Thesis submitted in fulfillment of the requirements for the degree

Dr. rer. pol.

on the topic

Reversing the Coarse: Statistical Methods for Recovering Structure from Aggregated, Sparse, and Discretized Data

to the Institute of Statistics

Faculty of Social Sciences, Economics, and Business Administration

University of Bamberg

submitted by

Michael Mühlbauer



Bamberg 2026

Cumulative dissertation
Michael Mühlbauer, *Reversing the Coarse: Statistical Methods for Recovering Structure from Aggregated, Sparse, and Discretized Data*

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Gutachter: Prof. Dr. Timo Schmid
Gutachter: Prof. Dr. Hans Kiesl
Tag der mündlichen Prüfung: 20.07.2026

This thesis has been submitted to the Faculty of Social Sciences, Economics and Business Administration of the University of Bamberg as a dissertation.

Reviewer: Prof. Dr. Timo Schmid
Reviewer: Prof. Dr. Hans Kiesl
Date of oral examination: 20 July 2026

Dieses Werk ist als freie Onlineversion über das Forschungsinformationssystem (FIS; https://fis.uni-bamberg.de) der Universität Bamberg erreichbar.
Das Werk steht unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Namensnennung 4.0
https://creativecommons.org/licenses/by/4.0/



URN: urn:nbn:de:bvb:473-irb-117096x

DOI: https://doi.org/10.20378/irb-117096

Publication List

The publications listed below are the result of the research carried out in this thesis titled "Reversing the Coarse: Statistical Methods for Recovering Structure from Aggregated, Sparse, and Discretized Data"

1. Mühlbauer, M. (2026). Estimating disaggregated mobility indicators using area-level models on survey data
   A version of this paper has been submitted for publication

2. Mühlbauer, M., & Schmid, T. (2025). Density estimation from aggregated data with integrated auxiliary information: Estimating population densities with geospatial data. https://doi.org/10.48550/arXiv.2508.03610
   A version of this paper has been submitted for publication

3. Mühlbauer, M., & Sönning, L. (2025). Ordinal random forests in language data analysis. https://doi.org/10.31234/osf.io/s4bv2_v1
   A version of this paper has been submitted for publication

Acknowledgements

First and foremost, I thank my supervisor, Prof. Dr. Timo Schmid, who gave me the opportunity to pursue this doctorate and provided both guidance and the freedom to stray off and follow my own ideas when I felt I needed to. I thank Prof. Dr. Hans Kiesl for agreeing to act as examiner for this thesis, after only knowing me from a short discussion at Statistische Woche. And I am grateful to Prof. Dr. Anne Leucht for her calm and balanced perspectives and for what I already learnt from her during my master's.

Handing in this dissertation feels like more than just completing an important piece of work—it feels like closing a chapter in the book of life. In this chapter I was lucky enough to get caught up in a web of friends whose energy and support carried me through to this point. Some of the most important knots in my little web are Chris & Ripple: one always has an attentive British ear for my struggles, the other does not understand but looks cute and wants a snack. Clary, you have done much more for me than just fixing me up when there's the occasional medical emergency; it is calming to know you never turn your phone on silent. Another one I was lucky enough to have with me in Bamberg is Lucki, a steady source of psyche. I also do not want to miss Kaja, Sara, Steffi, David, Lotte, Julia, Andi, Franzi, Martin, Robbie and Xaver.

I am indebted to the whole team of the Institute of Statistics, who accompanied me through the daily realities of research and teaching. Dr. Florian Meinfelder was my most avid (proof-) reader and a pillar of the institute. Florian Scholze and Dr. Guy Brunotte were always there when I needed a mathematical backbone to lean on. My colleague Julius stopped being a colleague a long time ago, because he became a friend. Dr. Martin Messingschlager was the Black Swan that set me on this course in the first place. I also thank Dr. Lukas Sönning; our discussions bridging linguistics and statistics were among the most stimulating of this doctorate. Let me not forget the HiWis at the Chair, some of whom had to endure my demands and temper during stressful projects; I hope that the delightful breaks we had could make up for it. Speaking of delightful breaks—thank you, Lea and Marta.

