Universität Trier

Analysing local-level rental markets based on the German Mikrozensus

Charlotte Articus
Hanna Brenzel
Ralf Münnich

Research Papers in Economics
No. 9/20

Analysing local-level rental markets based on the German Mikrozensus

Charlotte Articus*, Hanna Brenzel† and Ralf Münnich‡

November 27, 2020

Abstract

Recent developments on German real estate markets show a striking increase of rents, especially in larger towns. This development is, however, not homogeneous: The market dynamics vary between different parts of cities and prices develop highly heterogeneously. Therefore, small-scale results are of interest.

The German Mikrocensus, Germany's largest household survey, routinely provides information on housing, including rental prices. Since 2018 results are geo-coded, setting the prerequisite for obtaining results at a local level. This specifically strong regional disaggregation obviously results in small sample sizes for the entities of interest. While standard design-based estimators under these condition result in large standard errors, small are estimation techniques may be a solution to nevertheless obtain reliable estimates.

Therefore, the present paper explores the opportunity of obtaining very small-scale estimates of average rental prices based on the Mikrozensus employing small area estimation models. The study focusses on the City of Cologne, which provides a broad range of indicators that can be employed as auxiliary information in the models.

*Economic and Social Statistics Department, University of Trier, Germany. Email: articus@uni-trier.de

†Federal Statistical Office, Germany. Hanna.Brenzel@destatis.de

‡Economic and Social Statistics Department, University of Trier, Germany. Email: muennich@uni-trier.de

1 Introduction

In the last decade, rental prices in Germany have increased distinctly. This is especially true for large German cities, that have grown over-proportionally in recent years due to strong in-migration of younger age groups from rural regions (Statistisches Bundesamt, 2020). The resulting rising pressure on rental markets has fuelled a public debate on affordable housing. The topic has been identified as the crucial social question of our time (see Süddeutsche Zeitung (2018)).

The public debate and political targeting of related problems require valid and comprehensive information. In providing this foundation, it has to be taken into account, that rental markets develop highly heterogeneously. This is not only true for differences between urban and rural regions; the micro location within cities also strongly determines market conditions. Therefore, to complete the picture, very small-scale results are of interest.

The German Mikrozensus, Germany's largest household survey, routinely contains a special evaluation on housing. It is the only source that contains nationwide information on actually paid rents, including prices paid in ongoing tenancies. Additionally, a broad range of dwelling characteristics such as house and apartment sizes, year of completion and contract duration is available. In 2018, the survey was geocoded for the first time so that analyses at a very low resolution level become possible. Naturally, a strong regional disaggregation of the available sample results in small subsamples for the entities of interest. A natural solution is to apply Small Area Estimation (SAE), i.e. estimation methods that are specifically designed to produce reliable results in case of small subsamples (see Rao and Molina (2015) for a comprehensive overview). We, therefore, explore the potential of obtaining low scale estimates of average rental prices based on the Mikrozensus 2018 applying model-based small area estimation techniques. In this, we are interested in the question of how small is too small, i.e. how far can we disaggregate the available sample while still obtaining reliable results. We focus on the City of Cologne, Germany's fourth largest city. The city's administration provides a broad range of social indicators at the resolution level of boroughs, districts and even neighbourhoods as open data, so that model-based estimation techniques at a local level can be employed.

The paper is organized as follows. The next session presents the available data sources. Subsequently, we introduce the typical problem setting of SAE and the relevant small area estimators. We then present results on two different hierarchical levels within Cologne city. Finally, we close with a summarizing evaluation.

