wik

Working Paper

No 4

Costs of Very High Capacity Networks and Geographic Heterogeneity – a statistical assessment for Germany

Konrad Zoz
Gonzalo Zuloaga
Dr. Gabriele Kulenkampff
Dr. Thomas Plückebaum
Martin Ockenfels

December 2021

wik Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste

IMPRINT

WIK WORKING PAPER

WIK Wissenschaftliches Institut für Infrastruktur
und Kommunikationsdienste
Rhöndorfer Str. 68, 53604 Bad Honnef, Germany
www.wik.org

Person authorised to sign on behalf of the organisation:

General Manager and Director	Dr Cara Schwarz-Schilling
Director	Alex Kalevi Dieke
Director/Head of Department/ Networks and Costs	Dr Thomas Plückebaum
Director/Head of Department/ Regulation and Competition	Dr Bernd Sörries
Head of Administration	Karl-Hubert Strüver
Chairperson of the Supervisory Board	Dr Daniela Brönstrup
Registered at	Amtsgericht Siegburg, HRB 7225
Tax No.	222/5751/0722
VAT-ID	DE 123 383 795

Abstract

In this study, we analyse regional cost differences of fibre-based access networks. Our data base comprises a complete sample of Very High Capacity Network (VHCN) investment figures. By matching this data with the internationally standardised EUROSTAT and BBSR urban/rural typology classification, we show that such classification criteria do not sufficiently account for a large share of geographical differences in fibre-based access network costs. In order to better explain and/or identify regional differences in VHCN investment, we turn to spatial regression models to identify alternative influencing factors solely on the basis of publicly available data. We show that a handful of geographical factors are capable of explaining 95% of the differences in fibre investment requirements; the most relevant being (1) the size of demand (as number of access lines), (2) the street-based household density (defined as the number of households per kilometre of road in built-up areas), (3) a dispersion measure (approximated by the main road length per built-up area) and (4) the degree of urbanisation (measured by the share of built-up area in relation to the overall area). These results are consistent at different levels of spatial aggregation (e.g. from access areas to NUTS-3 level) and even after controlling for neighbouring effects. Thus, it is capable of predicting costs more precisely and at the level of the territorial unit, at which funds are bounded to be allocated. From a public policy perspective, the proper identification of areas, where the commercial roll-out is unlikely to occur, is key in preventing the widening of a digital gap without having a wasteful use of public funds.

Keywords: Very high capacity networks (VHCN), bottom-up cost models, statistical estimations, spatial analysis, NUTS-3, state aid

Contents

List of figures

List of tables

1 Introduction

In this paper, we investigate the heterogeneity of costs of Very High Capacity (VHC) networks in Germany and determine to what extent costs for VHC networks differ between urban and rural regions.

The data used in this study is taken from a bottom-up cost modelling task, exercised in a previous study.[1] It is derived from a complete sample of approximately 8000 MPoP areas in Germany, representing the investment requirements of a nationwide Fibre to the Home (FTTH) network (VHCN) at 100% homes-passed. The model results are derived from publicly available data on georeferenced buildings and household data and reflect regionally differentiated FTTH investment at the level of access areas.

In the first part of the study, we make a statistical assessment of the regional cost differences of access network areas, especially among rural areas. It is a common understanding that rural areas exhibit the lowest economic viability of a network roll-out. For this purpose, we start by defining “urban” and “rural” areas following the internationally standardised regional classifications determined by (1) EUROSTAT urban/rural typology and (2) the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) settlement structure. Since these statistics are originally created to shed light on economic and social issues related to rural areas, we analyse their suitability as a differentiation criterion for explaining regional differences in costs for FTTH access networks.

In the second part of the paper, we establish statistical estimation models that explain the observed regional differences in fibre-based access network costs. Based on publicly available data, we determine the most influential factors on fibre investment requirements and assess their cost prediction power at different levels of spatial aggregation.

1 Most of the data used in this study was collected and produced during a research project conducted between 2018 and 2020. For the description of outcomes see Kulenkampff G., Ockenfels M., Zoz K. and Zuloaga G. (2020): Costs of Broadband access networks (in German, English summary available), WIK Discussion Papers no. 473, Bad Honnef, December 2020; electronically available under: Kosten von Breitband-Zugangsnetzen (wik.org).

2 Definitions, Data and Methodology

2.1 VHCN -Definition

According to the digital strategy of the European Commission, by 2025 Very High Capacity Networks (VHCN) should provide broadband connections delivering:[2]

- One Gbps for all schools, transport hubs and main providers of public services and digitally intensive enterprises,
- Download speeds of at least 100 Mbps to be upgraded to 1 Gbps for all European households, and
- Uninterrupted 5G wireless broadband coverage for all urban areas and major roads and railways.

Article 2(2) of the EECC defines Very High Capacity Networks (VHCN)[3] as

- "either an electronic communications network which consists wholly of optical fibre elements at least up to the distribution point at the serving location"
- "or an electronic communications network which is capable of delivering, under usual peak-time conditions, similar network performance in terms of available downlink and uplink bandwidth, resilience, error-related parameters, and latency and its variation".

This EECC definition has been more detailed in the BEREC Guidelines on Very High Capacity Networks. There, a VHCN is specified as a network providing a fixed-line connection, which under usual peak-time conditions, delivers a downlink data rate of at least 1000 Mbps and a uplink data rate of at least 200 Mbps.[4] Thus, multiple access network architectures can be classified as Very High Capacity Networks, like: FTTH, FTTB and DOCSIS 3.1 and 4.0.

In this paper, we do not investigate cost differences of all these different Very High Capacity Networks architectures. Instead, we restrict our analysis to just one Very High Capacity Network architecture: FTTH/P2P. We consider cost differences between

2 European Commission: Broadband Strategy and Policy, 12 March 2021, https://ec.europa.eu/digital-single-market/en/broadband-strategy-policy [accessed 09 June 2021].

3 Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications.

4 Further performance thresholds (IP packet error ratio, IP packet loss ratio, round-trip IP packet delay, IP packet delay variation and IP service availability) are defined in the BEREC (2020) Guidelines on Very High Capacity Networks, BoR (20) 165, in paragraph 18.

network architectures as being less significant[5] compared to regional cost differences of the individual network architectures,[6] with the latter being the focus of this analysis.

2.2 Data and Methodology

This study builds upon two types of data:

(1) external geographical data.

(2) benchmark data set: VHCN investment figures (FTTH/P2P), derived from bottom-up modelling of 7871 access areas covering whole of Germany.

2.2.1 External geographical data

In the first part of the paper (Chapter 3), regional costs differences in fibre investments are analysed using the internationally standardised **regional classifications** developed by EUROSTAT and BBSR. These concepts establish criteria that classifies administrative areas as rural, urban or intermediate.[7]

The basic external geo-referenced data used in this study consists of:

MPoP (Metropolitan Point of Presences)**:** the MPoP represents the network sided termination point of subscriber access lines of the VHCN. Each MPoP is serving an individual access area. Here, MPoP are represented by MDF-locations of the existing access network in Germany.

Street data: GIS-street layer. Usually, trenches are constructed along streets. Accordingly, we use the street layer as a proxy for assessing trenching costs. Furthermore, this data is used to derive further explanatory variables.

Buildings: all address points in Germany - TOM TOM data with 22.7 million building addresses and in addition a comprehensive household and building data set in raster format.[8] The address points serve as geo-referenced demand locations.

Households: represent demand for VHCN access lines; available at grid cell level.

Settlement areas (build-up-areas)**:** vector data set on settlement areas (GIS-layer).

---

5 We limit our analysis to FTTH/P2P architecture. As shown in Jay et al. (2014): Comparing FTTH access networks based on P2P and PMP fibre topologies, Telecommunications Policy, 38(5), pp. 415-425, other Very High Capacity Network architectures show comparable results.

6 Hoernig et al. (2010), Jay et al. (2011), Jay et al. (2013).

7 Datasets are described and analysed in Sections 3.1.1 and 3.1.2 respectively.

8 Data sources are: WIGeoGIS and AZ-Direct Household and Building Data using 250 x 250 meter raster scale, comprising 41.5 million households.

Administrative areas: GIS-layer of different levels of aggregation (NUTS 1-3) and LAU (local administrative units).

Regiotype classifications: GIS-layer on EUROSTAT and BBSR classification of regiotypes.

Detailed information on this data is displayed in Table 2.1.

Table 2.1: List of primary data sources used to derive the statistical information used in this paper.

Data	Source	Comment
MPoP/MDF - locations	Web source based on DSL enabled MDF-locations list published by Federal Ministry of Economic Affairs and Energy (BMWI); http://selke.de/privates/hvt-standorte/	7971 data records with duplication of addresses, extraction of 7871 MDF-locations
Street data	TeleAtlas status 2018/04	External procurement and processing of TeleAtlas street layer for Germany Release 2018/04
Administrative Areas	Federal Agency of Cartography an Geodesy (BKG)	VG250-EW (Kompakt) Administrative areas of Germany including number of inhabitants, status 01.01.2018
Buildings	TOM TOM - address point data	about 22.7 million building locations
Buildings and Household data	WIGeoGIS and AZ-Direct Household and Buildings Grid data set	41.5 million households in Germany; Grid (250mx250m and 100mx100m), based on Bertelsmann Buildings Data set (BGD), status 2018
Settlement or build-up areas	WIGeoStreet from WIGeoGIS	Vector data set of build-up areas for Germany, status 2018
Urban/Rural Typology Classification	European Union’s Statistical Office (EUROSTAT)	For a detailed description see Section 3.1.1
Types of Settlement Structure	German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR)	For a detailed description see Section 3.1.2 and 3.1.1

The data presented in the Table 2.1 also serves as a basis for deriving further geo-referenced variables that we generate by applying GIS-tools. We use this data in the context of our statistical analysis. They include :

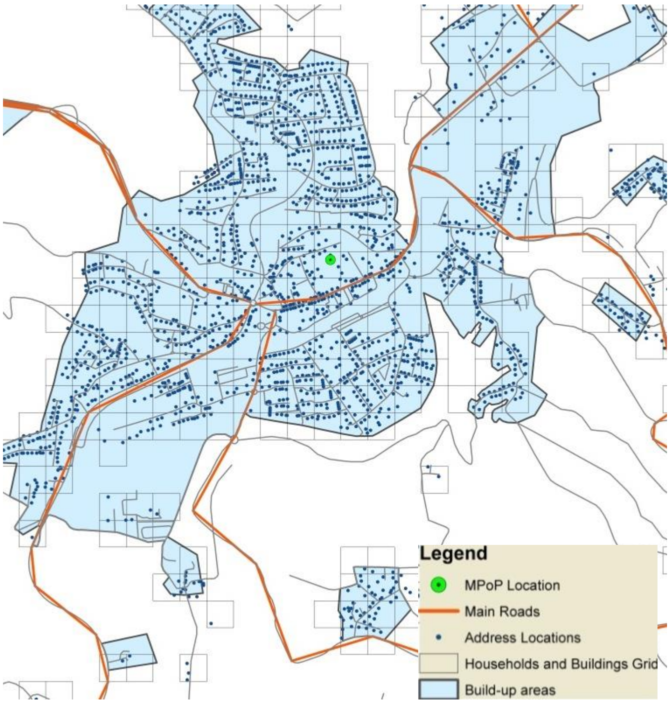
Access areas, which had been delineated by allocating the buildings to the closest MPoP (routing distance according to street layer).

Households per access area / buildings per access area. A geographical layer for the access areas is derived and then used to determine the number of buildings, the number of buildings by building type and the number of households and residents per access area.[9]

Number and size of built-up areas is derived by using the “built-up area layer” (area with buildings) of the geopackage WIGeoStreet from WIGeoGIS. By mapping the access area layer onto the built-up area layer, we are able to determine the number and size of built-up areas (settlement areas) per access area.

Aggregated **street lengths within and outside of settlement** for each access area. This data is derived by intersection of the street layer with the access areas and the built-up areas.

Figure 2.1: Graphical representation of explanatory data based on geographical data.



Source: WIK based on WIGeoStreet (WIGeoGIS; including street data from TeleAtlas and address point data from TOM TOM) and MDF/MPoP locations from the Federal Ministry of Economic Affairs and Energy (BMWI).