Finally, my family: my siblings David and Lea for supporting me, and Mum and Dad for allowing me to pursue the steep academic path which led me to the current chapter. Grandpa for being an inexhaustible source of calm in what feels like an increasingly accelerating world. I am ready to flip the page to the next chapter, because of all of you.

Contents

Introduction **1**

1 Estimating Disaggregated Mobility Indicators **5**
1.1 Introduction . . . 6
1.2 Data . . . 8
1.2.1 The MiD 2017 Dataset . . . 8
1.2.2 Official Statistics from the German Statistical System . . . 9
1.2.3 INKAR Regional Indicators . . . 10
1.3 Methodology . . . 10
1.3.1 Direct Estimation: The Hajek Estimator . . . 10
1.3.2 Variance Estimation via Taylor Series Linearization . . . 11
1.3.3 The Fay-Herriot Model . . . 13
1.3.4 The Log-Transformed Fay-Herriot Model . . . 15
1.3.5 Variable Selection . . . 16
1.4 Results . . . 19
1.4.1 Direct Estimates . . . 19
1.4.2 Variable Selection . . . 20
1.4.3 Model Estimates . . . 21
1.4.4 Model Diagnostics . . . 30
1.5 Discussion . . . 33
1.6 Conclusion . . . 34
A Appendix . . . 36
A.1 Direct Estimate Maps by Transport Mode . . . 36
A.2 Bootstrap Variance Estimation . . . 36
A.3 Cumulative Penalty Weights in Sparsity-Ranked LASSO . . . 38
A.4 Final Variable Set . . . 39

2 Density Estimation from Aggregated Data **49**
2.1 Introduction . . . 49
2.2 Methodology . . . 51

2.2.1 The Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator 51
2.2.2 The Augmented Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator . 55
2.2.3 Obtaining Auxiliary Densities 57
2.3 Simulation 58
2.3.1 Simulation Setup 58
2.3.2 Simulation Results 60
2.4 Evaluation Study on Population Data 64
2.4.1 Target Variable 64
2.4.2 Auxiliary Variable 64
2.4.3 Evaluation Results 66
2.5 Application: Brown Hare Population Control 70
2.5.1 Target Variable 70
2.5.2 Auxiliary Variable 70
2.5.3 Application Results 72
2.6 Conclusion 75

3 Ordinal Random Forests in Language Data 79
3.1 Introduction 79
3.2 Data 81
3.3 Methodology 83
3.3.1 Cumulative Link Mixed Models 83
3.3.2 Mixed-Effects Ordinal Forest 85
3.4 Modeling 90
3.4.1 Cumulative Link Mixed Model 91
3.4.2 mixfabOF 93
3.5 Results 93
3.5.1 Predictive Performance 93
3.5.2 Variable Importance 96
3.5.3 Partial Dependence 99
3.6 Discussion 107
B Appendix 110

Introduction

In empirical research, the data available to the analyst do not always match the resolution required for the question at hand. A transport planner needs district-level travel behavior, but the national mobility survey was not designed to support estimates at that level. An ecologist needs a continuous map of species abundance, but the hunting records only report totals per administrative region. A linguist needs to understand a speaker's latent preference for one of two synonymous expressions on a continuum, but the questionnaire offered only five ordered categories. In each case, the phenomenon of interest is richer than what the data can directly reveal.

This observation is not peculiar to any single discipline. Every measurement involves a choice of resolution: what to record and at what level of detail. Fisher (1922), in laying the foundations of modern statistical inference, promoted the concept of *sufficiency*—the question of whether a statistic computed from a sample retains all information about the parameter to be estimated in the sample. When resolution is lost, the compression is not sufficient in Fisher's sense: information about the quantity of interest is irreversibly discarded. The question then becomes not whether information has been lost, but how much, and whether what remains—combined with external knowledge—can still support the inference the analyst requires.

