), authors' calculations.

## 3 Example Analysis

In the following, we use the data that *Alice* has generated to study partnership behaviour. The dependent variable is the probability of having a partner at the time of the interview. The main independent variable is the sex ratio at the municipality level. Because the cohorts are very young, we generated the sex-ratio for the population ages 18-40 only. Control variables in our investigation are: education, age, and region. In order to regard the two-level sampling procedure, we allow a random intercept at the municipality level. Figure 2 plots the regional distribution of the sex ratio at the municipality level. Especially noticeable is the dark red coloured area in eastern Germany which indicates the well-known imbalances in the sex ratios that exists for younger cohorts in this part of the country.

**Fig. 2** Sex Ratio (males to females), ages 18 to 40 in 2008




Source: Data provided by Statistische Ämter des Bundes und der Länder (2015c), Shapefile available at BKG (2015), authors‘ illustration.

Table 5 reports the results from this investigation. We have computed a random intercept logistic regression model and allow a different intercept for every cluster (municipality). The control variables show the expected pattern. With increasing age, people have a higher chance of being partnered. We also find that eastern German men have lower chances of being partnered than do western German men, but there is no difference between eastern and western German women. Having a higher level of education apparently improves the chances on the partner market for both sexes. The odds of having a partner is elevated for highly educated men by 21 per cent and 37 per cent for women (compared to the low educated). The sex ratio at the municipality level was measured by an indicator that gives the proportion of males to females at

ages 18-40. This is a rough indicator of the sex-imbalances of the regional partner market. Nevertheless, we find a strong correlation between the sex ratio and the probability of having a partner for female respondents. If there are more than 106 men to 100 women in a region, the odds of having a partner increase by 33 per cent compared to a situation when the partner market is almost even. For males, we do not find any such association.

**Table 5** Results from a random intercept logistic regression. Dependent variable: Having a partner (1) versus not having a partner (0) at time of interview

| | Males | | Females | |
|---|---|---|---|---|
| | Odds Ratio | St. err. | Odds Ratio | St. err. |
| Age | 1.13*** | 0.01 | 1.08*** | 0.01 |
| Region | | | | |
| East Germany | 0.75*** | 0.07 | 1.11 | 0.11 |
| West Germany | Ref. | | Ref. | |
| Level of education | | | | |
| Currently studying | 0.57*** | 0.06 | 0.42*** | 0.05 |
| Low | Ref. | | Ref. | |
| Medium | 1.16* | 0.10 | 1.37*** | 0.13 |
| High | 1.21** | 0.10 | 1.37*** | 0.13 |
| Other | 0.71 | 0.25 | 0.58 | 0.19 |
| Sex ratio (male to female ratio) | | | | |
| below 1.00 | 1.02 | 0.11 | 1.10 | 0.12 |
| 1.00 – 1.03 | Ref. | | Ref. | |
| 1.03 – 1.06 | 1.08 | 0.11 | 1.34*** | 0.15 |
| 1.06 and larger | 1.00 | 0.11 | 1.33** | 0.16 |
| Constant | 0.05*** | 0.01 | 0.25*** | 0.05 |
| cluster_id (variance) | 0.10 | 0.03 | 0.16 | 0.04 |

Note: Robust standard errors. Low education is “Hauptschule” or less, medium eduction is “Realschule” and “POS 10. Klasse”, high education is “Hochschulreife”.
* p<0.1; ** p<0.5; *** p<0.01

## 4 Outlook

In this Technical Report, we demonstrated a potential strategy to access data from the German Family Panel for regional analysis. Geographically detailed data on the place

of residence are not available in the scientific-use files of the German Family Panel. The strategy that we propose in this document has two essential advantages. First, the procedure ensures universal application. Almost any geographically detailed contextual data set can be linked with research data in compliance with data protection rules using the described concept. Moreover, researchers do not have to travel to a Research Data Center, but can do their analyses from an office computer. Both factors will enhance the user friendliness and the usability in handling research data. However, there are also disadvantages. First, with our procedure the researcher is confined to using a categorical contextual variable for the empirical analysis. Related to this, one must decide on the classification of this particular variable in advance. Second, the data provided do not enable the detection of possible spacial autocorrelations. While the procedure reduces the logistic effort to conduct regional analysis, it still requires a knowledgeable person at the Research Data Center to merge the regional data and provide them to the user. Nevertheless, this procedure should incur lower costs for the researcher and the Research Data Center than an onsite visit of the researcher.