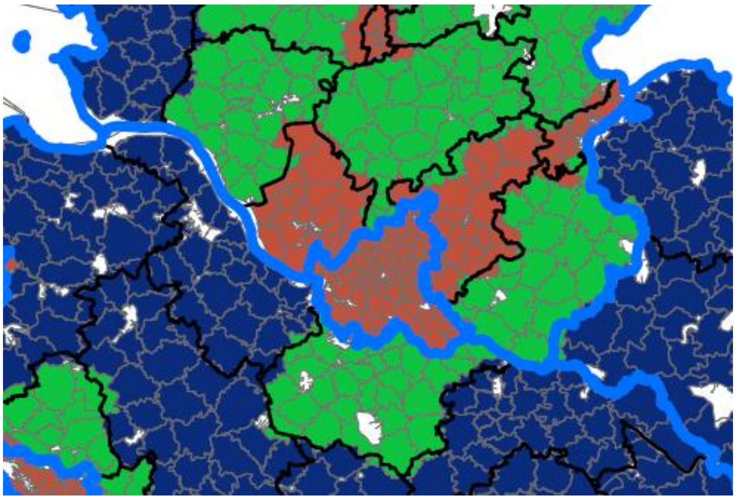
Regiotypes of access areas: The assignment of regiotypes to access areas is derived by intersection of the MPoP-location layer with the NUTS-3 layer (“Kreise”), in order to identify the administrative area(s) of the respective MPoP-locations.[10] Based on this

9 This is done by geographic intersection with the AZ-Direct Household and Building raster data set at the scale of 100m x 100m.

10 For this purpose, we used the product VG250-EW from the Federal Agency of Cartography an Geodesy (BKG).

information, we transferred the regiotype specific label to the access area of the corresponding MPoP location. This methodology allowed to generate a definite regiotype classification of access areas. As a result, each NUTS-3 region contains a group of access areas that, once matched with the EUROSTAT and BBSR data as described above, have the same regiotype class. This is illustrated in Figure 2.2.

Figure 2.2: Territorial demarcation of access areas, NUTS-3 areas and the assignation of EUROSTAT regiotype classifications.



Notes: Shaded areas (orange = "predominantly urban"; green= "intermediate"; and blue= "predominantly rural") for Hamburg and surrounding areas. Access areas delimited by grey lines, NUTS-3 areas by black lines. Source: Delimitation of access areas based on Kulenkampff et al. (2020), NUTS3 delimitation based on cartographical data from the Federal Agency of Cartography an Geodesy (BKG) VG250-EW (Kompakt) and regiotype classifications based on EUROSTAT and BBSR.

2.2.2 Investment figures of Very High Capacity Networks

The investment figures used in this analysis are derived from an emulated, bottom-up modelled access network for Germany. The essential geographical data is taken from the same publicly available source as introduced in the previous section. The modelling is based on a scorched node greenfield approach.[11]

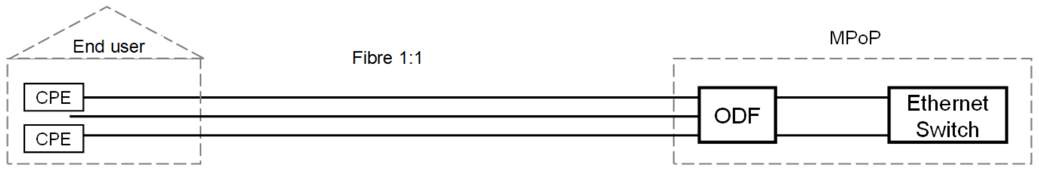
The scope of bottom-up modelling captures the costs for the value chain from the customer premises equipment (CPE) up to network-sided equipment (Ethernet Switch) at the MPoP as illustrated in Figure 2.3 for each individual access area.[12]

The investment was calculated under the assumption of 100% homes passed and 90% homes connected.

11 The underlying study on bottom-up investment calculations also captured FTTC (Fibre to the Cabinet) and FTTS (Fibre to the Street) architectures.

12 Details can be found in Kulenkampff et al. (2020).

Figure 2.3: Demarcation of FTTH/P2P network architecture and investment considered.



The cost figures available are derived from a non-confidential data base only:

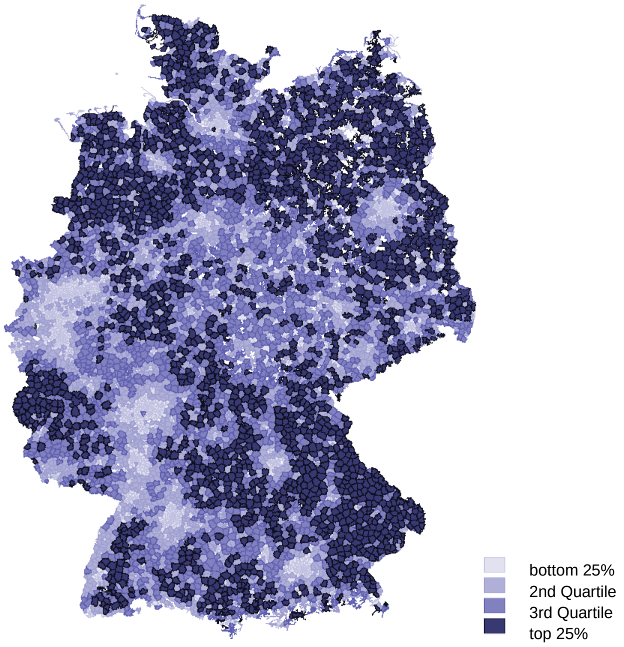
- 7,871 DSL-MDF locations in Germany, which served as scorched nodes (each MDF location establishing an MPoP of an individual VHCN access area).[13]
- Demand and its location was captured by making use of building and household data and address locations
- GIS-street data for the routing tasks.[14]

These bottom-up investment figures are calculated at the level of access areas. The spatial distribution of these investment figures for 7,871 access areas in Germany is depicted in Figure 2.4. A first look at the figure suggests a relationship between the location of some of the least populated areas in Germany and highest costs of fibre deployment per access line, e.g. in the parts of east Bavaria, or regions in the north-west and north-east of Germany. Whether this impression can be supported by statistical analysis under consideration of regiotype classifications is subject of the following sections.

13 It can be argued that FTTH/P2P rollout is not necessarily restricted to the existing layout of copper based access networks, especially if investment is made by alternative operators. Nonetheless, we are confident that our approach is capable of producing reliable results because: (1) MDF-locations, in most cases, still serve as an appropriate MPoP-location due to their placement in the centre of settlement areas, (2) our endogenous delineation of the access area by assigning end-user locations (address data) according to the shortest distance to the MPoP location (and not taking the delineation of the historical copper network architecture).

14 The address points were connected to the streets in Germany (using TeleAtlas street layer for whole of Germany), preparing for the routing tasks in the subsequent network computations. For a detailed description of data preparation and processing see Kulenkampff et al. (2020) appendices A1 and A2, pp.103-121.

Figure 2.4: Bottom-up modelled FTTH/P2P investments per access line and access area.



3 Regional cost differences

3.1 Regional classifications and their application to Germany

We consider the following two EU-wide standardised methodologies in this study:

- **EUROSTAT: Urban-Rural Typology Classification** of the European Union's Statistical Office.[15]

- **BBSR: Settlement Structure Classification** of the German Federal Institute for Research on Building, Urban Affairs and Spatial Development.[16]

Both regiotype classifications apply to territorial demarcations at a disaggregated level; at higher levels of aggregation they are aligned with the NUTS demarcations.[17]

15 https://ec.europa.eu/eurostat/web/rural-development/methodology, version: NUTS-2016, [accessed 22 Feb 2020].

16 BBSR - Raumbeobachtung - Downloads (bund.de), [accessed 06 May 2021].

17 The NUTS (nomenclature of territorial units for statistics) classification is a hierarchical system dividing up the economic territory of the EU for the purpose of generating harmonized regional statistics. It is worth noticing that "Degree of Urbanisation", which was proposed for endorsement by European Commission is available only at a higher level of granularity, see: European Commission – Eurostat and DG for Regional and Urban Policy – ILO, FAO, OECD, UN-Habitat, World Bank (2020). Yet, NUTS-3 level is the minimum level of territorial demarcation required to uniquely match either regiotype classification with the corresponding bottom-up figures produced at the level of access

3.1.1 EUROSTAT Urban/Rural Typology

The EUROSTAT *urban-rural typology* is a subnational statistic of the EU, which relies on population density information at a high level of granularity, namely grid cells of 1 km², in order to classify administrative units at higher levels of territorial aggregation: NUTS-3 regions.

According to EUROSTAT (2018) *Methodological Manual on Territorial Typologies*, continuous grid cells of 1 km² are defined as "urban clusters", if (1) their population density is at least 300 inhabitants per km², and (2) they have a minimum population of at least 5000 inhabitants. Rural areas are then all other areas outside urban clusters.

On basis of this 1 km² grid cell classification, larger territorial units can be classified. For the *urban-rural typology* classification statistic, this is applied on NUTS-3 administrative regions. Depending on the share of population that lives within the identified "urban clusters" of the respective administrative region, they are attributed as:

- **predominantly urban,** if the share of population that is living in urban clusters is over 80%,
- **intermediate,** if the share lies between 50%-80%,
- **predominantly rural,** if the share is lower than the 50%.[18]

3.1.2 BBSR Types of Settlement Structure

BBSR types of settlement structure[19] is an alternative regional classification methodology It classifies regions into four groups: "metropolitan", "urban", "rural" or "sparsely populated" based on (1) the share of city population (defined at LAU-level)[20] and (2) population density. The classification is derived from information available at the level of local administrative units (LAU). This is shown in Table 3.1.

---

areas. At the next level of granularity, local administrative units (LAU) - which in Germany correspond to the local municipalities ("Gemeinden") - do not uniquely identify access areas. This is reflected in the fact that there are more local municipalities (11 087) than number of access areas (7871) in our data set, compared to the number of NUTS-3 regions (401) in Germany. In the case of Germany, the NUTS-3 level corresponds to the national demarcation of district areas ("Kreise"/ "kreisfreie Städte").

18 In order to account for the presence of a city, EUROSTAT reclassifies "predominantly rural" regions containing a city of more than 200 000 inhabitants as "intermediate". Similarly, "intermediate" regions containing a city of more than 500 000 inhabitants become "predominantly urban". In both cases, the city must represent at least 25 % of the region's total population.
(see EUROSTAT 2018, Methodological manual on territorial typologies).

19 BBSR, see Siedlungsstrukturelle Regionstypen Europas.

20 This classification methodology requires the definition of "large or large medium-sized cities". According to the BBSR classification of cities and municipalities, "large cities" are local municipalities ("Gemeinden") with at least 100 000 inhabitants and at least a mid-central function; whereas "large medium-sized cities" are local municipalities with at least 50 000 inhabitants and also at least a mid-central function.

Table 3.1: BBSR regiotype classification criteria.

	Share of population living in large or large medium-sized cities (%)	Population density	
		inhab/km²	applied to
Metropolitan	-	>300	city inhab < 300 000
	-	>150	city inhab > 300 000
Urban	>45	>150	-
	<45	>150	areas without cities
Rural	>45	<150	-
	<45	>100	-
Sparsely populated	<45	<100	-

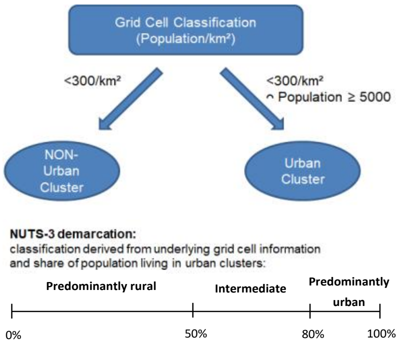
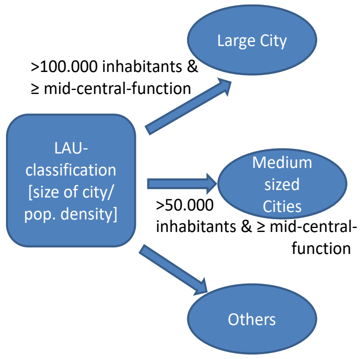
Notes: Regions at NUTS-3 level. BBSR criteria: (1) share of population living in “large or large medium-sized” cities (defined at LAU-level) and (2) population density. Source: WIK based on BBSR regiotype classification criteria.