The concept of *coarsened data* provides a formal framework for this class of problems. Heitjan and Rubin (1991) define coarsening as a general mechanism in which the observed datum is a set-valued function of the true value: instead of observing the precise quantity of interest, the analyst observes only that it belongs to some region of the sample space. This formulation generalizes the classification of missing data mechanisms introduced by Rubin (1976)—where an observation is either fully present or entirely absent—to the richer setting in which partial, set-valued information is available (cf. Gill et al., 1997; Heitjan, 1994). Censoring, rounding, grouping, and spatial aggregation are all instances of this single phenomenon. Resolution can also be lost at the level of the study design rather than the individual observation: when a survey is designed for national-level inference, its sample may be too sparse in fine-grained geographic domains to support reliable direct estimates, even though each response is recorded in full detail. Here the coarsening is not a property of the datum but of the design—the geographic resolution at which the survey can speak with acceptable precision is coarser than the

resolution the analyst requires. In both cases—whether the coarsening acts on the datum or on the design—the available data do not carry enough information to pin down the quantity of interest without additional assumptions. In the fields most relevant here, small area estimation addresses the sparsity of survey samples in fine-grained domains (Rao & Molina, 2015); spatial disaggregation methods address the loss of spatial precision through aggregation (Groß et al., 2017); and latent variable models address the discretization of continuous processes into ordinal categories (Buczak, 2025c; Liddell & Kruschke, 2018; McCullagh, 1980).

When the coarsening operates on the observation itself, recovering the fine-grained quantity is, in a precise sense, an inverse problem. The forward map—from the true value to the coarse datum—is well defined, but the reverse is not always: because the coarsening is not injective—many fine-grained values map to the same coarse datum—multiple truths are consistent with the same observation. When, instead, the resolution loss stems from the coarseness of the sampling design, the challenge is not inversion but stabilization: the target parameter is well defined and the direct estimator is unbiased, yet too imprecise to be useful on its own.

A common strategy for addressing both challenges is *borrowing strength*—using models to link related sources of information, so that the estimate for any one target benefits from what is known about others (Ghosh & Rao, 1994). In small area estimation, this takes the form of mixed models that link domain-specific estimates through shared parameters (Fay & Herriot, 1979; Rao & Molina, 2015); in spatial disaggregation, auxiliary information such as remote sensing imagery constrains the otherwise underdetermined density recovery (Stevens et al., 2015); and in ordinal data analysis, latent variable structure and mixed effects recover continuous variation from categorical responses.

The three chapters correspond to three distinct mechanisms through which resolution is lost:

(i) **Sampling sparsity.** When a survey is designed for national- or regional-level inference, its geographic resolution is too coarse for the finer domains the analyst may require. District-level direct estimates may then be based on very few observations, resulting in unacceptably large sampling variances. The true domain parameter exists but cannot be estimated with sufficient precision from the available sample alone.

(ii) **Spatial aggregation.** When individual-level geocoordinates are collapsed into regional totals—for reasons of data collection, administrative convention, or confidentiality—the analyst observes only that each unit lies within a particular area, losing all information about its precise location. The continuous spatial density exists but is hidden behind the coarse partition.

(iii) **Ordinal discretization.** When a respondent's position on a continuous latent scale is recorded through an ordered categorical response, the measurement instrument imposes a coarse binning that preserves only the rank order among response categories (Stevens,

1946). The underlying preference or judgment exists on a continuum, but only a discretized version is observed.

Despite the diversity of application domains—transport, demography, ecology, and linguistics—the inferential challenge is structurally related in all three cases. The observed data represent an incomplete reduction of a richer underlying structure, and the goal is to draw inferences—whether estimating a fine-grained quantity or understanding the dynamics between covariates and the response—that the coarsened data alone cannot support. In the absence of such additional information, the coarsening is genuinely non-invertible: a sparse sample admits a wide range of domain means; an area-level count is consistent with infinitely many spatial distributions within the area; an ordinal category is compatible with any latent value between two thresholds. What strengthens inference in each chapter is the availability of covariates, auxiliary densities, or structural assumptions that narrow the space of plausible solutions.