2 Data

The annual German Mikrozensus is the largest household survey conducted by official statistics in Germany. Its purpose is to deliver detailed information on the socio-economic situation of the society including a broad range of topics such as employment, education, health, and living conditions. The survey is designed to cover the resident population of Germany and is drawn as a one-step cluster sample. Sampling units are clusters of apartments, that are either comprised by several neighbouring buildings or – in large buildings – by all or

some apartments of a building. Overall, a cluster comprises on average nine apartments. A regional stratification and strata for different building sizes are implemented to secure respective coverage of the sample. All households and individuals in the sampled clusters enter the sample. In 2018, approximately 375,000 households or 752,000 individuals were questioned. Participation is mandatory. See Statistisches Bundesamt (2019a) for further information on the Mikrozensus.
Every fourth year, most recently in 2018, the Mikrozensus contains a special evaluation on housing. It delivers detailed information on rental prices and dwelling characteristics (see Statistisches Bundesamt (2019b). The official results are published on level of the German Länder and for a small range of selected cities and large administrative or statistical subentities. Nationwide results on a lower resolution level are not routinely provided.
For the study conducted here, Horvitz-Thompson estimates of average rental prices per square meter were calculated at level of the 86 districts as well as 294 neighbourhoods nested in these districts. Furthermore, variance estimates were obtained. The estimates are based on 3130 observations. Overall, 17 out of 86 districts were either unsampled so that no design-based estimate could be obtained or information was suppressed due to confidentiality reasons. At the level of the neighbourhoods, only for 98 out of 294 areas a result could be made available. For the remaining entities the coefficient of variation (CV) ranges from 0.01 to 0.18 (mean: 0.06) for the districts and from 0.01 to 0.26 (mean: 0.08) for the neighbourhoods.
Auxiliary information, which is required if model-based estimators are to be employed, can be obtained from publicly available data collections supplied by official statistics. The City of Cologne provides public access to indicators on the very fine resolution scales of 9 boroughs, 86 districts and 294 neighbourhoods (City of Cologne, 2020). We gathered 154 and 112 indicators on level of the districts and neighbourhoods, respectively. These contain register- or census-based information on the population, such as the age structure, household compositions, migration and duration of residence, nationality and confession, voting behaviour, etc.

3 Small Area Estimation

The standard Fay-Herriot (FH) model as first suggested by Fay and Herriot (1979) is given by (see Rao and Molina (2015), section 4.2 and 6):

$$\begin{aligned}\hat{\mu}_i^{\text{Dir}} &= \boldsymbol{x}_i^T\boldsymbol{\beta} + v_i + e_i \quad \text{for } i = 1, \ldots, m \\ v_i &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_v^2) \\ e_i &\overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e,i}^2)\end{aligned}$$

$\hat{\mu}_i^{\text{Dir}}$ is the direct estimate obtained from the sample realized in area i. $\mathbf{x}_i$ is a p-vector of auxiliary information. $\boldsymbol{\beta}$ is an p-vector of regression coefficients. $\sigma_{e,i}^2$ is the (known) design variance of direct estimates $\hat{\mu}_i^{\text{Dir}}$. v_i is an area-specific random effect. It is assumed that e_i and v_i are independent.
The Empirical Best Linear Unbiased Predictor (EBLUP) of the parameter of interest μ under this model, i.e. the Fay-Herriot-estimator, is given by Rao and

Molina (2015)

$$\begin{aligned}\hat{\mu}_i^{\text{FH}} &= \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \hat{v}_i \\ &= \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_i(\hat{\mu}_i^{\text{Dir}} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) \\ &= \gamma_i\hat{\mu}_i^{\text{Dir}} + (1-\hat{\gamma}_i)\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\end{aligned}$$

with

$$\hat{\gamma}_i = \frac{\sigma_v^2}{\sigma_{e,i}^2 + \hat{\sigma}_v^2}.$$

Note that $\hat{\mu}_i^{\text{FH}}$ can be expressed as a composite estimator of the synthetic estimator $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ obtained from the fixed part of the model and the direct estimator $\hat{\mu}_i^{\text{Dir}}$. Weights are given by the area-specific shrinkage factor $\hat{\gamma}_i$, that is the model variance $\hat{\sigma}_v^2$ relative to the total variance $\sigma_{e,i}^2 + \hat{\sigma}_v^2$.
The aim, of course, is to stabilize the estimation, that is to yield estimates with a far smaller variance in the context of small sample sizes. Note, however, that this comes with the price of loosing the property of design-unbiasedness. We, hence, face a trade-off between bias and variance (that is solved optimally when estimating the EBLUP). The relevant measure to judge the quality of model-based small area estimates, therefore, is the Mean squared error (MSE), that is $\text{MSE}(\hat{\mu}_i^{\text{FH}}) = \text{E}(\hat{\mu}_i^{\text{FH}} - \mu_i)^2$. For variance components estimated by REML, it can be estimated as