3.1.3 Classification Comparison

On the basis of these regiotype classifications, we compare EUROSTAT and BBSR methodologies, in order to identify to what extent different concepts lead to diverging categorisations in Germany. In a first step, we make this comparison at the level of access areas. In a second step, more aggregated levels (NUTS-3) are considered.

Contrast of concepts of regional classification:

Table 3.2: Comparison of EUROSTAT and BBSR classification concepts.

	EUROSTAT	BBSR
Database for initial assignation	Grid cells of 1 km^2	Local administrative units (LAU)
Criteria for NUTS-3 level assignation	Population density (based on grid cell categories)	mixed criteria: 1) population density (inhab./km^2) 2) share of population living in large or medium sized cities (%)
Number of regiotypes	3 ("predominantly urban", "intermediate", "predominantly rural")	4 ("metropolitan", "urban", "rural", "sparsely populated")
		

Source: WIK based on EUROSTAT and BBSR regiotype classifications

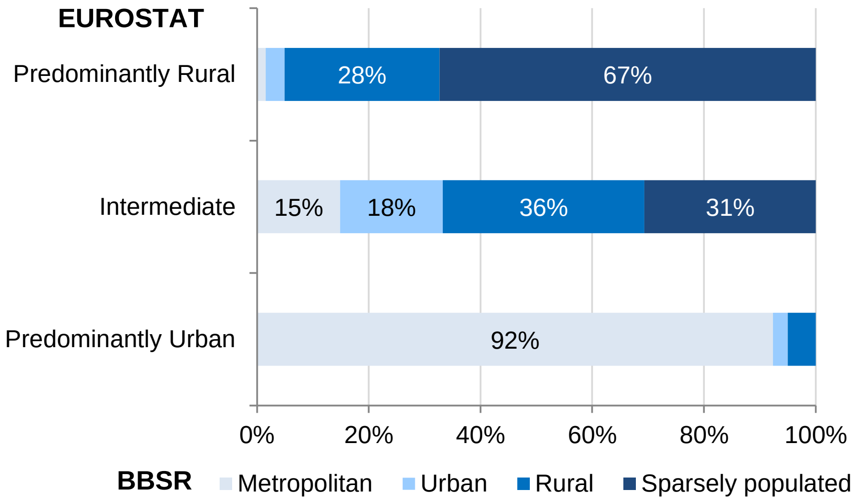
Classification of access areas:

With regard to EUROSTAT, our data base for Germany shows that 2115 access areas are considered urban, 3629 intermediate and 2113 rural.

Figure 3.1 displays the overlap between both classification methodologies, which can range from 0% to 100%, where 100% means that all the areas identified by EUROSTAT as “rural”, “intermediate” or “urban” are also identified by BBSR under an equivalent regiotype class. It reveals that from the total of 2115 access areas classified in EUROSTAT as “predominantly urban” regions, most of them (92%) correspond to BBSR “metropolitan” regions. This reflects a relative high level of congruency among both classification methodologies regarding the identification of the most urbanized regions.[21]

21 As the BBSR has more access areas with the highest urban class compared to its counterpart EUROSTAT, from the BBSR perspective, there is a significant number of access areas in “metropolitan” regions that cannot be found in EUROSTAT “predominantly urban” regions. There are 541 access areas, which – under EUROSTAT classification – are instead listed in “intermediate” regions, making up 15% of this category.

Figure 3.1: Comparison of the assignation of NUTS-3 regions (N=401) and their corresponding access areas (N=7871) according to the EUROSTAT classification compared to the BBSR classification.

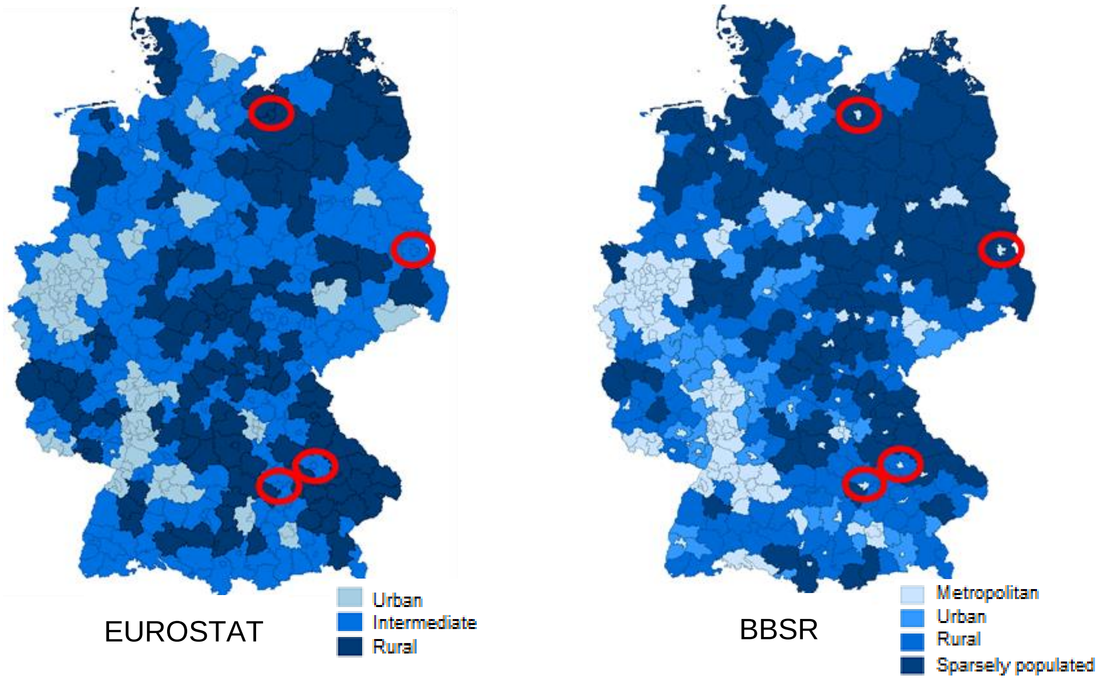


Source: WIK based on EUROSTAT and BBSR regiotype classifications. Delimitation of access areas based on Kulenkampff et al. (2020).

For the remaining categories, we reveal differences in the classification between EUROSTAT and BBSR (see "intermediate" and "predominantly rural" in Figure 3.1). It does not seem to exist a clear consensus regarding the identification of rural regions in Germany. For instance, there is a noticeable number of access areas (1115) assigned to "sparsely populated" regions under the BBSR classification, whereas in EUROSTAT, they are considered to be within "intermediate" regions. They account for 31% of all access areas of this category.

By comparing the geographical presentation of EUROSTAT with the respective BBSR map (Figure 3.2) at NUTS-3 level, it becomes clear how some of the findings from the numerical analysis are reflected in a geographical dimension. For instance, BBSR classifies several districts in east Germany (mostly surrounding Berlin) as "sparsely populated" regions, while EUROSTAT considers them to be "intermediate" regions. This is explained by the different definition of "sparsely populated" regions applied by BBSR compared to EUROSTAT "predominantly rural". BBSR classifies these eastern districts as "sparsely populated" given that less than 45% of their population lives in "large or large medium-sized" cities. Furthermore, they display an average population density below 100 inhabitants/ km^2. In contrast, EUROSTAT classifies the same districts as "intermediate regions" since between 50% to 80% of their population lives in "urban clusters", which they defined as 1 km^2 grid cells with population densities of at least 300 inhabitants per km^2 and a minimum population of at least 5000 inhabitants.[22]

Figure 3.2: Geographical comparison of regiotype classifications (based at NUTS-3 level, N=401) between EUROSTAT and BBSR.



Notes: Selective illustration of 'small' NUTS-3 areas (red circles). NUTS-3 delimitation is based on cartographical data from the Federal Agency of Cartography an Geodesy (BKG) VG250-EW (Kompakt).

Figure 3.2 highlights a further systematic difference between these two methodologies. This difference refers to NUTS-3 regions in Germany that represent "autonomous cities" (or "kreisfreie Städte").[23]

The reasons for this difference is that EUROSTAT adjusts small regions (or "autonomous cities" in Germany) by using a size-of-area-related criterion, while BBSR adjusts small regions by applying a population-related criterion. This simple

22 See Sections 3.1.1 and 3.1.2. This outcome is not only caused by different threshold values but also by fundamental methodological differences: EUROSTAT methodology classifies 1 km^2 grid cells first and applies further criteria for classifying NUTS-3 areas, whereas BBSR takes higher levels of territorial aggregation as a starting point for its classification (LAU-level).

23 Autonomous cities account for 107 from its 401 NUTS-3 regions. While BBSR considers "autonomous cities" to be the highest urban class, EUROSTAT only considers less than a half of "autonomous cities" to be "urban" (51). For example, "autonomous cities" like Freiburg (south-west of Germany), Kassel (centre of Germany) or Regensburg (south-east of Germany) are not considered to be "predominantly urban" regions by the EUROSTAT classification.

methodological difference is responsible for relevant differences in the classification outcome of 51 out of the 401 NUTS-3 regions in Germany.[24]

3.2 Fibre investment heterogeneity under application of the EUROSTAT and BBSR regional classifications

In this section, we make a statistical assessment of regiotype-specific investments in VHCN. For this purpose, we match the available bottom-up modelled investment figures with the internationally standardised EUROSTAT and BBSR regiotype classifications.

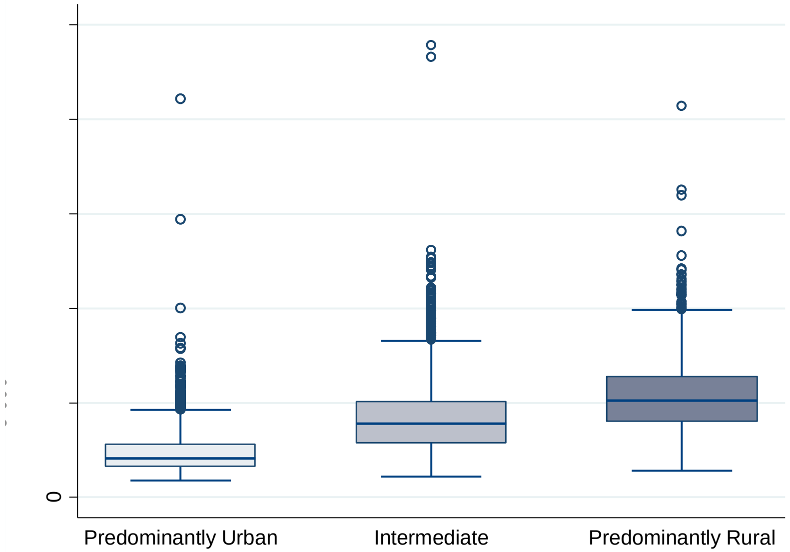
3.2.1 Differences between-regiotypes

Under EUROSTAT classification, “predominantly urban” regions in Germany have an average investment requirement of 2353 € per access line. This is based on 2115 access areas, and accounts for 27% of the total access areas in Germany. For “intermediate” regions, which comprise 46% of all access areas, the average investment requirement increases to 4154 € per access line. For “predominantly rural” regions, the average investment requirement rises further up, reaching 5316 € per access line, making fibre deployment in this region, on average, more than two times more expensive than in “predominantly urban” regions. These expensive regions comprise 2120 access areas, accounting for 27% of all access areas in Germany. The summary statistics and box-plot charts with the data distribution according to EUROSTAT and BBSR classifications are presented in Figure 3.3.

24 According to the BBSR data for Germany, these small NUTS-3 regions (or “autonomous cities”) were not combined with any neighbouring regions, keeping their original classification. This is because BBSR applies this adjustment for regions that are not necessarily small in surface but in population (below 100 000 inhabitants), and all these cities have a population above 100 000 inhabitants.