The first two chapters both address spatial inference under limited information. Chapter 1 begins with a survey not designed for district-level inference—the German household travel survey *Mobilität in Deutschland 2017*—and asks how district-level mobility indicators can be recovered when direct estimates for many of the target domains are unreliable. The target quantities are person-kilometers traveled per day, disaggregated by 400 German districts and four transport modes, yielding 1,590 estimation domains of which only about a third achieve acceptable precision under direct estimation. The chapter develops a log-transformed Fay-Herriot area-level model (Fay & Herriot, 1979) that borrows strength through covariates sourced from administrative registers and the INKAR regional indicator database. From an initial pool of over 1 000 candidate variables, a three-phase pipeline combining LASSO, sparsity-ranked LASSO, and cross-validation on the Fay-Herriot model itself selects the final covariate set. The resulting estimates reduce coefficients of variation by 16–44% and reveal pronounced spatial patterns in multimodal travel behavior—an east–west gradient in car dependence, the concentration of public transit usage in metropolitan networks—that direct estimation alone cannot resolve. Chapter 2 operates at a deeper level of spatial coarsening: here, individual geocoordinates have been entirely collapsed into regional totals, and the continuous spatial density must be reconstructed from aggregate counts alone. The starting point is the GRSST estimator (Groß et al., 2017), which treats spatial aggregation as a measurement error problem and uses an iterative algorithm akin to the stochastic expectation–maximization procedure to estimate the underlying density. Within large administrative areas, however, this estimator produces homogeneous density plateaus due to the absence of sub-regional information. The chapter proposes the AGRSST estimator, which integrates auxiliary spatial information—nighttime light satellite imagery for population densities, vegetation indices for wildlife abundance—through a correlation-weighted convex combination that improves precision when the auxiliary data are informative and provides robustness when they are not.

Where the first two chapters address coarsening along the spatial dimension, Chapter 3 turns to a fundamentally different axis: the coarsening imposed by the measurement instrument itself. When a speaker of Maltese English rates a lexical pair on a five-point scale, the response discretizes what is, in principle, a continuous latent preference. Rating scales are prevalent in language research, yet the standard practice of converting ordinal categories to numeric scores and applying linear regression can yield misleading results when the underlying relationships are non-linear or when the category distribution is severely imbalanced (Liddell & Kruschke, 2018). The chapter demonstrates a modelling approach based on the frequency-adjusted borders ordinal forest (Buczak, 2025b), which, like the cumulative link mixed model (McCullagh, 1980), operates on a continuous latent score underlying the ordinal response. Employing the recently developed mixed-effects extension (mixfabOF; Buczak, 2025c) to handle hierarchical data structures, the chapter contributes previously unavailable partial dependence routines for model interpretation. Applied to lexical preference ratings in Maltese English, the mixed-effects ordinal forest outperforms the parametric alternative in predictive accuracy and reveals non-linear, non-monotonic age trajectories in language change—fine-grained patterns of variation that models operating on numeric scores systematically miss.

Across all three chapters, the methodological contribution lies not in proposing entirely new estimators, but in augmenting established frameworks—the Fay-Herriot model, the GRSST algorithm, the ordinal forest—with tools that extend their practical reach: principled mechanisms for integrating auxiliary information in the first two chapters, and previously unavailable partial dependence routines for model interpretation in the third. The title of this thesis reflects this shared agenda: *Reversing the Coarse* refers both to the recovery of resolution lost by measurement or by design, and to the change of course the analyst must make—from expecting finer-grained data to recovering structure and insight from what is available.

Chapter 1

Estimating Disaggregated Mobility Indicators using Area-Level Models on Survey Data