$$\widehat{\text{MSE}}_{\text{REML}}(\hat{\mu}_i^{\text{FH}}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$$

with

$$g_{1i}(\sigma_v^2) = \frac{\sigma_v^2\sigma_{e,i}^2}{\sigma_v^2 + \sigma_{e,i}^2}, \tag{1}$$

$$g_{2i}(\sigma_v^2) = \left(\frac{\sigma_{e,i}^2}{\sigma_v^2 + \sigma_{e,i}^2}\right)^2 \boldsymbol{x}_i^T\left(\sum_{j=1}^m \frac{\boldsymbol{x}_j\boldsymbol{x}_j^T}{(\sigma_{e,j}^2 + \sigma_v^2)}\right)^{-1}\boldsymbol{x}_i \tag{2}$$

$$g_{3i}(\sigma_v^2) = \frac{\sigma_{e,i}^4}{(\sigma_v^2 + \sigma_{e,i}^2)^3}\frac{2}{\sum_{j=1}^m \frac{1}{(\sigma_v^2+\sigma_{e,j}^2)^2}}. \tag{3}$$

See Datta and Lahiri (2000), Datta et al. (2005) and Rao and Molina (2015, Chapter 6.2.1) for details.

4 Results

4.1 Results at District level

We estimated a FH model for the rents per square meter at level of the 86 districts. We employed simple step-wise selection procedures to select a set of covariates for the model, using the conditional AIC as suggested by Vaida and Blanchard (2005) as a model selection criterion. This resulted in a model using the following four indicators:

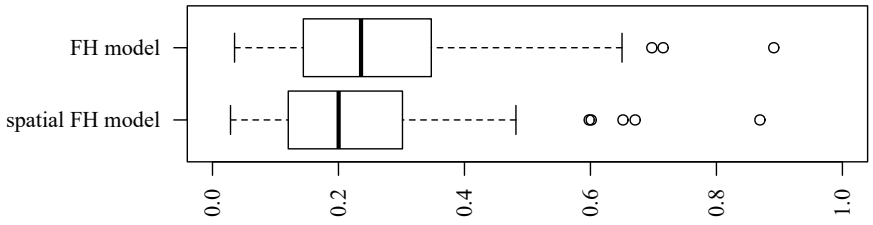

Figure 1: shrinkage factors of small area models

Households: Share of unmarried couples without children	HH
Inhabitants/hectare	DENS
Share of votes for the liberal party in the federal election (2013)	LIB
Share of inhabitants in age group 0 to < 18	YOUTH

Table 1: Covariates in the model

Further, we estimated the spatial extension of the FH model. When defining neighbouring areas, the river Rhein was treated as a separating line, to the effect that districts on different sides of the river are never considered as neighbouring. This corresponds to the strong dividing effect of the river in terms of living environment and infrastructure.
Fixed model coefficients are given in Table 2. The estimated model variance σ_v^2 is 0.0899 for the FH model and 0.073 for the spatial extension of the FH model.

	beta	std.error	tvalue	pvalue
(Intercept)	9.02	0.71	12.75	2.98^{-37}
HH	0.16	0.09	1.82	6.8^{-2}
DENS	0.01	0.00	2.62	8.7^{-3}
LIB	0.27	0.04	6.21	5.1^{-10}
YOUTH	-0.12	0.03	-4.46	8.2^{-6}