## 5 Acknowledgment

We would like to thank our colleague Sebastian Klüsener (Max Planck Institute for Demographic Research) who made significant contributions and Rüdiger Lenke (Ludwig Maximilian University of Munich / pairfam) for his many helpful comments. All remaining errors are ours. For language editing, we are grateful to Renée Lüskow.

**References**

Arbeitsgruppe Regionale Standards. (2013). GESIS-Schriftenreihe 12. Regionale Standards. http://www.ssoar.info/ssoar/handle/document/34820. Accessed 01. October 2015.

BKG (2015). Bundesamt für Kartographie und Geodäsie. Dienstleistungszentrum. Open Data Karten. Verwaltungsgebiete 1:250.000. http://www.geodatenzentrum.de/auftrag1/archiv/vektor/vg250_ebenen/2008/vg250_2008-12-31.lamgw.shape.ebenen.zip. Accessed 01. October 2015

BMJV (2015). Gesetz über die Statistik für Bundeszwecke. http://www.gesetze-im-internet.de/bstatg_1987/index.html. Accessed 01. October 2015.

Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer F.J., Walper, S., Alt, P., Buhr, P., Castiglioni, L., Finn, C., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., Peter, T., Salzburger, V., Schmiedeberg, C., Schubach, E., Schütze, P., Schumann, N., Thönnissen, C., Wilhelm, B. (2015). The German Family Panel (pairfam). GESIS Data Archive, Cologne. ZA5678 Data file Version 6.0.0, doi:10.4232/pairfam.5678.6.0.0. Accessed 01. October 2015.

Cavoukian, A., Emam, K.E. (2011). Dispelling the myths surrounding de-identification. Anonymization remains a string tool for protecting privacy. Information and Privacy Commissioner (Ontario, Canada). https://www.ipc.on.ca/images/Resources/anonymization.pdf, Accessed 30. September 2015.

Christians, H. (2006). Möglichkeiten kleinräumiger Analysen auf Basis des Mikrozensus. In Forschungsdatenzentrum der Statistischen Landesämter (Ed.), *Amtliche Mikrodaten für die Sozial- und Wirtschaftswissenschaften. Beiträge zu den Nutzerkonferenzen des FDZ der Statistischen Landesämter 2005* (pp. 81-91). Düsseldorf: Forschungsdatenzentrum der Statistischen Landesämter.

Frick, J.R., Goebel, J., Haas, H., Krause, P., Sieber, I., Engelmann, M. (2010). Verfahren für den Datenschutz beim Zugang zu den SOEP-Daten innerhalb und außerhalb des DIW Berlin. https://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.347090.de/soep_datenschutzverfahren.pdf. Accessed 28. September 2015.

Kreyenfeld, M., Huinink, J., Trappe, H., Walke, R. (2012). DemoDiff: A Dataset for the Study of Family Change in Eastern (and Western) Germany. *Schmollers Jahrbuch*, 132(4), 653-660.

Rat für Sozial- und Wirtschaftsdaten (Ed.). (2012). *Georeferenzierung von Daten Situation und Zukunft der Geodatenlandschaft in Deutschland*. Berlin: SCIVERO Verlag.

Schmiedeberg, C. (2015). Regional Data in the German Family Panel (pairfam). pairfam Technical Paper No. 07.

Statistische Ämter des Bundes und der Länder (2015a). Regionaldatenbank Deutschland. Sozialversicherungspflichtig Beschäftigte am Wohnort nach Geschlecht und Nationalität - Stichtag 30.06. - regionale Tiefe: Gemeinden, Samt-/Verbandsgemeinden. Tabelle 254-13-5. https://www.regionalstatistik.de/link/tabelleErgebnis/254-13-5?type=dataset. Accessed 01. October 2015.

Statistische Ämter des Bundes und der Länder (2015b). Regionaldatenbank Deutschland. Straßenverkehrsunfälle, verunglückte Personen regionale Tiefe: Gemeinden, Samt-/Verbandsgemeinden. Tabelle 302-11-5. https://www.regionalstatistik.de/link/tabelleErgebnis/302-11-5?type=dataset. Accessed 01. October 2015.

Statistische Ämter des Bundes und der Länder (2015c). Regionaldatenbank Deutschland. Bevölkerungsstand: Bevölkerung nach Geschlecht und Altersgruppen (17) - Stichtag 31.12. - regionale Ebenen. Tabelle 173-21-5-B. https://www.regionalstatistik.de/link/tabelleErgebnis/173-21-5-B?type=dataset. Accessed 01. October 2015.

Stauder, J. (2011). Regional Ungleichheit auf dem Partnermarkt? Die makrostrukturellen Rahmenbedingungen der Partnerwahl in regionaler Perspektive. *Soziale Welt*, 62, 41-69.