Data Observer

Philipp Breidenbach* and Lea Eilers

RWI-GEO-GRID: Socio-economic data on grid level

https://doi.org/10.1515/jbnst-2017-0171

1 Small-Scale Regional Data in Germany

Quantitative research in economics and related social sciences has increasingly made use of small-scale regional data, allowing for the analyses of neighborhood effects and controlling for local conditions. In Germany, existing regional data from statistical offices mostly do not fulfill the specific demands for research on small areas. Data are commonly not provided below the level of municipalities (*Gemeinden*) or counties (*Kreise*). Hence these data cannot be used for research on inner-municipality or -county developments. Furthermore, data based on administrative boundaries come along with major shortcomings: (1) the regional units are very unequal in their size, and (2) territorial reforms (*Gebietsstandsreformen*), which were common over Germany within the last decades, impede comparisons over time.[1]

The RWI-GEO-GRID provided by the FDZ Ruhr am RWI, contains data which can overcome the above-mentioned shortcomings. The dataset is based on a grid which is uniformly defined by 1×1 kilometer raster cells. In contrast to administrative delineation, these grids are time-consistent and equally spread across the entire territory of Germany. The grid level is defined according to the EU directive standardized European projection system INSPIRE (Infrastructure for Spatial Information in Europe). INSPIRE ensures that different data on the same projection can be merged to each

1 Since the German reunification, the number of counties (including county-level cities) has decreased from 543 in 1990 to 401 in 2017, whereas the number of municipalities has decreased form 16,128 to less than 12,000 in the same period. These territorial changes mostly affect regions in East Germany.

***Corresponding author: Philipp Breidenbach,** RWI – Leibniz-Institute for Economic Research, Essen, German, E-mail: breidenbach@rwi-essen.de
Lea Eilers, RWI – Leibniz-Institute for Economic Research, Essen, German, E-mail: lea.eilers@rwi-essen.de

other.[2] In total, there are around 362,000 1-km^2 grid cells for Germany. Residential or commercial property are located in more than 218,550 grid cells, and for those socio-economic data are provided in the RWI-GEO-GRID dataset.

To the best of our knowledge, the RWI-GEO-GRID is unique in its composition of socio-economic data and spatial resolution for Germany. These data allow research insights into inner-city or -county distributions which cannot be contained from higher aggregated data.

As an example, Figure 1 visualizes the population distribution in Hamburg and its surrounding counties, and it highlights two main advantages. Firstly, a clear north-south divide is observable in the population distribution in Hamburg. The north is characterized by a high share of inhabitants, while the south is characterized by the industrial zone of the harbor leading to a lower share of inhabitants. Compared to administrative data, the grid-level data clearly present the inner-city distribution, where administrative data would only be able

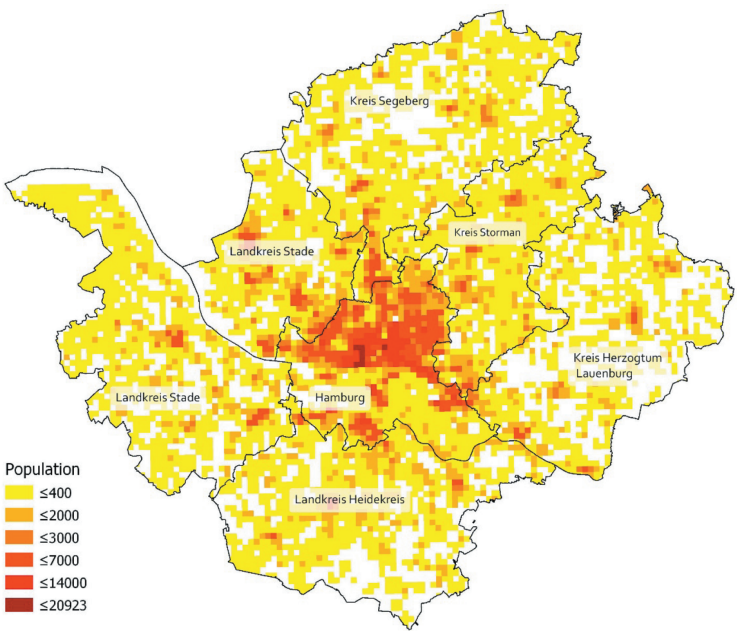

Figure 1: Population on grid level city of Hamburg and surrounding counties in 2015.