Table 2: Model coefficients FH model

The estimated model variances determine the weight of the synthetic estimator in the model-based estimation. Figure 1 depicts the resulting distribution of shrinkage factors $\hat{\gamma}_i$ for both models. It can be taken from this plot that, overall, reliance on the model is large. This result is even more pronounced for the spatial FH model.
Generally, this indicates that a large part of the variability in the data can be explained through the model. While this is a promising result, as it is a prerequisite for the intended reduction in variance, the reliance on the model also comes at the cost of a possible bias. Thus, a close evaluation of the model

is important.
In Figure 2 the small area-estimates are plotted against the design-based direct estimates. Thus, the unbiased but potentially imprecise direct estimates are employed to judge the bias of model-based estimates (see BROWN et al. (2001)). If these are unbiased, too, data points can be expected to scatter randomly around the identity line. To further analyse large deviations between model-based and design-based estimates, data points with a deviation larger than the standard deviation of the direct estimate are indicated by a cross. These differences may either stem from a large deviation between synthetic estimate and direct estimate or from a large variance of direct estimates Additionally, we plotted the regression line for regressing $\hat{\mu}_i^{\text{Dir}}$ on $\hat{\mu}_i^{\text{FH}}$.

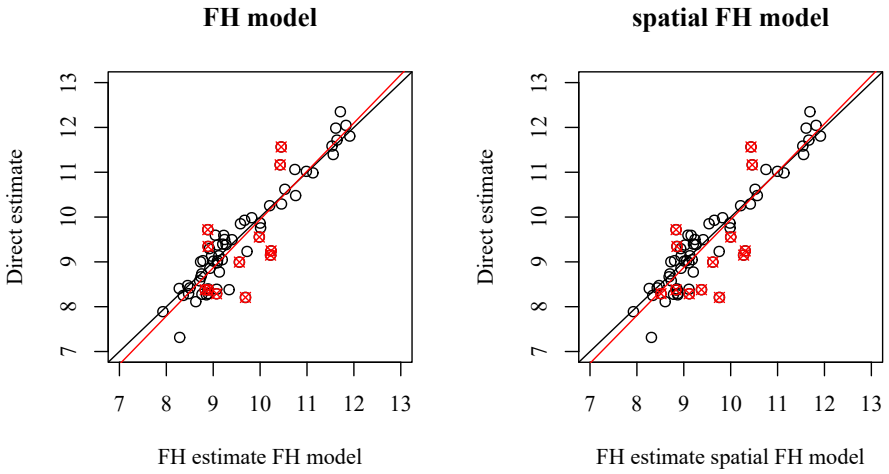

Figure 2: Graphical bias diagnostics

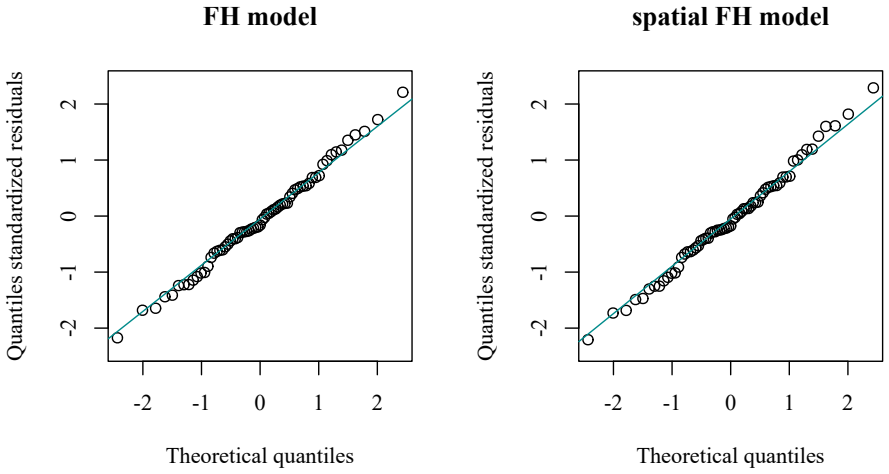