Abstract

Large-scale mobility surveys, such as the Mobility in Germany 2017 survey (*Mobilität in Deutschland 2017*, MiD), are often designed to produce reliable statistical indicators at the level of specific subpopulations defined by geographical regions (e.g., states, districts, counties). After the data have been collected, there is often a need for indicators at a lower geographical level, for which direct estimation may not provide a satisfactory degree of precision due to insufficient sample sizes. This paper demonstrates the application of a log-transformed Fay-Herriot area-level model to estimate district-level person-kilometers traveled per day across four transport modes—walking, cycling, car travel, and public transit—using MiD data. The 1 590 estimation domains span 400 German districts crossed with four modes. Covariates are obtained from the German statistical system and the Indicators and Maps for Spatial and Urban Development (*Indikatoren und Karten zur Raum- und Stadtentwicklung*, INKAR) regional indicator database. They are selected using a three-phase procedure incorporating standard Least Absolute Shrinkage and Selection Operator (LASSO), sparsity-ranked LASSO for interactions and polynomials, and cross-validation on the Fay-Herriot model. The variances of the direct estimates are estimated using Taylor series linearization. Two distributional assumptions, the Gaussianity of the random effects and the residuals, are made in the modeling process. As the untransformed data do not fulfill these assumptions, a logarithmic transformation is applied, which achieves good fit to normality. The results show significant spatial variation in person-kilometers across German districts and transport modes. The use of the model reduces coefficients of variation by 16–44% compared to direct estimation.

Keywords: Mobility Indicators, Small Area Estimation, Fay-Herriot Model, Variable Selection, Multimodal Transport

1.1 Introduction

A detailed understanding of mobility patterns is crucial for transport planning and climate-related policy. The transport sector—including both freight and passenger movement—accounts for approximately one quarter of greenhouse gas emissions in the European Union, making mobility policy a central component of climate strategy (European Commission, 2020). Effective policy interventions require spatially disaggregated mobility indicators that capture regional heterogeneity in travel behavior—information that national-level statistics cannot provide.

Mobility can be measured using indicators such as the average number of trips per person per day, modal split, or average trip length. The granularity of these indicators can be increased by disaggregating across socio-economic groups, geographic areas, or transport modes. The specific indicator we focus on is person-kilometers traveled per day, which we call mean person-kilometers (MPK). This indicator appears regularly in major surveys such as the National Household Travel Survey (NHTS) (McGuckin & Fucci, 2018) and provides insight into regional travel demand when disaggregated spatially. We extend this logic to joint geographic and modal disaggregation: estimating MPK separately for walking, cycling, car travel, and public transit at the German district level. This multimodal approach produces four distinct policy-relevant indicators from a single methodological framework, enabling targeted interventions such as cycling infrastructure investments or public transit coverage improvements.

Our data source is the large-scale German mobility survey Mobility in Germany 2017 (*Mobilität in Deutschland 2017*) (MiD), which provides detailed trip records for over 316 000 respondents (Nobis & Kuhnimhof, 2018). The MiD is designed to provide precise direct estimates at the national level and for 42 in-depth sampling areas that sponsored extended data collection (Eggs et al., 2018). For the remaining districts, sample sizes are insufficient for reliable direct estimation. This limitation motivates the use of model-based methods that borrow strength from auxiliary data.

Small area estimation (SAE) methods are designed to provide precise predictions for domains with small sample sizes by incorporating auxiliary information. Rao and Molina (2015) provide a comprehensive treatment of SAE theory, while Morales et al. (2021) offer an application-oriented textbook focusing on implementation in the `R` programming language (R Core Team, 2026). Additional methodological developments are covered in Jiang and Lahiri (2006) and Pratesi (2016). Tzavidis et al. (2018) present a framework for the complete SAE process from user needs assessment through method evaluation. Applications of SAE to mobility research have emerged over the past two decades. Long et al. (2009) applied a spatially correlated variant of the Battese-Harter-Fuller unit-level model (Battese et al., 1988) to estimate household travel characteristics at the census tract level using NHTS data for Des Moines, Iowa. Vaish et al. (2010) applied a survey-weighted hierarchical Bayes method to the 2001 NHTS to esti-

mate the prevalence of high daily person-miles of travel at the state level, achieving 31–48% reductions in prediction interval widths compared to design-based survey estimates. Salon (2016) estimated pedestrian and cyclist activity at the neighborhood scale for California census tracts using a Broad Area Ratio Estimator. More recently, Anjoy (2023) estimated commuting patterns in India, and Al-Khasawneh and Cirillo (2024) addressed the challenge of estimating household trips for census tracts in Maryland, using synthetic population generation and regression models for tracts lacking survey coverage, and applying the Fay–Herriot (FH) model to improve precision in tracts with limited samples. Within the MiD context specifically, Bäumer et al. (2018) applied unit-level SAE models to estimate mobility indicators at the district level, combining survey-reported individual characteristics with district-level auxiliary data from official statistics. This approach was extended to the MiD 2023 data by Bäumer and Pfeiffer (2025), who incorporated additional address-level covariates (slope, public transit accessibility, location within municipality) and benefited from improved auxiliary data availability through the 2022 German Census. However, both reports employ a limited set of covariates selected based on domain knowledge rather than systematic variable selection, and do not report model diagnostics or precision estimates.