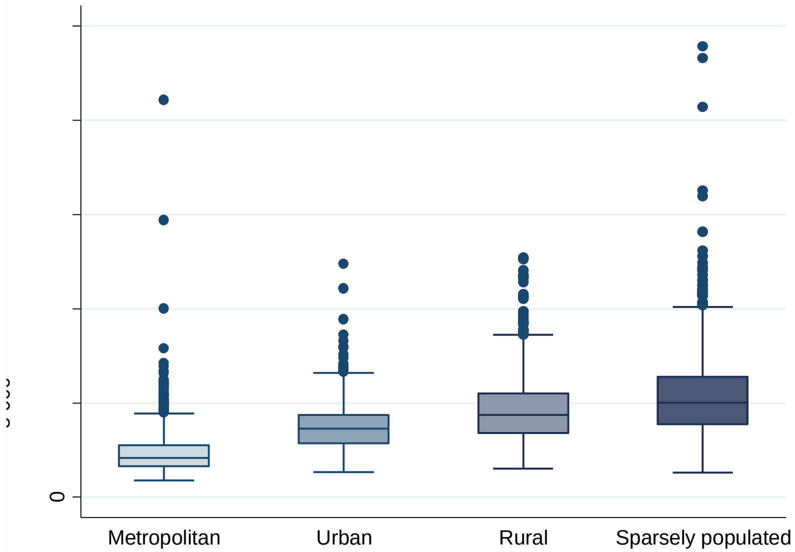
Figure 3.3: Average values and statistical distribution of investment figures per access line and access area grouped by EUROSTAT and BBSR regiotype classifications.

EUROSTAT



	Predominantly Urban	Intermediate	Predominantly Rural
Nr. of access areas	2 115	3 636	2 120
in %	27%	46%	27%
Mean (in €)	2 353 €	4 154 €	5 316 €
Standard deviation (in €)	1 186 €	1 781 €	1 893 €
min. value (in €)	904 €	1 109 €	1 403 €
1% percentile (in €)	961 €	1 360 €	1 895 €
99% percentile (in €)	6 121 €	9 505 €	10 848 €
max. value (in €)	21 092 €	23 922 €	20 689 €

BBSR



Metropolitan	Urban	Rural	Sparsely Populated
2 527	796	2 005	2 543
32%	10%	25%	32%
2 302 €	3 760 €	4 579 €	5 254 €
1 056 €	1 276 €	1 597 €	2 019 €
904 €	1 336 €	1 523 €	1 300 €
967 €	1 486 €	1 806 €	1 878 €
5 479 €	7 957 €	9 158 €	10 867 €
21 092 €	12 378 €	12 718 €	23 922 €

In contrast to EUROSTAT, BBSR highest urban class “metropolitan” comprises a higher number of access areas (32% instead of 27% under EUROSTAT classification), but shows slightly lower average investment per access line of -2% (BBSR: 2302 €, EUROSTAT: 2353 €).[25] In line with EUROSTAT, with each BBSR regiotype class, investment requirements increase progressively reaching, on average, 5254 € per access line in BBSR “sparsely populated” regions. Because of the additional fourth regiotype class, BBSR shows a more smooth increase in average investment per line from the most dense “metropolitan” to “sparsely populated” class whereas EUROSTAT reveals a noticeable gap between “predominantly urban” and “intermediate” areas.

3.2.2 Differences within regiotypes

As illustrated in the box-plot charts in Figure 3.3, there is a noticeable spread range within each regiotype class. This holds for EUROSTAT as well as for BBSR. It is most obvious for EUROSTAT regiotype classification within the “predominantly urban” cluster: the maximum invest per access line (21 092 €) is 9 times higher than the average of this cluster (2353 €) and even 23 times higher than the minimum investment value (904 €). This pattern applies, to almost the same extent, to all regiotype classifications shown in Figure 3.3, including those derived from the BBSR classification. Even without taking extreme values as basis for this analysis, the spread within both, EUROSTAT and BBSR classifications, is still substantial, with the most expensive access areas (at the 99% percentile) being 5 to 7 times higher than the most cheapest access areas (at the 1% percentile) of their own group.

3.3 Measuring goodness-of-fit of EUROSTAT and BBSR regional classifications at NUTS-3 and access area level with respect to fibre investments

In order to evaluate, to which extent the aforementioned regional classifications developed by EUROSTAT and BBSR are suited for explaining regional differences in fibre investment requirements, we measure to which extent EUROSTAT and BBSR regional classifications are able to predict the variations we observe in fibre investment requirements data.

For this purpose, we quantify the *goodness-of-fit* following the R-squared measure, which serves as an indicator (ranging from 0 to 1) reflecting the proportion of data that

25 It should be noticed that BBSR “metropolitan” regions show lower average investment figures compared to its EUROSTAT counterpart “predominantly urban”, although BBSR “metropolitan” regions gather a larger number of access areas. This is due to the composition and not to the quantity of access areas, as BBSR and EUROSTAT regiotype classes consist more or less of different regions with different levels of investment (see Figure 3.3). Neither of these two classifications perfectly assign a regiotype with investment figures following a steadily increasing investment function, as evidenced by the wide spreads shown in Figure 3.3. Thus, the average investment of a regiotype class depend less on the number of access areas and more on its composition.

is explained (or fitted) by the model. Here, R-squared is estimated by regressing logged fibre investment per access line (in €) on dummy-variables for each regiotype class.[26] Our analysis covers EUROSTAT as well as BBSR urban/rural classifications.

The *goodness-of-fit* at both levels, access areas and NUTS-3,[27] is presented in Table 3.3. Of course, we expect that the aggregation of access areas onto NUTS-3 level reduces the variation ("information loss") in the investment figures, balancing the investment heterogeneity of access areas within each NUTS-3 region. For BBSR, we can clearly observe the expected increase of R-squared (from 0.48 to 0.72), whereas EUROSTAT only shows a slight increase (from 0.41 to 0.48). Thus, NUTS-3 level produces a higher value of R-squared compared to the level of access areas. Apparently, BBSR classification is better suited for explaining the regional differences in fibre investment requirements in Germany as compared to EUROSTAT. This might be explained by methodological differences in the assignation of 'small' NUTS-3 regions analysed in Section 3.1.3.

Table 3.3: Comparison of the goodness-of-fit (R-squared) at different aggregation levels (NUTS-3 vs. access area) and under different regional classifications (EUROSTAT vs. BBSR).

	NUTS-3		Access Areas	
	EUROSTAT	BBSR	EUROSTAT	BBSR
R-squared	0.4443	0.7225	0.4077	0.4804
Nr. of observations	401	401	7871	7871

3.4 Summary of heterogeneity of VHCN investment from a regiotype classification perspective

The first part of the analysis in this section illustrates the magnitude of geographical heterogeneity in investment requirements for fibre deployment at 100% homes-passed in Germany, and describes how imprecise current NUTS-3 level regiotype classifications are in capturing regional investment variations either at the level of access areas or even at NUTS-3 level. The findings can be summarised as follows:

- Among comparable regiotypes, both classifications produce comparable average values of investment requirements.

26 The model includes a constant.

27 In order to determine the average investment requirement per access line at the level of NUTS-3, investment values for each access area are summed at the NUTS-3 level and then divided by the corresponding number of access lines within the NUTS-3 region.

- Under both classifications, the average values for each regiotype are statistically significantly different ($p<0.000$) from each other.
- There is evidence that regiotype classifications employed by EUROSTAT and BBSR are, on average and overall, consistent with the expected progression in investment requirements from the most urban to the most rural areas.
- Nonetheless, the investment heterogeneity observed within each category is large (see EUROSTAT and BBSR quartile distribution, median, minimum and maximum for each regiotype category depicted in the box-plot charts in Figure 3.3).

4 Statistical estimation models

In the following, we outline our statistical approach for estimating fibre investment requirements. We start with a literature review and continue with our assessment of cost drivers we intend to consider in our estimation model.

Our model is entirely based on publicly available georeferenced data only and relies on a complete sample for Germany (7871 access areas). The results are controlled with regard to interfering neighbouring effects ("spatial autocorrelation") and are assessed regarding its predicting power.

The bottom-up modelled invest per VHCN access line figures serve as a benchmark to assess the fit of the model we are suggesting.

4.1 Geographical factors influencing fibre investment requirements – recent literature

NGA networks entail high fixed costs principally from intensive trenching and display therefore diminishing unit costs. According to OECD (2011), fibre deployment is only feasible or profitable in areas where potential demand is high and concentrated.[28] But obviously, these characteristics are not sufficient in order to explain regional differences in VHCN investment needs.

In recent empirical studies, Fourberg and Korff (2020) outline that geographical factors such as population density, municipality area and ground ruggedness relate to the expansion of fibre projects. Similarly, Sahebali et al. (2021) claim that regions with low investment levels are characterized by lower density of houses, greater distance to the next house, longer distance to backbone and more water bodies.

28 In this line, population density and loop length are often seen as main cost drivers. See also OECD (2011-06-20), "Next Generation Access Networks and Market Structure".

Mitcsenkov et al. (2013) argue that in presence of an uneven population density and irregular street system, the use of models based on area-wide average parameters only, without considering their local characteristics, tend to lead to inaccurate estimates, especially in non-urban regions where the population density fluctuates highly within an area. This is mostly the case in absence of geospatial information.[29]

Also Phillipson (2015) shows that population density alone loses precision in rural areas as households tend to be less evenly distributed.

Kulenkampff et al. (2020) identify explanatory variables like “subscriber density in built-up areas” and “trench length in feeder-cable segment” as the most relevant network-based elements driving the costs of fibre-based access networks.

4.2 Assessment of geographical cost drivers for VHCN investment on the basis of data available

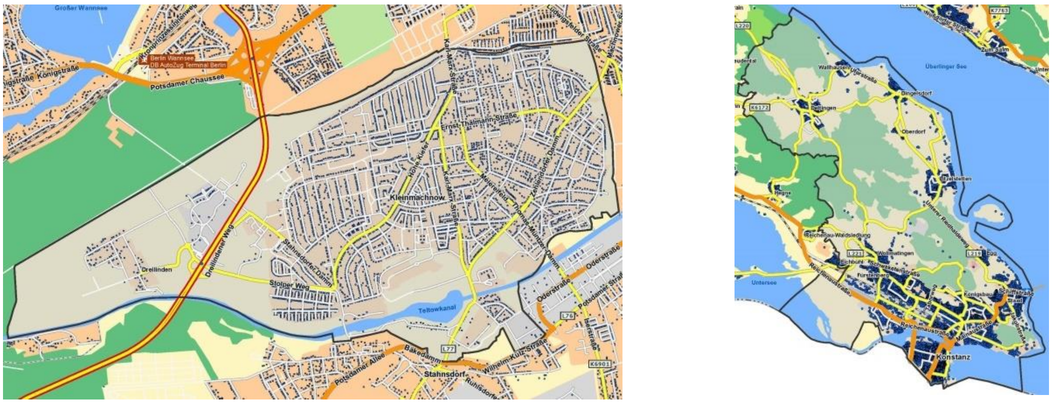
Given the findings from recent literature, and based on our knowledge on cost of access networks, the **number of access lines** and **trench length** constitute essential cost drivers. In the following, we analyse the suitability of publicly available data for the purpose of VHCN invest estimation. For this purpose, we use the investment figures outlined in Section 2.2.2.

Access area vs built-up-area:

Having started with a simple density measure (access lines per access area), we noted that access areas with almost identical densities (lines/km^2) showed a noticeable spread in VHCN invest per line. This is illustrated in Figure 4.1, displaying the municipalities of Konstanz and Kleinmachnow, both with identical densities, but with the latter showing an invest per line which is about twice the invest of Konstanz. This finding suggests to rather use density figures relating to built-up-areas instead of density figures relating to the total access area.

29 The FTTH Council (2017) implemented a cost model using real GIS-data to optimize a network design for some regions to then extrapolate the resulting fibre costs to the remaining regions based on an statistically estimated cost/density relationship. In contrast to FTTH Council (2017), our statistical estimation is based on complete sample (all access areas in Germany). More importantly, our approach shifts the focus from an area-based density to a street-based density and considers multiple additional factors that reflect the heterogeneity of settlement structures.

Figure 4.1: Settlement structure of municipalities: identical density but doubled invest per access line.