Figure 3: Normal QQ-plot

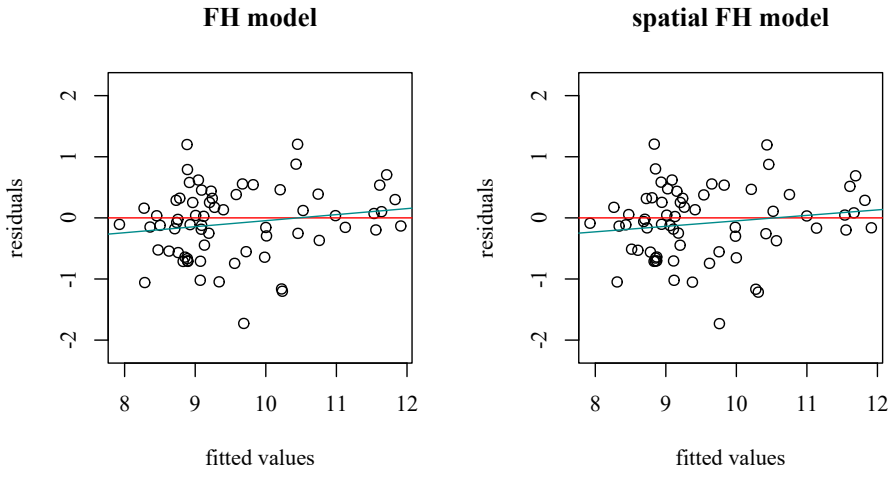

Figure 4: Residuals vs fitted values

With Figure 3 and Figure 4 two standard instruments of graphical residual analysis are employed to check the validity of model assumptions. The normal qq-plot in Figure 3 evaluates the normality assumption for the error terms. The result is fairly good for both the FH model and the spatial FH model, with only a slightly larger dispersion in the right tail of the distribution of standardized residuals. In Figure 4 residuals are plotted against fitted values. Again, no strong patterns are detectable.

Figure 5 shows the gain in accuracy realized through the use of small area estimation techniques. We can take from this plot that the improvement is large. Median variance of direct estimates is 0.29 and the distribution is skewed to the right with a maximum of 2.47. This could be reduced to a median MSE of 0.09 and 0.08 with the FH model and the spatial FH model, respectively. The improvement is also distinctly visible in the tails of the distribution.

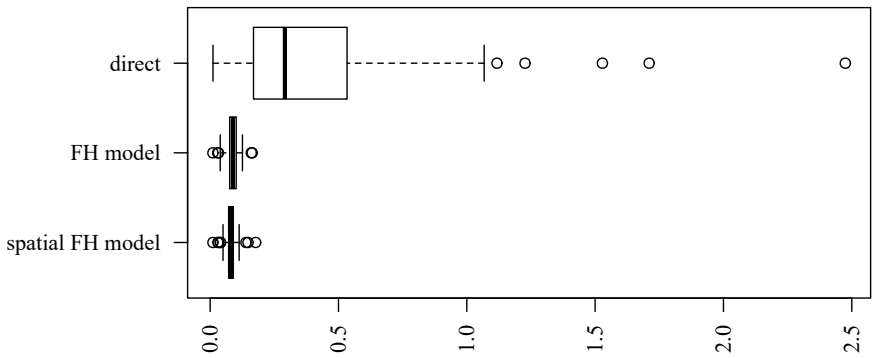

Figure 5: Variance of direct estimates and MSE of model-based estimates

Finally, the results for average rents per square meter obtained from both the FH model and the spatial FH model are illustrated in figure 6. Estimated prices range from 7.9 to 11.9 Euro/square meter. The map clearly shows the particularly high prices in the center and the southern and, even more pronounced, western districts of the city. Furthermore, the dividing line of the river Rhein is distinguishable; prices are generally higher west of the river.