The present paper makes three contributions to SAE in mobility research. First, we develop a multimodal estimation framework that jointly models district-level MPK across four transport modes, defining 1 590 estimation domains (400 districts $\times$ 4 modes, minus 10 out-of-sample domains). Second, we propose a three-phase variable selection pipeline that combines standard least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) for main effect selection with ranked sparsity LASSO (Peterson & Cavanaugh, 2022) for interactions and polynomials, followed by cross-validation on the FH model itself (Ren et al., 2022) to select the final specification. Third, we integrate auxiliary data from the German statistical system (Destatis) and the Indicators and Maps for Spatial and Urban Development (*Indikatoren und Karten zur Raum- und Stadtentwicklung*) (INKAR) regional indicator database, providing a substantially richer covariate set than previous MiD-based analyses. The log-transformed FH model (Fay & Herriot, 1979) achieves precision improvements of 16–44% across transport modes compared to direct estimation.

The remainder of this paper is structured as follows. Section 1.2 introduces the MiD survey data and auxiliary data sources. Section 1.3 describes the methodology including direct estimation, variance estimation, the log-transformed FH model, and the variable selection procedure. Section 1.4 presents estimates and model diagnostics. Section 1.5 addresses limitations and directions for future research. Section 1.6 summarizes the findings.

1.2 Data

1.2.1 The MiD 2017 Dataset

The MiD is a nationwide German survey on various aspects of the mobility of people living in Germany. The survey is conducted by the commercial research institute Institute for Applied Social Sciences (*infas Institut für angewandte Sozialwissenschaft GmbH*) (infas) on behalf of the Federal Ministry for Digital and Transport (*Bundesministerium für Digitales und Verkehr*) (BMDV) (formerly the Federal Ministry of Transport and Digital Infrastructure). Further project partners include the IVT Research GmbH (IVT) and the German Aerospace Center (*Deutsches Zentrum für Luft- und Raumfahrt e. V.*) (DLR). An overview of the survey's results is found in Nobis and Kuhnimhof (2018); a detailed report of the methodology is documented in Eggs et al. (2018). Additional information and documentation can be obtained on the project's official website (Infas, 2023). Previous iterations of the survey were conducted in 2002 and 2008, with the most recent iteration, MiD 2023, published in November 2025 (Infas et al., 2025). This work uses the 2017 iteration, as our analysis was completed before the 2023 data became available; however, the methodological framework developed here—particularly the variable selection pipeline and multimodal estimation approach—is directly applicable to the newer data. The 2017 wave provides detailed data (Infas et al., 2019) on a range of socio-economic indicators for 316 361 individuals from 156 420 households. Respondents documented their movements by reporting a total of 960 619 trips. The MiD sampling design is complex. An introduction to the full sampling design of MiD is beyond the scope of this article, and the interested reader is referred to Eggs et al. (2018). Some features of the design include a triple-frame approach in which sampling units are drawn from the official German residence registration register as well as from a system that generates mobile phone numbers by random digit dialing and one that does so for landline phone numbers. Components of multistage sampling are also implemented, but the exact design is neither clear nor reproducible. Dataset packages from MiD exist for three levels of spatial resolution, the one used in this work being the middle one: B.2 Regional. This analysis draws on both person-level and trip-level data from this package. The person dataset provides person-level sampling weights `P_HOCH`, which account for unequal inclusion probabilities due to the sampling design and include calibration to known population totals on multiple geographical levels. These weights are used throughout the analysis for direct estimation and variance calculation. The trip dataset provides the reported journeys with distance and mode information. The official MiD documentation provides a general overview of how post-stratification was carried out, but the level of detail is insufficient for reproduction. It is established that the weights can be used in the Horvitz–Thompson (HT)-type estimators introduced in Section 1.3.1 (Horvitz & Thompson, 1952) at the German national level (Nobis & Köhler, 2018).