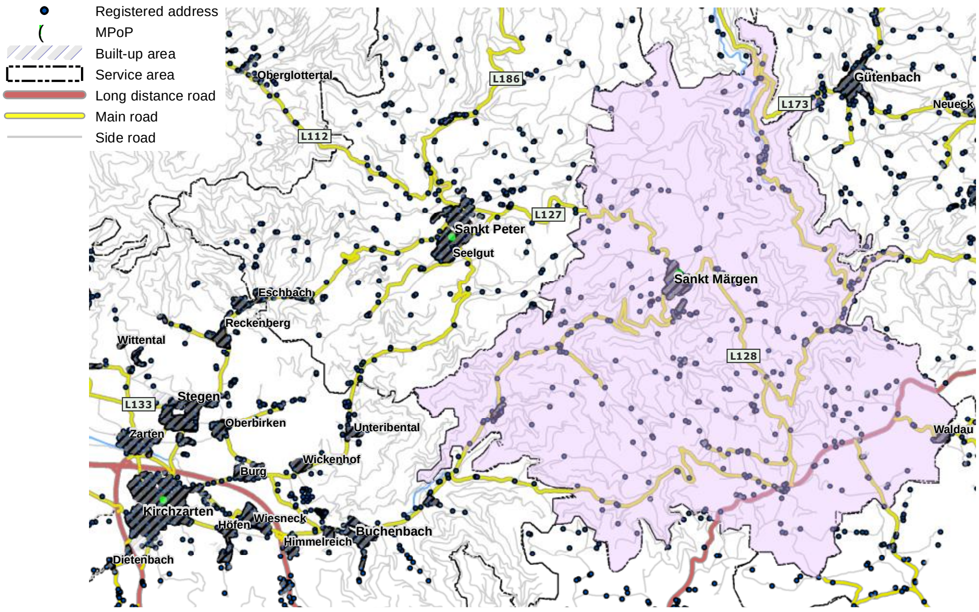
Kleinmachnow *Konstanz*

Source: WIK based on geographical data from WIGeoStreet.

Appropriate type of street length:

Furthermore, how access line demand is distributed within the access area is relevant in the determination of costs per access line (see locations displayed as blue dots in Figure 4.2).

Figure 4.2: Settlement structure: built-up areas, road types and household distribution



Notes: Demarcation at the level of access areas using WIGeoStreet-Data and MDF/MPoP locations from the Federal Ministry of Economic Affairs and Energy (BMWI) for the Sankt Märgen Area, Baden-Württemberg, Germany.

Figure 4.2 illustrates the aforementioned georeferenced data in a map for one of the most expensive access areas in Germany, according to our data. It depicts the uneven distribution of households within access areas. Access line demand is scattered over the whole access area. Within this access area, the village of Sankt Märgen represents a (relatively small) built-up area. This example reveals that the aforementioned indicator "number of access lines per built-up-area" is not sufficient in order to explain regional cost differences. In case of our example outlined in Figure 4.2 the street length seems to be an important variable, too, which should be considered as a cost driver. Within built-up areas, most streets are relevant for trenching, thus, the number of households per kilometre of street seem to be an adequate cost factor in these areas. Yet, in areas with highly disperse households not all streets are relevant, as depicted in Figure 4.2. To make use of the available GIS-data on street length (see Section 2.2.1), selections have to be made, in order to consider only those types of streets that are of importance for trenching in areas with widely spread households. Here, we selected main roads as most suitable dispersion variable and thus neglected all other types of roads like highways, side roads or agricultural or forest pathways.[30]

Appropriate density definition:

Regarding population density as a cost driver, we investigate in the appropriateness of different density definitions.

On the basis of our available data, we generated a complete sample of all three density variables and determined their R-squared statistic:

- number of households per land area: 0.698
- number of households per built-up area: 0.725
- number of households per street-kilometre in built-up area: 0.8593

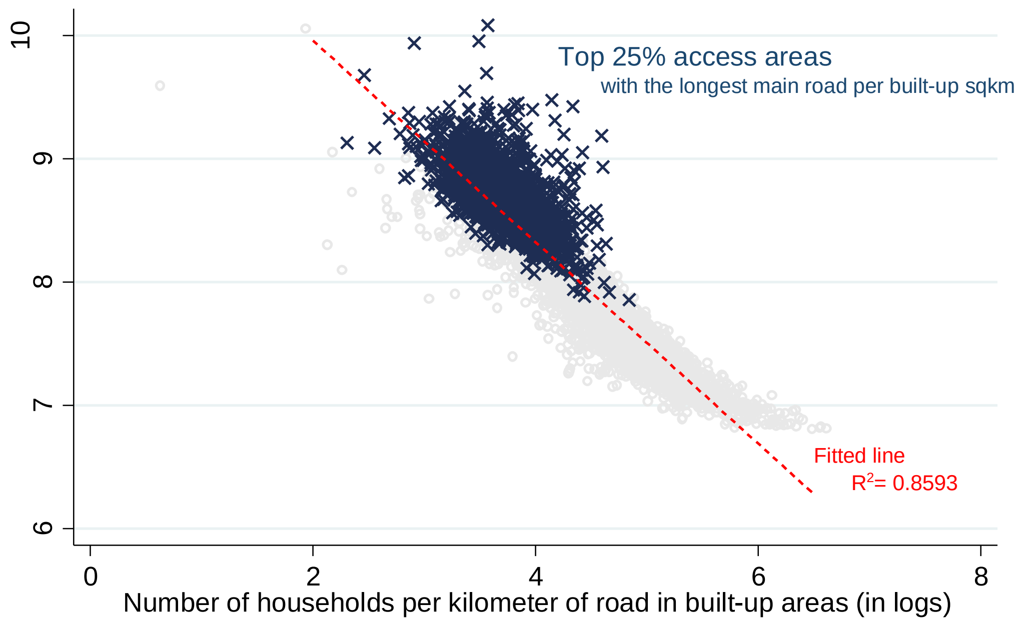
The proportion of cost variation that is explained by *the number of households per kilometre of road in built-up areas* amounts to 85,9% (R-squared 0,859). This is significantly higher than the proportion of cost variation explained by the traditional area-wide based household densities often used in the literature (see R-squared statistics listed above).[31]

Similarly to Mitcsenkov et al. (2013), we also expect that this approach increases the precision of estimating network costs.

30 As shown in Figure 4.2, not all streets are of relevance for network deployment. Fortunately, the differentiation of three street types (long distance roads (in red), main roads (in yellow) and side roads (in grey)) allows us to focus on road types that best capture, to a certain degree, the roads, which are most relevant for a network roll-out within the access area. Main roads typically are better suited in connecting local communities. For VHCN investment estimations this is particularly important in thin populated areas where we find main roads connecting towns to its neighbours (Phillipson, 2015).

31 FTTH Council (2017) uses the number of households per built-up area. Fourberg and Korff (2020) and Sahebali et al. (2021) use area-wide based population densities.

Figure 4.3: Relationship between street-based household density, main road length and fibre investment costs (in natural logarithm. N= 7871 access areas).



In addition, we test the hypothesis that the main road length has particular importance in less urbanized regions. Actually, those access areas with the longest main road per km^2 of built-up areas (top 25%) have the highest fibre investment requirements per access line (see the blue crosses in Figure 4.3). This finding supports our hypothesis that access areas with dispersed households, villages and built-up areas, measured by the relative length of main roads, tend to be more expensive to deploy, as they require longer loops to connect the demand within a km^2 compared to less disperse areas.

4.3 A spatial regression model for fibre investment requirements

In this section, we combine the analysed geo-referenced factors in a multiple regression model. We establish a model with multiple costs drivers, in order to improve the prediction quality of geographical differences of VHCN investment requirements per line. For that purpose, we determine first which set of variables and empirical model fits jointly best[32] and identify which geographical factor among them is most influential on investment requirements. Then, we test the strength of the estimated relationships, and finally evaluate its predicting power. With additional control mechanisms we take account of neighbouring effects, which due to autocorrelation may degrade the quality of our results.

32 Kulenkampff et al. (2020) identify the most relevant network elements of FFTH/ P2P access networks in Germany based on bottom-up measures granting a technical basis for the empirical estimation.

4.3.1 Regression model design: four-factor-regression-analysis

On the basis of the aforementioned studies and analysis of cost drivers, we selected the following explanatory geographical variables:

The first variable ***number of access lines*** is included with the explicit purpose of controlling for differences in population size across different access areas.

The second explanatory variable ***number of households per kilometre of road in built-up areas*** captures the density effect (or "economies of scale") on fibre investment requirements particularly in the drop-cable segment. This density is a street-based measure – in contrast to the traditional area-based measures. In the preliminary assessment shown in Figure 4.3, this variable has proven to be closely related to fibre investments in absence of any other explanatory variable.

The third and fourth explanatory variables refer to the ***main road length per built-up area***, which reflects the degree of household dispersion in an access area. As shown in Figure 4.3, access areas with longer length in their main roads relative to their built-up areas are associated with higher fibre investment requirements.[33]

The last explanatory variable ***share of built-up area to total area*** is a measure of the degree of urbanisation of an access area. Whereas it is true that dense areas tend to be more urbanized, this is not necessarily the case for a wide range of less dense areas, as the degree of urbanisation, expressed as a percentage from total access area, has a more spatial focus and therefore supplements additional information in the case of an uneven density distribution.

The log-log spatial error model (SEM) is outlined in the following equation:[34]

$$\begin{aligned}\ln(FTTH\ P2P\ Invest)_i = \beta_0 + \beta_1 \ln(number\ of\ access\ lines)_i &+ \\ \beta_2 \ln(household\ per\ km\ of\ road\ in\ built\ up\ areas)_i &+ \\ \beta_3 \ln(main\ road\ length\ per\ built\ up\ area)_i &+ \\ \beta_4 \ln(main\ road\ length\ per\ built\ up\ area)^2{}_i &+ \\ \beta_5 \ln(share\ of\ built\ up\ area\ to\ total\ area)_i &+ u_i\end{aligned} \tag{I}$$

with $u_i = \rho W_i u_j + \varepsilon_i$

u_i treats the spatial dependence as a nuisance[35]; and ε_i is the independently and identically distributed error. *W* is the spatial weight matrix. The subscript *i* denotes a territorial demarcation at the level of access area.

---

33 A quadratic term allows to capture the exponential progression of the relationship between household dispersion and fibre investment requirements (after controlling for the differences in number of access lines within an access area).

34 Estimations from a log-log model (see betas in Equation I) can be interpreted as elasticities.

Alternative or additional factors not specified in Equation (I), which potentially influence the required level of investment in Germany, are investigated within a robustness analysis.[36] Further details of this analysis can be drawn from Table 7.3 in the technical Annex.[37]

4.3.2 Regression analysis and preventing interference of neighbouring effects

In order to control for spatial error interactions and to prevent their interference in the estimation and performance of the model, ***disturbances of neighbouring access areas*** that are spatially dependent on each other (see u_j in u_i in Equation I)[38] are considered in the spatial error model. Of course, this methodology requires information about the spatial contiguity of all access areas in Germany, which we generated through a spatial weight matrix (denoted with W in u_i in Equation I). Figure 4.4 presents the graphical illustration of our definition of neighbourhood.

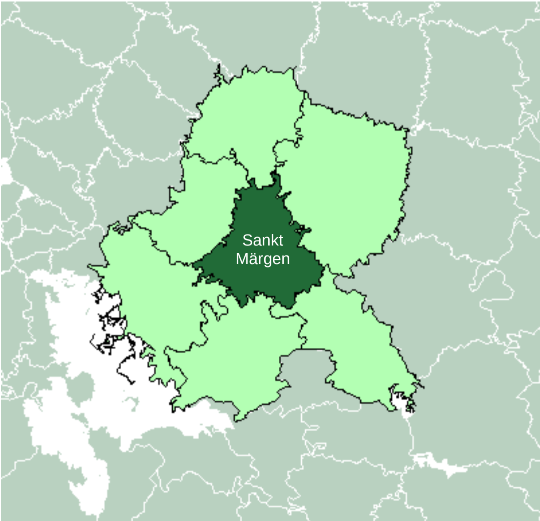
35 The spatial dependence is treated as a nuisance (disturbance), in the sense that it reflects spatial autocorrelation in i) measurement error or ii) in variables that are otherwise not crucial to the model (Anseling and Bera, 1998, p. 249).

36 This includes several bottom-up emulated network parts such as house distance to the street, number of distribution points, number of sleeves per km, the average cost of trench per meter (in drop- and feeder segment), the deviation of feeder length from the regional average and the ratio of feeder- to drop-segment length, etc. Their impact on the specified model is reported in Table 7.3 in the technical Annex.

37 Results from the robustness analysis suggest that the identified variables in Equation (I) are relatively robust against the inclusion of additional and alternative covariates. For forty-one out of the forty-eight estimated coefficients, the maximal coefficient variation is not higher than +/-10%.

38 Sahebali et al. (2021) empirically investigate the spatial effects driving very high capacity fibre-based network roll-out in the Netherlands using explorative spatial data analysis. They found that fibre networks are based on geographical clustering, taking place as a neighbourhood driven-process.