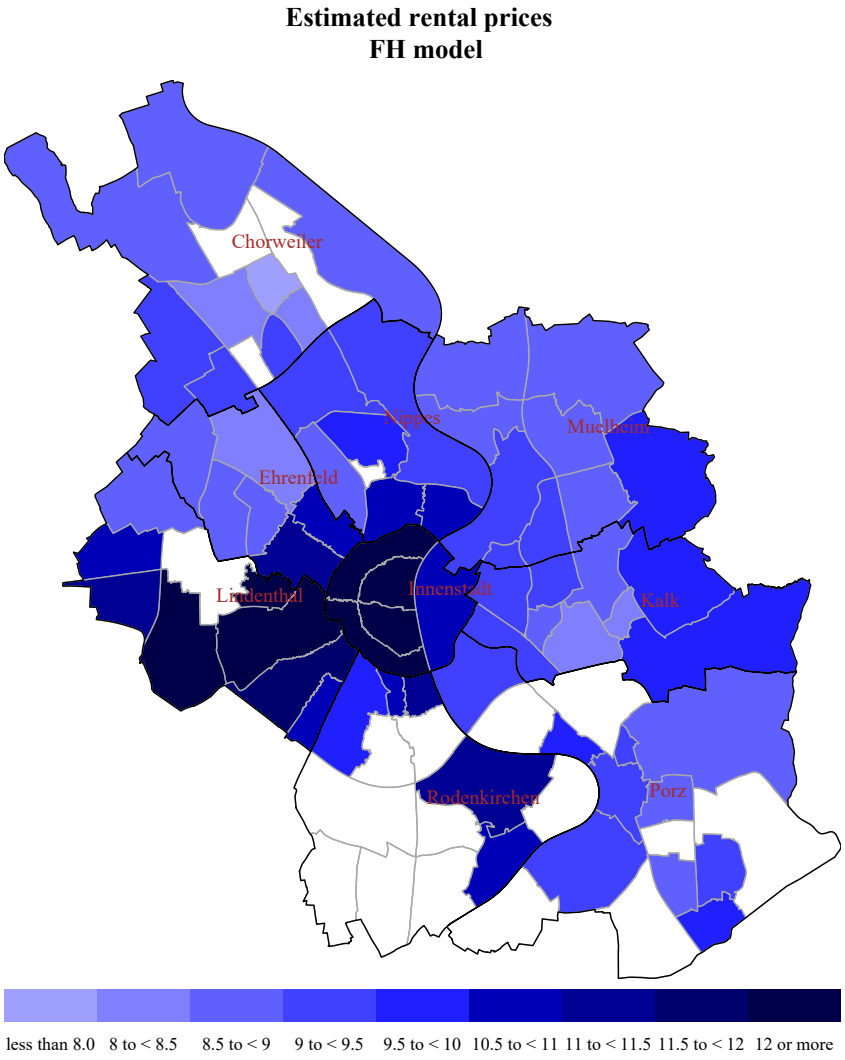

Figure 6: FH estimates of average rental prices at district level

Figure 6 also identifies the districts, for which no information is available, either because the area-specific sample size was too small to allow disclosure of results or because it was indeed unsampled. Model-based small area estimation offers the opportunity to predict the missing information from the synthetic model, i.e. to obtain

$$\hat{\mu}_i^{\text{synth}} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}. \tag{4}$$

Figure 7 depicts these results.