2 INSPIRE allows the use of differently sized grids. The edge length of one kilometer is one option.

to show the average number of inhabitants. Secondly, administrative borders, such as the border of Hamburg, do not reflect the borders of residential estates, the so-called exurbs, like Norderstedt in the north of Hamburg. The extent of such exurbs, which can only be guessed in administrative data, can easily be identified using small-scale grid-level data.

2 Data Source and Anonymization

The data is collected by the commercial micro- and geo-marketing provider *Micromarketing-Systeme und Consult GmbH (microm)*. Its main business is target marketing, the analysis of local market conditions, and multi-channel marketing. To this end, microm uses more than one billion individual data points for the aggregation of their dataset. These stem from various data sources.[3] The data points are available for all 40.9 million households in Germany, while the final data product contains information on approximately 20 million houses (microm 2016).

For data privacy reasons, houses within a residential environment are summed up to a virtual micro-geographic segment (so-called micro-cells) which on average comprises eight, and five households at the minimum. Houses with at least five households become a distinct micro-cell, while houses with less than five households are combined with similar houses in the same street. Combined houses are as close as possible in spatial terms. Hence, the derived data do not contain information on the individual level, since they are summed up to the micro-cells.

3 Data Description

For each grid cell, the dataset provides an extensive range of household, demographic, mobility and development information. The data are available for the years 2005 and 2009–2015. Future years are intended to be appended by

3 Most of the data are collected from companies acting in data intensive environments such as Creditreform, CEG Consumer Reporting, bedirect, AZ Direct arvato services, and Deutsche Post. Data also stem from official institutions such as Kraftfahrt-Bundesamt, the Statistical Offices of the Federation and the Länder (*Statistische Ämter des Bundes und der Länder*), and the Federal Employment Agency (*Bundesagentur für Arbeit*).

availability. The dataset RWI-GEO-GRID is organized in four packages with different scopes.

Table 1 gives an overview of the different packages and the information they contain. Each dataset contains **basis-data** which are comprised of four categories: number of households, number of commercial enterprises, number of houses (including pure commercial buildings), and number of residential buildings (excluding pure commercial buildings).

The package **HOUSEHOLD** contains information on *household structure, children, the unemployment rate, the purchasing power, payment defaults, foreigners* and *ethnicities*. The variable *household structure* differentiates between singles, couples and families and is largely based on the household size as well as the number of children (RWI/microm 2017f). The variable children specifies the average share of children in a household (RWI/microm 2017c). The variable *unemployment rate* is defined as the share of the unemployed in the population

Table 1: Overview on available data.

	Variable	Description
Basis (included in all packages)	Number of households*	Absolute number*
	Number of commercial enterprises	Absolute number*
	Number of houses (including pure commercial buildings)	Absolute number*
	Number of residential buildings (excluding pure commercial buildings)	Absolute number*
Mobility	Car capability	Index scores
	Car brands	Score values
	Car segments	Score values
Building Development	House type	Percentage per household
Household	Purchasing power	In Euro*
	Household structure	Percentage per household
	Children	Number per household*
	Unemployment rate	Share of the unemployed in the population
	Ethno	Percentage per household
	Foreigners	Percentage per household
	Payment default	Percentage per household
Population	Gender and age structure	Share of inhabitants w.r.t. sex and 17 age groups*
	Population structure	Absolute number of inhabitants*

**Anonymized in the Scientific Use File if the variable has a value of 5 or less in a specific grid.*

of those working or searching for employment (RWI/microm 2017l). The variable *purchasing power* reflects the household income. It comprises information on labour supply, capital wealth, rental and leasing income minus taxes and social security contributions, including social transfers such as unemployment benefits, child-allowances and pensions (RWI/microm 2017j). The variable *payment default* describes the statistical probability of payment default for each house in Germany. The houses are grouped into 9 risk groups (RWI/microm 2017g). The variables *share of foreigners* (RWI/microm 2017k) and *ethno* (RWI/microm 2017d) are based on an analysis of fore- and surnames with respect to their linguistic origin. The evaluation is based on lists of names and their origin, and only refers to the head of the households. Thus, this does not allow for conclusions about the actual number of persons with a foreign origin, how long the individuals have already been Germany, or how strong their social integration is.[4]

The second package, **BUILDING DEVELOPMENT**, contains information on different house types. The variable *house type* indicates the size of a house and is based on the number of the households and the number of firms in a house. There are 7 house types in the data: single- and two-family homes in homogenous road sections, single- and two-family homes in heterogeneous road sections, 3–5 family homes, 6–9 family homes, blocks of flats with 10–19 households, multi-storey buildings with 20 and more households, mainly commercially used houses (RWI/microm 2017e).