Figure 4.4: Graphical illustration of the contiguity weight matrix with neighbours of first order.



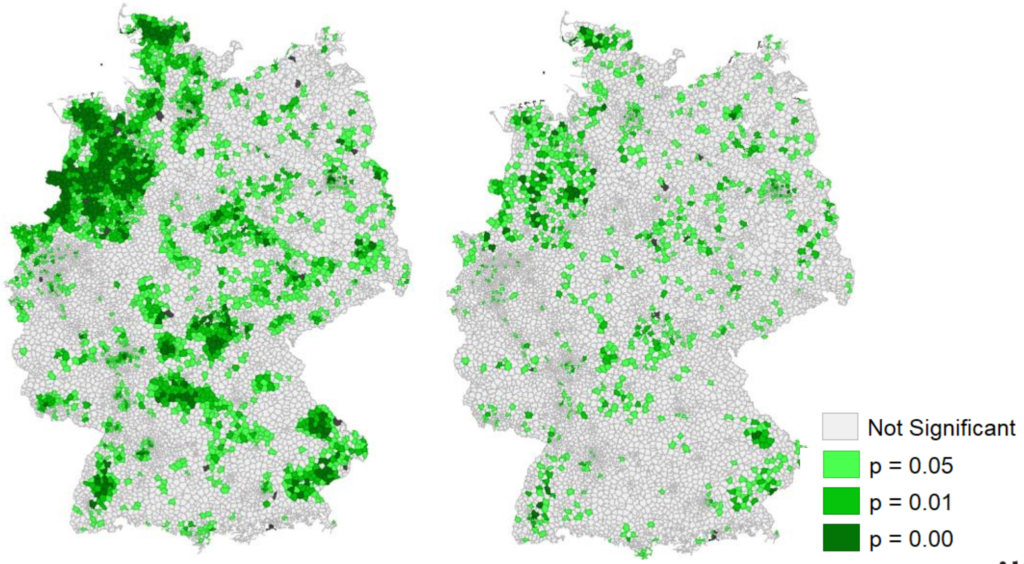
Notes: Delimitation of access areas based on Kulenkampff et al. (2020). Illustration generated using the software program GeoDa™.

By using SEM,[39] we expect to minimize the detrimental effects of spatial autocorrelation in the residuals. In order to verify the improvement of results we make use of *"Local Moran's I Test"*.[40] The findings are displayed in Figure 4.5. It illustrates the magnitude of the spatial autocorrelation between neighbouring access areas that emerges from the unexplained part of a model (residuals highlighted in green). The comparison of the initial results before implementation of the SEM (left) and after (right) the implementation of SEM suggests a noticeable amount of spatial autocorrelation in the residuals.

39 Other spatial models (see Pisati, M., 2012), e.g. spatial autoregressive models (SAR) or spatial autoregressive models with spatial autoregressive disturbances (SARAR) were tested. In contrast to the statistical significance of ρ in the spatial error model (SEM), the statistical significance of the autoregressive terms in other spatial models mentioned above are very unstable and sensitive to the specified model. Indeed, in this case, there is no theoretical background supporting the use of models with autoregressive spatial effects (such as SAR and SARAR) as it would not be consistent with the data generating process of the depend variable (as the bottom-up emulated investment requirements per access area do not explicitly consider an spatial interaction or spill-over effect among access areas).

40 We implement the Local Moran's I Test using the software program GeoDa™ (see Anselin et al., 2006 and Anselin, 1995).

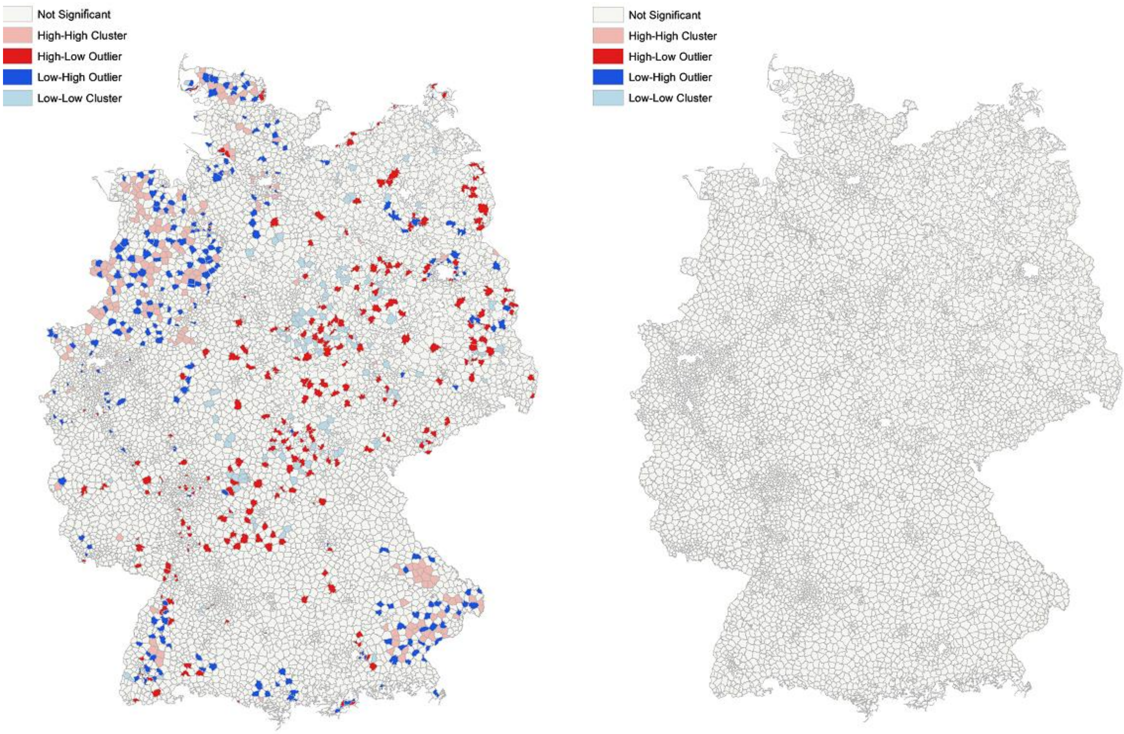
Figure 4.5: Significance Map: spatial autocorrelation of OLS-residuals (left figure) and SEM-residuals (right figure) at the level of access areas based on Local Moran's I Test of first order neighbourhoods.



It is worth noticing that the significance map is likely to produce many false positives in case of the commonly applied p-value (0,05). In order to correct for this, we apply the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995).[41] As a result of the FDR-correction, the remaining spatial autocorrelation in the spatial error model (SEM) depicted on the right map in Figure 4.5 must be considered as false positive. Therefore, our initial findings on remaining spatial autocorrelation after using SEM must be rejected. The results for SEM with and without FDR are presented in Figure 4.6.

41 C. de Castro and Singer (2006) found that in comparison to more conservative approaches (such as Bonferroni bound), FDR provides a significant gain in identifying meaningful clusters / spatial associations.

Figure 4.6: FDR “false discovery rate” – Significance Map: remaining spatial autocorrelation of SEM-residuals without (left figure) and with (right figure) FDR.



Notes: FDR “false discovery rate” – Significance Map derived on the basis of ArcGIS built-in command.

4.4 Regression results of the spatial model

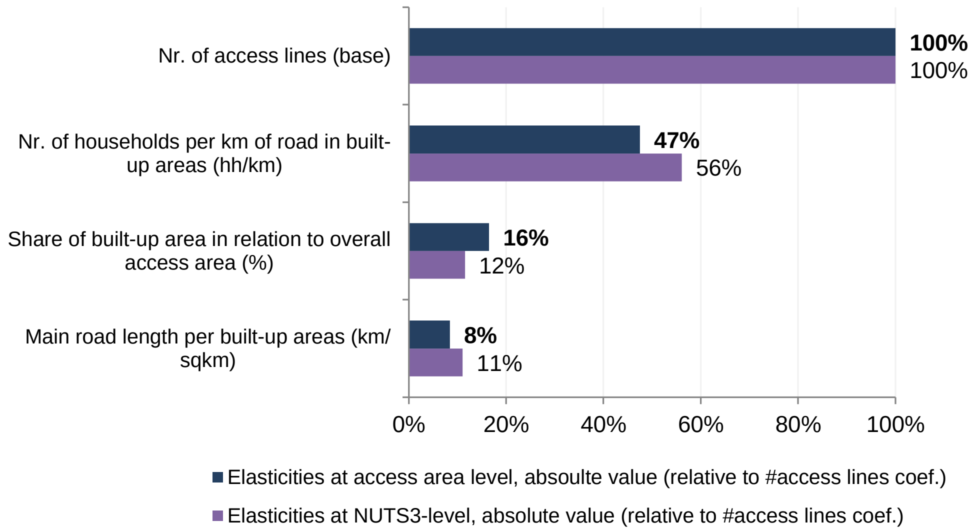
The spatial error model presented in Equation (I) is fitted using the generalized spatial two-stage least squares (GS2LS) estimation.[42] Instead of presenting the regression output tables, we present an indexation of the regression coefficients. The coefficient for the number of access lines serves as a basis. All variables are geo-referenced figures. The comparison between the results for the explanatory variables introduced in previous chapters is shown in Figure 4.7. The corresponding output tables are presented in Table 7.1 in the technical Annex.

Regression results displayed in Figure 4.7 present the **number of access lines** as the most influential variable, which is followed by the street-based **density** measure, operationalized by the *number of households per kilometre of road in built-up areas. This* is the most influential geographical factor that can be derived from publicly available data. Next according to magnitude of influence, comes **the degree of**

42 STATA’s command “spregress” and the option “gs2sls” (see STATA Manuals: spregress - Spatial autoregressive models, p. 5 or Drukker and Raciborski, 2013).

urbanization, measured by *share of built-up area in relation to overall access area*. This is closely followed by **household dispersion**, captured by the *main road length per built-up area.* It is worth mentioning that the household dispersion has an exponential effect,[43] which suggests an increasing importance of this variable with rising levels of household dispersion.[44]

Figure 4.7: Determining the required level of FTTH/P2P investment: An impact comparison among the identified geographical factors (in absolute % relative to #access lines coefficient).

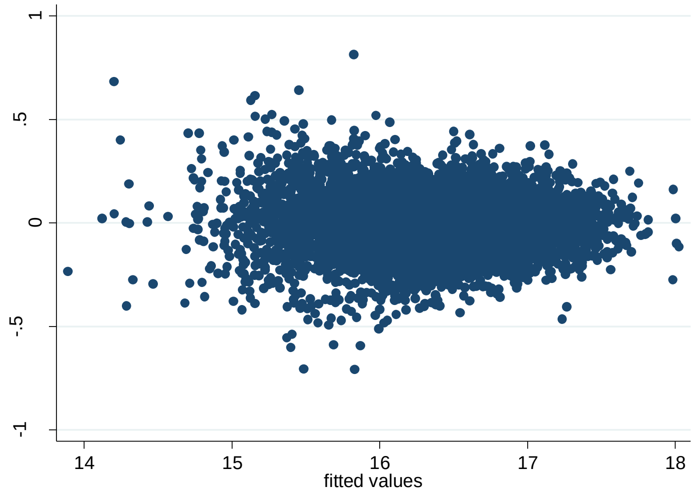


Jointly, these factors are capable of explaining 95% (R-squared 0.947) of the investment variations at the level of access areas (see Table 7.1 in the technical Annex). Additionally, when applying the regression model at the higher level of territorial aggregation (i.e. NUTS-3), the results for each identified geographical factor also remain stable (see Figure 4.7).

To assess the validity of the statistical model and to control the assumptions made on the error term, a graphical analysis of residuals ("residual vs. fit" plot) at the level of access areas is presented in Figure 4.8. Here, we observe that residuals gravitate around zero (on average), suggesting that the part of investment requirements not explained by the model does not show any systematic deviation and is consistent with a random error.

43 This refers to the quadratic term of *main road length per built-up area* in Equation (I).
44 This is in line with Phillipson (2015).

Figure 4.8: Residual analysis of the spatial error model: residuals vs. fitted values.