The MiD provides the data to estimate our parameter of interest: MPK traveled per day, disag-

gregated by travel mode. The estimation domains are defined by the cross-classification of district of residence (*Kreis*) and main travel mode (*Hauptverkehrsmittel*), yielding 1 604 domain cells across 401 districts. Travel modes are grouped into four categories: walking (*zu Fuß*), cycling (*Fahrrad*), motorized individual transport (*MIV*, combining car drivers and passengers), and public transport (*ÖPV*). Trip-level data are merged with person-level data to obtain both person characteristics and trip-related covariates at the domain level. Observations with missing values or error codes in key variables are excluded, and sampling weights are proportionally adjusted to preserve population totals. 10 domains are excluded where no sampled respondent reported any trip by that mode; they constitute out-of-sample domains for which the Slud-Maiti back-transformation does not provide estimates (Kreutzmann et al., 2019; Slud & Maiti, 2006). Combined with one district exclusion due to auxiliary data misalignment (see Section 1.2.3), the final analysis comprises 1 590 domains across 400 districts. Trips are attributed to the respondent's district of residence, aligning the outcome variable with residence-based auxiliary covariates and survey calibration.

While the MiD provides the outcome variable and domain structure, we exclude MiD-derived covariates from the auxiliary variable pool. Survey-derived covariates introduce measurement error that violates the FH model's assumption of error-free area-level covariates (Ybarra & Lohr, 2008). We therefore rely exclusively on administrative data sources for auxiliary information.

1.2.2 Official Statistics from the German Statistical System

Small area estimation methods require auxiliary data that are available for all domains of interest. We draw on official statistics from the German statistical system, which provides district-level structural indicators as potential covariates for the FH model. These are register-based statistics (*laufende Statistiken*) compiled annually from tax records, vehicle registrations, population registers, and other administrative processes. The primary data source is the Regional Database Germany (*Regionaldatenbank Deutschland*) (RDB), a joint database of federal and state statistical offices (Statistische Ämter des Bundes und der Länder, 2025), supplemented by the GENESIS-Online Database (*Gemeinsames neues statistisches Informationssystem*) (GENESIS) of the Federal Statistical Office of Germany (Destatis) (2025) for vehicle registration data. For brevity, we refer to these data collectively as Destatis throughout this paper. Data retrieval was automated using the `restatis` `R` package (Buhl et al., 2025).

We acquire 38 statistical tables from the 2017 reference year, organized into eight thematic domains, to serve as candidate auxiliary variables for the model (a complete listing is available at https://github.com/muehlbauer-m/sae-mobility-germany). The selection of thematic domains reflects categories of auxiliary variables commonly employed in travel behavior research, such as socioeconomic and household characteristics (e.g., Stead, 2001) and built environment mea-

sures (e.g., Ewing & Cervero, 2010). Our auxiliary variables include population and household statistics, income and economic indicators, mobility-relevant variables, demographic indicators, built environment characteristics, tourism statistics, and social infrastructure. All variables refer to the 2017 reference year to match the MiD survey period.

1.2.3 INKAR Regional Indicators

To capture broader regional characteristics beyond Destatis, we incorporate data from INKAR, a comprehensive indicator database maintained by the Federal Institute for Research on Building, Urban Affairs and Spatial Development (*Bundesinstitut für Bau-, Stadt- und Raumforschung*) (BBSR) (2025). INKAR aggregates data from multiple federal agencies—including the Federal Statistical Office, the Federal Employment Agency, the Federal Motor Transport Authority, and the Federal Office for Cartography and Geodesy—into a unified system of regional indicators harmonized to consistent territorial boundaries. Data access was implemented programmatically using the `bonn R` package (Marbach, 2026), which provides an interface to the INKAR API.