The third package is **DEMOGRAPHY**. This group contains information on the *population by age and gender* (RWI; microm 2017i) as well as the absolute number of *inhabitants* (RWI/microm 2017h). The variable *gender and age structure* indicates the share of inhabitants in an area and can be differentiated with respect to sex and 17 age groups. Age between 20 and 75 is divided into categories of 5 years each to build the age groups. For the elderly, the group is "75 years and older". For children, microm uses the following categories: infants in the age range 0–3 years, 3–6 year, 6–10 years, 10–15 years, 15–18 years and 18–20 years.

The fourth package is **MOBILITY**. This package contains information *on car capability, car brands* and *car segments*. The variable *car brand* indicates the relative market share of 14 different brands compared to others within a geographic area (RWI/microm 2017a). For the variable *car segments*, cars have been aggregated to classes that allow for conclusions about the socio-economic

4 For the variable "ethno" the following linguistic origins can be distinguished: Turkey; Italy; Greece; Spain/Portugal/Latin America; ethnic German repatriates from the former Soviet Union (*Spätaussiedler*); Eastern Europe; the Balkans; Africa south of the Sahara; non-European Islamic states; South-, East- and Southeast-Asia (microm Consumer Marketing 2016, p. 40).

status. The dataset comprises of 12 car segments: mini cars, compact cars, lower mid-range cars, mid-range cars, upper mid-range cars, top-of-the-range cars, ATVs, cabriolets, estate cars, vans, utility vehicles, other vehicles (RWI/microm 2017b).

4 Advantages and Analyses Potential

The data can be used for analyses in a wide range of fields with social and political relevance. For example, access to such small-scale data opens the possibility to locate distress areas precisely, and it allows to identify and to analyze different outcomes based on neighborhood characteristics. Moreover, small-scale data are crucial for politicians to define deprived areas, and react appropriately as a consequence. Relevant analyses on e. g. neighborhood effects, segregation and urban planning can highly benefit from the use of the RWI-GEO-GRID.

One application of the dataset is the population projection "RWI-GEO-GRID-POP-FORECAST" (Breidenbach et al. 2017). This projection of the natural population development on the grid level up to the year 2050 reveals detailed variations in the population loss and aging dynamics on a small regional scale.

Moreover, the data can be merged to existing datasets which contain precise geographical information.[5] Some examples illustrate the wide range of analyses which benefit from these data. Hentschker and Mennicken (2017) control for heterogeneity of patients in different hospitals by adding the purchasing power of a patient's postcode area. Schaffner and Treude (2014) use the data to identify ethnic enclaves and analyze labor market outcomes of foreigners conditional on being located in such an enclave or not. Micheli (2016) as well as Frondel et al. (2017) use the data to explain different apartment purchasing and rental prices on the local level. This combination with data on real estate advertisements is also used by Eilers (2017) in order to gain more precise estimates for a rental price index on a small-scale regional level.

Furthermore, the grid-level data can be used to converse data on a higher aggregated level from one spatial unit to another. Using grid-level data, county data, for example, can be broken down to lower levels such as postcodes. This is very useful since there are only scarce information available on the postcode level. In addition to this cross-sectional transformation from one spatial unit to

5 The underlying data provided by microm have been merged to the German Socio-Economic Panel (SOEP), called SOEP-microm, as documented in (Goebel et al. 2014).

another, the data also allow a harmonization of spatial units over time. Territorial reforms form a severe challenge for any regional research over a longer time period. The grid data allow to transform territorial boundaries from different years into a harmonized base.[6]

5 Data Availability

There are two versions of the RWI-GEO-GRID dataset. First, the Scientific Use File (SUF), which has a stronger anonymization if less than 5 inhabitants or households live in one grid cell. For these cases, absolute values are censored and the dataset only contains those variables which are provided as shares. Second, the full dataset is available in the Data Secure Room of the FDZ Ruhr in Essen. The data can be obtained as a Stata® dataset (.dta), R dataset (.rds), Excel (.xlsx) sheet or .csv file. Data access to both versions requires a signed data use agreement. Both versions are restricted to non-commercial research and only researchers of scientific institutions are eligible to apply for data access. The SUF may be used at the workplace of the users.