4.5 Regression-based predictions of fibre investment requirements aggregated at NUTS-3 level

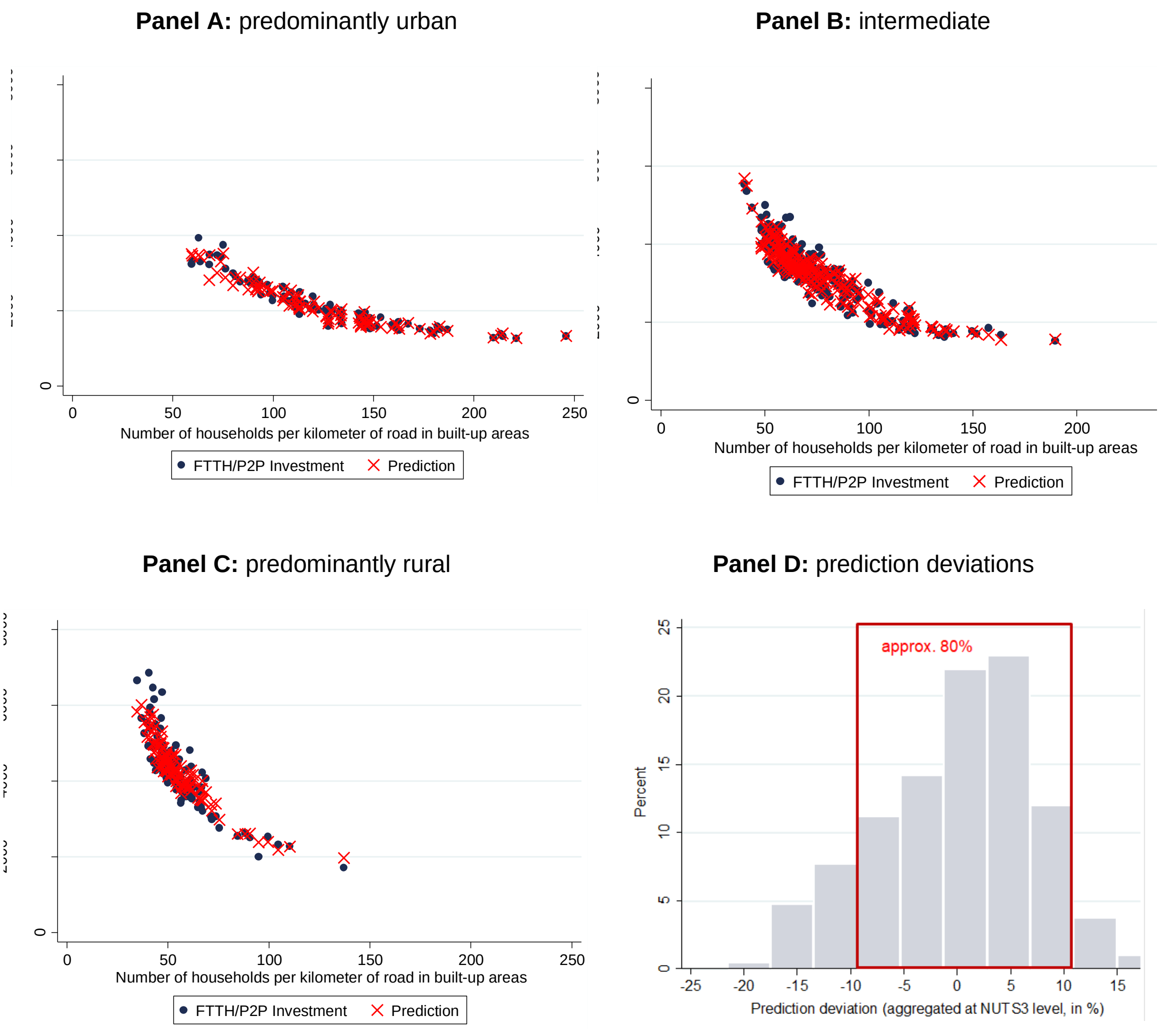
Based on the regression equation specified in Section 4.3 and on its estimation results at level of access area presented in Section 0, we assess its prediction accuracy at the next administrative demarcation level represented by NUTS-3.

Thus, investment estimates determined per access area are aggregated at NUTS-3 level[45] and plotted against their corresponding actual values (derived from bottom-up modelling). Results are presented by regiotype class (urban/ intermediate/ rural) in Figure 4.9 respectively.

In Figure 4.9, Panel A-C, we can observe that most of actual investment values (displayed in blue) lie closely to their predicted values (in red). This finding holds for the entire range of the most relevant geographical factor (*number of household per kilometre of road in built-up areas*), which in this case starts from roughly under 50 households/km in rural regions and ends above the 200 households/km in urban regions. From Figure 4.9, Panel D, we observe that for the majority of NUTS-3 regions (approx. 80%) predictions are close to their actual value (which present a deviation not higher than +/-10%) regardless of their regiotype classification (urban/ intermediate/ rural). This holds despite the fact that each regiotype class has its own and distinct steepness in their relationship (as shown in Figure 4.9, Panel A-C) with the level of investment requirement per access line.

45 In order to determine the average investment requirement per access line predicted by the model at the level of NUTS-3, investment estimates for each access area are summed at the NUTS-3 level and then divided by the corresponding number of access lines within the NUTS-3 region. For that purpose, regression estimates from the log-log model specified in Equation I are back transformed from logs into EUR without applying any non-parametric retransformation method (i.e. smearing adjustment).

Figure 4.9: Prediction vs. actual values by regiotype class in € values (Panel A, B and C) and the overall distribution of prediction deviations (Panel D) aggregated at NUTS-3 level (N= 401).



Nonetheless, there is a lack of accuracy with regard to the upper bound of the investment axis, as shown in Figure 4.9, Panel C. Here, we observe a significant discrepancy between predicted and actual values. For this reason, we also investigate in an analysis of sub-segments according the EUROSTAT categories urban, intermediate and rural. The derived coefficients for the three sub-samples show slight differences only. The detailed findings are presented in Figure 7.1 in the Appendix.

In summary, and in line with the high explanatory power of the model (R-squared), we have established a model which is capable of explaining and predicting, with a high level of precision, the underlying geographic heterogeneity in the investment requirements per access line in Germany. Most important to notice is that this is done by using a reduced number of publicly available geographical variables only.

5 Conclusions

In this paper, we described and explained noticeable differences regional differences in VHCN invest per line in Germany. In more detail, our paper provides the following insights:

- For Germany, we show that internationally standardised urban/rural typology classifications (EUROSTAT and BBSR) exhibit a significant spread in the investment costs of Very High Capacity (VHC) access networks. Nevertheless, this type of categorisation of NUTS-3 regions is too rough and imprecise that it cannot be used as an indicator for profitability.

- Our statistical regression model improves the accuracy of the investment estimation by incorporating the influence of multiple geographical factors, particularly by capturing the effect of street-based density (measured by the *number of households per kilometre of road in built-up areas*) instead of relying on an area-based measure as main explanatory variable. Further relevant explanatory factors are the degree of urbanisation (as *share of the built-up area to the overall area*) and an indicator of demand dispersion (*main road length per built-up areas*), which are of particular importance in areas where density is not evenly distributed.

- The validity of the results obtained for the aggregation level of access areas remains valid even if we apply the regression analysis at the higher aggregation level of NUTS 3 (as shown in Figure 4.7).

- More importantly, by using the aforementioned regression results at the highest level of granularity at hand (access area), we are capable of estimating and predicting, with a higher level of precision compared to alternative publicly available indicators, the required investment figures at different levels of territorial demarcations, including NUTS-3 regions.

For local authorities and politicians, these findings are of value, if they need to evaluate profitability of fibre projects and associated funding requests. These valuations require an assessment of profitability (gaps) and information on other relevant indicators besides investment (i.e. expected revenues, etc.). In this context, our results provide a more accurate point of reference for investment figures in Very High Capacity Networks (VHCN), at different levels of territorial aggregation compared to regional classifications. This holds for access areas and district areas (NUTS-3), as we have shown, and we expect the results to hold also for any other territorial area demarcation that can be derived from boundaries of network access areas within Germany.[46]

Current funding practise in Germany is managed at different administrative levels, varying from state to state ("Bundesland" to "Bundesland").[47] In order to support public decision-making, a project-specific assessment is required. Our access area based cost estimation might be suitable for evaluating individual state aid projects at different levels of territorial aggregation.

46 As the model is calibrated to reflect the particular regulation and costs in Germany, we expect that in the near future, we are able to extend the statistical model such that predictions of investment needs can also be drawn for areas outside Germany.

47 Details on call for tenders for state aid funded VHCN projects are available under https://www.breitbandausschreibungen.de/publicOverview.

6 References

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*, 27(2), 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. & Bera, A. L. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullah, Amman and Giles, David E.A., (Eds.), *Handbook of Applied Economic Statistics* (pp. 237–289). New York: Marcel Dekker.

Anselin, L., Syabri I. & Kho Y. (2006). GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, 38(1), 5-22. https://doi.org/10.1111/j.0016-7363.2005.00671.x.

[dataset] BBSR Bonn. (2018). *Siedlungsstrukturelle Regionstypen Europas* [Types of Settlement Structures in Europe]. Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR). Retrieved from https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/downloads/downloadsReferenz2.html#doc2825458bodyText1. Accessed May 6, 2021.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series* B, 57(1), 289–300.

BEREC. (2020). BEREC Guidelines on Very High Capacity Networks. Body of European Regulators for Electronic Communications (BEREC). BoR (20) 165

[dataset] Federal Ministry of Economic Affairs and Energy (BMWi). (2013). *DSL-Hauptverteilerliste*. Retrieved from http://selke.de/privates/hvt-standorte/. Accessed October 24 , 2019.

Caldas de Castro, M. and Singer, B. H. (2006). Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Test in Local Statistics of Spatial Association. *Geographical Analysis,* 38(2), 180-208. https://doi.org/10.1111/j.0016-7363.2006.00682.x

Drukker D.M., Prucha I. R. & Raciborski, R. (2013). Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. *The Stata Journal*, 13(2), 221–241. https://doi.org/10.1177/1536867X1301300201

European Commission. (2013). EU Guidelines for the application of State aid rules in relation to the rapid deployment of broadband networks. *Official Journal of the European Union*, 2013/C 25/01.

European Commission. (2020). State Aid SA.52732 (2020/N) – Germany: National gigabit scheme Germany. *The European Union*, C(2020) 7859 final

European Commission. (2021). *Broadband Strategy and Policy*. Retrieved from https://ec.europa.eu/digital-single-market/en/broadband-strategy-policy. Accessed June 9, 2021.

European Commission – Eurostat and DG for Regional and Urban Policy – ILO, FAO, OECD, UN-Habitat, World Bank. (2020). *A recommendation on the method to delineate cities, urban and rural areas for international statistical comparisons*. UN Statistical Commission, 51th session, 3 – 6 March 2020.

European Union. (2018). Directive 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications. *Official Journal of the European Union,* L 321/36.

Federal Ministry of Transport and Digital Infrastructure (BMVI) (2020). *Rahmenregelung der Bundesrepublik Deutschland zur Unterstützung des flächendeckenden Aufbaus von Gigabitnetzen in „grauen Flecken“* [Framework Regulation of the Federal Republic of Germany to support the deployment of gigabit-capable networks covering all “grey areas”] Retrieved from https://www.bmvi.de/SharedDocs/DE/Anlage/DG/Digitales/gigabit-rahmenregelung.pdf?__blob=publicationFile. Accessed October 20, 2021.

FTTH Council Europe. (2017). *The Cost of Meeting Europe’s Future Network Needs*. FTTH CE, Comsof 2017-04-12. Retrieved from https://www.ftthcouncil.eu/documents/Reports/2012/Cost_Model_Report_Full_Version.pdf. Accessed May 7, 2021.

Fourberg, N. & Korff, A. (2020). Fiber vs. Vectoring: Limiting Technology Choices in Broadband Expansion. *Telecommunications Policy*, 44(8). https://doi.org/10.1016/j.telpol.2020.102002.

Hoernig, S.; Ilic, D.; Neumann, K.-H.; Peitz, M.: Plückebaum, T.; Vogelsang, I. (2010): Architectures and competitive models in fibre networks, Bad Honnef, December 2010, http://www.wik.org/uploads/media/Vodafone_Report_Final_WIKConsult_2011-01-10.pdf.

Jay, S., Neumann, K.-H., Plückebaum, T. (2011): Implikationen eines flächendeckenden Glasfaserausbaus und sein Subventionsbedarf, WIK Diskussionsbeitrag Nr. 359, October 2011.