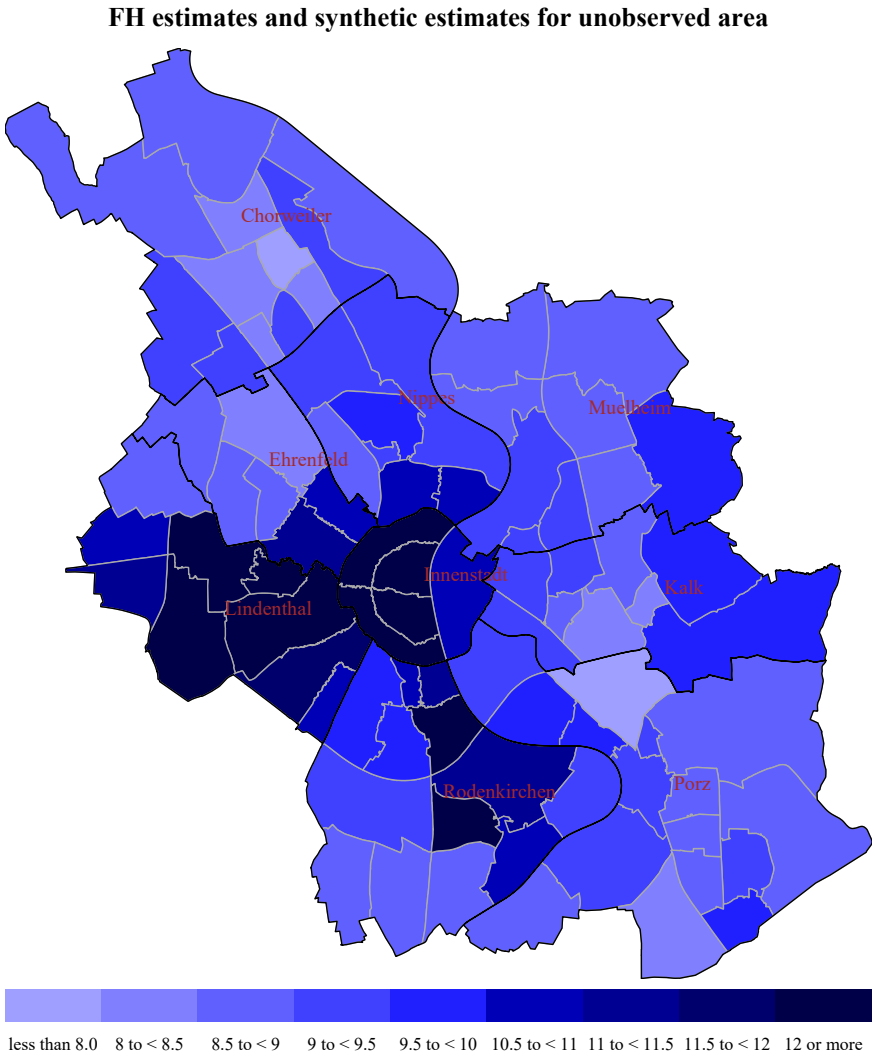

Figure 7: FH estimates of average rental prices at district level

4.2 Results at level of the Neighbourhoods

In a second step, model-based estimates are obtained on the even finer resolution level of the 294 neighbourhoods. As mentioned above, direct estimates were only available for 98 of the 294 neighbourhoods. While unsampled and suppressed cells are to be expected on this very low level of analysis, the particularly large amount of affected areas in combination with the overall moderate level of CVs is due to the sampling design; the selection of sampling units comprising neighbouring apartments results in regional clusters of relatively homogeneous dwellings. This is a strong limitation of the present study, that has to be kept in mind. Because of the large number of areas that are excluded from the analysis and the resulting island character of many areas, we do not estimate the spatial

extension of the FH model at this level.
As above, the model is chosen applying step-wise selection procedures. The following indicators were chosen:

Migration to Cologne in proportion to inhabitants	MIGR
Confession: share of protestants	CONF
Share of inhabitants born in Cologne	BIRTH
Share of one-person-Households	HH

Table 3: Covariates in the model

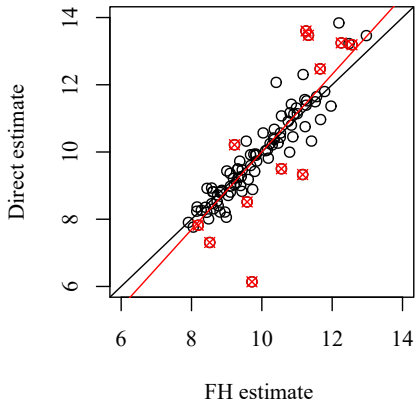

Figure 8: Graphical bias diagnostics

As in the preceding Section, with Figure 8 we evaluate the bias of small area estimates by plotting them against the available direct estimates. While the majority of points scatters randomly around the identity line, the plot reveals biased results for specifically low- and high-priced areas. Results are further evaluated with a normal qq-plot and a plot of residuals vs. fitted values in 9. The bias in the especially high and low priced areas is detectable here, too.