Since the data are purchased from microm, users are charged with a processing fee of 100 € plus VAT. Data access is provided by the Research Data Centre Ruhr at the RWI – Leibniz-Institute for Economic Research (FDZ Ruhr). Data access can be applied for online at http://fdz.rwi-essen.de/application.html. The application form includes a brief description and title of the project, potential cooperation, information on the applying department, expected duration of data usage as well as further participants in the project.

References

Breidenbach, P., M. Kaeding, S. Schaffner (2017), Population Projection for Germany 2015-2050 on Grid Level (WI-GEO-GRID-POP-FORECAST). Jahrbücher für Nationalökonomie und Statistik *forthcoming*.

Eilers, L. (2017), Is My Rental Price Overestimated? A Small Area Index for Germany. Ruhr Economic Papers 734, RWI.

Frondel, M., A. Gerster, C. Vance (2017), The Power of Mandatory Quality Disclosure: Evidence from the German Housing Market. Ruhr Economic Papers 684, RWI.

6 The described distribution keys for these transformations over time and conversions over spatial units are produced at the RWI and will be published as RWI-GEO-BRIDGE.

Goebel, J., C.K. Spieß, N.R. Witte, S. Gerstenberg (2014), Die Verknüpfung Des SOEP Mit MICROM-Indikatoren: Der MICROM-SOEP-Datensatz. SOEP Survey Papers. Series D – Variable Descriptions and Coding 233. German Institute for Economic Research.

Hentschker, C., R. Mennicken (2017), The Volume-Outcome Relationship Revisited: Practice Indeed Makes Perfect. Health Services Research.

Micheli, M. (2016), Local Governments' Indebtedness and Its Impact on Real Estate Prices. Ruhr Economic Papers, RWI.

microm Consumer Marketing (2016), Datenhandbuch: Arbeitsunterlagen Für Microm MARKET & GEO. Neuss, microm GmbH, Neuss.

RWI; microm (2017a), Socio-Economic Data on Grid Level (SUF 5.1). Car Brands. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:pkwmarken:suf:V5:1.

RWI; microm (2017b), Socio-Economic Data on Grid Level (SUF 5.1). Car Segments. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:pkwseg:suf:V5:1.

RWI; microm (2017c), Socio-Economic Data on Grid Level (SUF 5.1). Children. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:kinder:suf:V5:1.

RWI; microm (2017d), Socio-Economic Data on Grid Level (SUF 5.1). Ethno. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:ethno:suf:V5:1.

RWI; microm (2017e), Socio-Economic Data on Grid Level (SUF 5.1). House Type. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:haustyp:suf:V5:1.

RWI; microm (2017f), Socio-Economic Data on Grid Level (SUF 5.1). Household Structure. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:hstruktur:suf:v5:1.

RWI; microm (2017g), Socio-Economic Data on Grid Level (SUF 5.1). Payment Index. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:zahlindex:suf:v5:1.

RWI; microm (2017h), Socio-Economic Data on Grid Level (SUF 5.1). Population. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:einwohner:suf:V5:1.

RWI; microm (2017i), Socio-Economic Data on Grid Level (SUF 5.1). Population by Age and Gender. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:einwGeAl:suf:V5:1.

RWI; microm (2017j), Socio-Economic Data on Grid Level (SUF 5.1). Purchasing Power. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:kaufkraft:suf:v5:1.

RWI; microm (2017k), Socio-Economic Data on Grid Level (SUF 5.1). Share of Foreigners. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:auslaender:suf:v5:1.

RWI; microm (2017l), Socio-Economic Data on Grid Level (SUF 5.1). Unemployment Rate. RWI-GEO-GRID. Version: 5.1. RWI – Leibniz Institute for Economic Research. Dataset. http://doi.org/10.7807/microm:alq:suf:V5:1.

Schaffner, S., B. Treude (2014), The Effect of Ethnic Clustering on Migrant Integration in Germany. Ruhr Economic Papers 536, RWI.