Jay, S.; Neumann, K-H.; Plückebaum, T. (2014): Comparing FTTH access networks based on P2P and PMP fibre topologies, *Telecommunications Policy*, 38(5), pp. 415-425. DOI: 10.1016/j.telpol.2013.04.010

Jay, S., Neumann, K.-H., Vogelsang, I., et al. (2011): Architectures and competitive models in fibre networks, in: Vodafone, *The Policy Paper Series*, 13, June 2011 .

Kulenkampff, G., Ockenfels, M., Zoz K., Zuloaga G. (2020): Costs of Broadband access networks, WIK Discussion Papers no. 473, Bad Honnef, December 2020.

Kulenkampff, G., Plückebaum, T., Zoz, K. (2019): Analytisches Kostenmodell für das Anschlussnetz AKM-AN Version 3.0, WIK-Consult, Bad Honnef, 14.01.2019, Referenzdokument im Auftrag der Bundesnetzagentur https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Marktregulierung/Massstaebe_Methoden/Kostenmodelle/Anschlussnetz/20190114_AKM_AN_RefDokpdf.pdf?__blob=publicationFile&v=2 .

Mitcsenkov, A., Kantor, M., Casier, K., Lannoo, B. Wajda K., Chen J. & Wosinska L. (2013). Geometric versus Geographic Models for the Estimation of an FTTH Deployment. *Telecommunication Systems,* 54(2),113-127.

OECD. (2011). Next Generation Access Networks and Market Structure, *OECD Digital Economy Papers*, No. 183. Paris: OECD Publishing. http://dx.doi.org/10.1787/5kg9qgnr866g-en

Phillipson, F. (2015, July). *Estimating FttH and FttCurb Deployment Costs Using Geometric Models with Enhanced Parameters*. The 20th European Conference on Network and

Optical Communications, NOC 2015, London. https://doi.org/10.1109/NOC.2015.7238613

Pisati, M. (2012, September). *Spatial Data Analysis in Stata. An Overview*. Italian Stata Users Group Meeting 2012, Bologna.

Sahebali, M.W.W., Sadowski, B.M., Nomaler, O. & Brennenraedts, R. (2021). Rolling out of fibre optic networks in intermediate versus urban areas: An exploratory spatial analysis in the Netherlands. *Telecommunications Policy*, 45(5). https://doi.org/10.1016/j.telpol.2020.102080

StataCorp. (2015). *Stata Statistical Software*. Release 14. College Station, TX: StataCorp LP.

StataCorp. (2021). Stata Spatial Autoregressive Models Reference Manual. Release 17. *A Stata Press Publication*. College Station, TX: StataCorp LLC.

Statistical Office of the European Union (EUROSTAT). (2018). Methodological Manual on Territorial Typologies. Publications Office of the European Union, 2018. Luxembourg.

[dataset] Statistical Office of the European Union (EUROSTAT). (2019). *Urban-Rural Typology Classification.* Version NUTS-2016. Retrieved from https://ec.europa.eu/eurostat/web/rural-development/methodology. Accessed February 22, 2021.

Žukowski, C., Mukhopadhyay, A., Payne, D.B. & Ruffini, M. (2021). Cost Analysis of Rural Roll-Out using long reach PON: Trading off Upfront Cost Under Uncertainty of User Take-up Rate. *Journal of Optical Communications and Networking*, 13(5), 69-84 https://doi.org/10.1364/JOCN.415806.

7 Technical Annex

Table 7.1: Regression results (p-values): Statistical significant geographical factors determining the level of FTTH/P2P investment requirements at the level of access area.

Dependent Variable: ln (FTTH/P2P Investment requirements)	Ordinary-Least-Squares (OLS) Model		Regression Model with Spatial Autocorrelated Errors		Delta OLS to Spatial Coeff.
	t-value	p-value	z-value	p-value	in %
ln (nr. of access lines)	(289.92)	0.000	(313.56)	0.000	1.4%
ln (households per km of road in built-up areas)	(-75.39)	0.000	(-91.67)	0.000	5.9%
ln (share of built-up area to total access area)	(-32.89)	0.000	(-46.18)	0.000	-13.1%
ln (main road length per sqkm of built-up area)	(24.67)	0.000	(26.13)	0.000	23.2%
ln (main road length per sqkm of built-up area)^2	(9.32)	0.000	(18.68)	0.000	18.2%
Rho (ρ)			(68.89)	0.000	
Observations (N)	7856		7856		7856
R-squared	0.947				
ovtest F-stat	1.261				
ovtest p-value	0.286				

Constant is included but not reported. Polynomials are mean-centered.

Note: Constant included but not reported. Polynomials are mean-centred.

Table 7.2: Model comparison: general vs. region-specific model.

	General spatial model	Region-specific spatial model	Errors: General to Specific Ratio
Total RMSD (root mean squared deviation)	0.12923	0.12893	1.002
urban	0.11896	0.11916	0.998
intermediate	0.13226	0.13156	1.005
rural	0.13373	0.13370	1.000
Total MAPE (mean absolute percentage error)	9.91%	9.85%	1.006
urban	9.35%	9.37%	0.997
intermediate	9.94%	9.85%	1.010
rural	10.43%	10.34%	1.009

Note: RMSD is based on logged residuals. MAPE is based on unlogged residuals.

Table 7.3: Robustness analysis: Results presented as deviation in % from the Basis-Model (OLS-Regression, see in first column).

SPECIFICATION	Basis	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9
VARIABLES		without outliers	incl. Ø house distance to the street	incl. #distr. points	incl. #sleeves/ km	incl. Ø trench cost/m (drop)	incl. Ø trench cost/m (feeder)	incl. dev of feeder length from the regional average	incl. ratio feeder-to-drop length	Y= invest/ access line
ln (nr. of access lines)		0.0%	0.1%	1.0%	-5.2%	0.5%	-0.5%	-1.3%	0.6%	
ln (households per km of road in built-up areas)		0.0%	-2.0%	15.4%	-13.1%	-0.2%	2.3%	-3.7%	0.0%	2.0%
ln (share of built-up area to total access area)		0.0%	6.5%	-9.4%	-10.8%	2.2%	14.4%	-7.9%	-3.6%	3.6%
ln (main road length per sqkm of built-up area)		0.0%	-6.9%	-70.3%	2.0%	-6.9%	3.0%	-4.0%	-2.0%	2.0%
ln (main road length per sqkm of built-up area)^2		0.0%	-15.4%	-69.2%	0.0%	-7.7%	-3.8%	0.0%	-3.8%	3.8%
Constant		0.0%	-3.7%	-0.8%	0.1%	3.8%	-5.4%	-4.3%	-0.5%	-0.9%
Observations	7856	7856	7856	7856	7841	7856	7856	7856	7856	7856
R-squared	0.947	0.947	0.949	0.956	0.955	0.947	0.947	0.948	0.947	0.938
ovtest p-value	0.286	0.286	0.055	0	0	0.014	0.331	0.185	0.015	0
ovtest f-stat	1.261	1.261	2.530	14.62	10.55	3.534	1.142	1.607	3.475	30.87
vif_1	5.809	5.809	5.949	5.882	5.918	5.891	7.741	6.221	6.368	5.239
vif_2	3.776	3.776	3.820	4.412	4.656	3.941	5.469	4.192	4.083	3.127
vif_3	3.767	3.767	3.777	4.386	4.303	3.769	3.941	4.044	3.767	3.038
vif_4	3.151	3.151	3.212	3.802	3.154	3.463	3.868	3.185	3.206	1.020
vif_5	1.084	1.084	1.119	2.111	1.295	1.511	3.166	1.639	3.060	
vif_6			1.070	1.230	1.082	1.143	1.088	1.084	1.122	

Notes: Basis refers to the OLS estimates (as diagnostic statistics are not available for the SEM). Polynomials are mean-centred. Coefficient deviations (in %) from the original OLS-Regression are reported for all identified explanatory variables. Tested covariates are included one by one (not cumulative). Diagnostic statistics(VIF and RESET-Test) are included for each test; whereas VIF is reported for each explanatory variable.

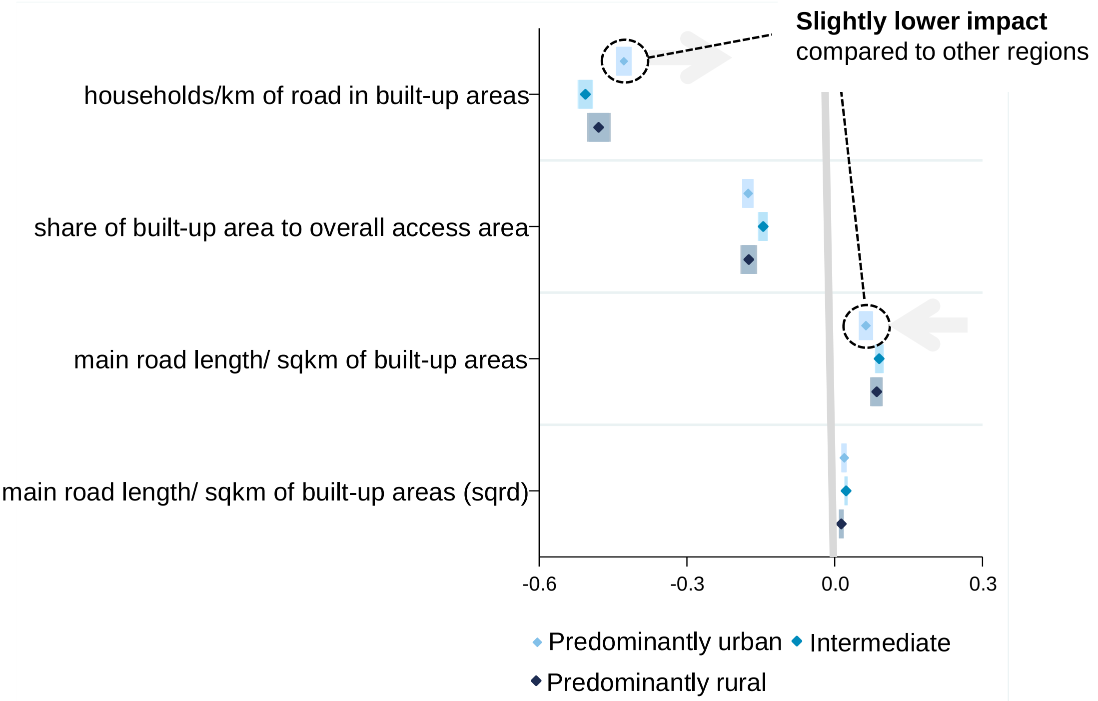
Table 7.4: Regression results: Statistical significant geographical factors determining the level of FTTH/ P2P investment requirements at NUTS-3 level.

Dependent Variable: ln (FTTH/P2P Investment requirements)	Ordinary-Least-Squares (OLS) Model		Regression Model with Spatial Autocorrelated Errors		Delta OLS to Spatial Coeff.
	t-value	p-value	z-value	p-value	in %
ln (nr. of access lines)	(179.00)	0.000	(218.12)	0.000	0.8%
ln (households per km of road in built-up areas)	(-28.13)	0.000	(-28.76)	0.000	7.6%
ln (share of built-up area to total access area)	(-3.80)	0.000	(-9.00)	0.000	-60.3%
ln (main road length per sqkm of built-up area)	(13.40)	0.000	(7.86)	0.000	46.8%
ln (main road length per sqkm of built-up area)^2	(1.50)	0.135	(3.79)	0.000	-37.0%
Rho (ρ)			(27.17)	0.000	
Observations (N)	401		401		401
R-squared	0.987				
ovtest F-stat	1.220				
ovtest p-value	0.303				

Service level data is aggregated at NUTS3-level. Constant is included but not reported. Coefficients estimated but not reported. Polynomials are mean-centered.

Note: Access area level data is aggregated at NUTS-3 level. Constant included but not reported. Polynomials are mean-centred.

Figure 7.1: Spatial regression coefficients and their 95% confidence interval by urban/rural typology groups (N urban =2114, N intermediate=3629, N rural= 2113).



Note: The constant and number of access lines are included but not reported.

WIK Wissenschaftliches Institut für Infrastruktur
und Kommunikationsdienste
Rhöndorfer Str. 68, 53604 Bad Honnef, Germany
www.wik.org

ISSN 2750-5448 (Online)