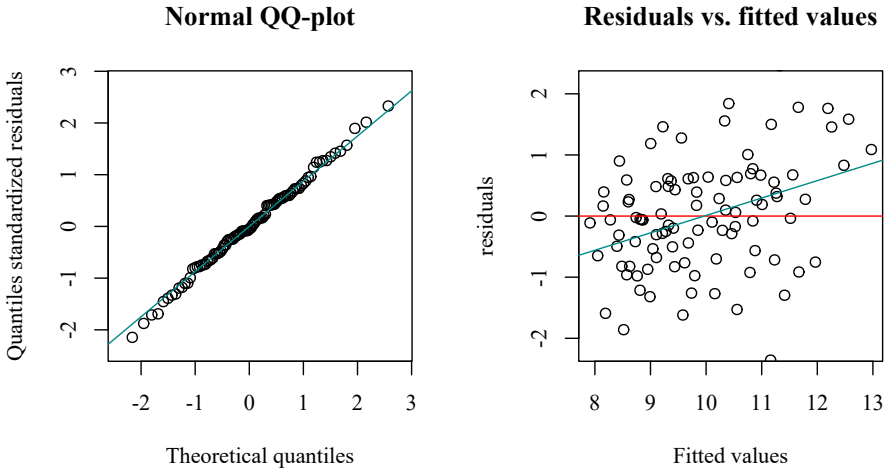

Figure 9: Diagnostic plots

5 Conclusion

This study has evaluated the opportunities of deriving very low scale estimates of rental prices based on the German Mikrozensus. Area-level small area models were employed to obtain reliable results on level of districts and neighbourhoods in the city of Cologne. The study revealed that the usability of the survey data is restricted by the sampling design of the Mikrozensus. On the very low level of neighbourhoods, the cluster design results in many unsampled areas while at the same time the sample does not reflect the heterogeneity of observations in sampled areas. Thus, we conclude, that the data is not suitable for evaluations at this very fine resolution level and chose not to present respective results here. We do, however, think that it is worthwhile to further pursue the approach of employing Mikrozensus survey data to obtain valuable information on rental markets in large German cities on the already very high resolution level of the districts. As stated above, it is the only way to obtain information on established and ongoing tenancies. At the same time, the study has shown that suitable methods are available and high-quality auxiliary information with good explanatory power is openly provided by city statistics. Correspondingly, the analysis on level of districts shows that a large gain in accuracy can be realized through the employment of small area models. The Mikrozensus further contains information on central dwelling characteristics. In a next step, this potential should be exploited by differentiating between apartment classes to yield an even more detailed picture of local rental markets.

Acknowledgements

The research was supported by the DFG Research Unit FOR 2559 MikroSim. The first author is financed by the Nikolaus Koch Stiftung within the project REMIKIS. The authors also thank the City of Cologne for the very valuable supply of open data.

References

Brown, G., **Chambers, R.**, **Heady, P.** and **Heasman, D.** (**2001**): *Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS*. Proceedings of Statistics Canada Symposium 2001, Statistics Canada.

City of Cologne (**2020**): *Statistische Daten*. `https://www.stadt-koeln.de/politik-und-verwaltung/statistik/statistische-daten-thematische-karte`, accessed: 2020-10-20.

Datta, G. S. and **Lahiri, P.** (**2000**): *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems*. Statistica Sinica, 10, pp. 613–627.

Datta, G. S., **Rao, J. N. K.** and **Smith, D. D.** (**2005**): *On measuring the variability of small area estimators under a basic area level model*. Biometrika, 92 (1), pp. 183–196.

Fay, R. E. and **Herriot, R. A.** (**1979**): *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data*. Journal of the American Statistical Association, 74 (366), pp. 269–277.

Rao, J. N. K. and **Molina, I.** (**2015**): Small Area Estimation. Wiley series in survey methodology, New York: John Wiley & Sons, 2 ed.

Statistisches Bundesamt (**2019**a): Mikrozensus 2018. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2019**b): Wohnen in Deutschland. Zusatzprogramm zum Mikrozensus 2018. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2020**): *Städte-Boom und Baustau: Entwicklungen auf dem deutschen Wohnungsmarkt 2008 – 2018*. Accessed: 2020-10-20.
URL `https://www.destatis.de/DE/Presse/Pressemitteilungen/2019/12/PD19_N012_122.html`

Süddeutsche Zeitung (**2018**): *Deutschlands Mietmarkt ist kaputt*. Accessed: 2020-10-20.
URL `https://projekte.sueddeutsche.de/artikel/wirtschaft/miete-wohnen-in-der-krise-e687627/`

Vaida, F. and **Blanchard, S.** (**2005**): *Conditional Akaike information for mixed-effects models*. Biometrika, 92 (2), pp. 351–370.