We downloaded all 562 indicators available at the district level (*Kreise*) for the 2017 reference year to match the MiD survey period. The indicators are organized into 41 thematic categories covering unemployment, employment structure, demographics, housing, education, transport accessibility, settlement structure, social transfers, income, public finances, healthcare, land use, and economic output. After removing duplicates and indicators with data quality issues, 540 unique indicators remained for analysis (the complete list with English translations is available at https://github.com/muehlbauer-m/sae-mobility-germany). Of these, 424 (79%) had 2017 data directly available; for the remaining 116, we used the nearest available year. Data quality was high: 480 indicators (89%) had complete coverage across all 400 districts, and 526 (97%) had less than 10% missing values. Note that INKAR uses harmonized 2019 territorial boundaries, which contain 400 districts rather than the 401 districts in the MiD 2017 data; this difference arises from the 2021 merger of the independent city of Eisenach into the Wartburgkreis district (Thüringer Landtag, 2019). One MiD domain (Eisenach) is therefore excluded from the final analysis due to missing auxiliary data.

1.3 Methodology

1.3.1 Direct Estimation: The Hajek Estimator

The design-based paradigm assumes a total finite population $\mathcal{U}$, which can be split into D subpopulations (domains) $\mathcal{U}_d \subset \mathcal{U}$ such that $\mathcal{U} = \cup_{d=1}^{D} \mathcal{U}_d$. A fundamental estimator for the domain

total $\tau_d = \sum_{i \in \mathcal{U}_d} y_i$ is provided by Horvitz and Thompson (1952)

$$\hat{\tau}_d^{HT} = \sum_{i \in \mathcal{S}_d} y_i \cdot w_i,$$

where $i \in \{1, 2, \ldots, n_d\}$ denotes the index of the i-th elementary sampling unit in domain d, n_d is the number of units sampled in domain d, $w_i = \frac{1}{\pi_i}$ is the sampling weight or inverse of the first-order inclusion probability π_i of sampling unit i, and y_i is the target variable's value for sampling unit i. The sub-population samples $\mathcal{S}_d$ sum up to the full sample $\mathcal{S}$ so that $\mathcal{S} = \cup_{d=1}^{D} \mathcal{S}_d$.

When the population count N_d in domain d is unknown, the domain mean μ_d can be estimated by the weighted domain sample mean (Särndal et al., 1992, Result 5.8.1, p. 185), commonly referred to in the SAE literature as the Hajek (HJ) estimator (e.g., Morales et al., 2021) in reference to Hájek (1971):

$$\hat{\mu}_d^{HJ} = \frac{\sum_{i \in \mathcal{S}_d} y_i \cdot w_i}{\sum_{i \in \mathcal{S}_d} w_i}. \tag{1.1}$$

This estimator allows the estimation of the mean $\mu_d = N_d^{-1} \sum_{i \in \mathcal{U}_d} y_i$ without requiring knowledge of the total population count N_d in domain d; moreover, it is often preferable to HT-based alternatives of the form $N_d^{-1} \hat{\tau}_d^{HT}$ even when N_d is known (Särndal et al., 1992, p. 182). Under mild conditions (n_d is large and $\pi_i > 0 \; \forall \; i \in \mathcal{U}_d$), $\hat{\mu}_d^{HJ}$ is an approximately unbiased estimator of μ_d, so that $\mathbb{E}\left[\hat{\mu}_d^{HJ}\right] \approx \mu_d$ (for proof, see e.g., Morales et al., 2021, pp. 23).

1.3.2 Variance Estimation via Taylor Series Linearization

Variance estimates of the HJ estimator are needed to evaluate precision and as inputs for the FH model. Standard variance formulas for ratio estimators (see Särndal et al., 1992, p. 185) require second-order inclusion probabilities π_{ij}, which are unavailable for the MiD. Simplified formulas assuming independent sampling do not apply to complex survey designs. We therefore employ first-order Taylor series expansion (linearization) as implemented in the `survey` package for `R` (Lumley, 2004, 2010), following the framework of Binder (1983). This approach approximates variance estimation for any smooth statistic through variance estimation of weighted score totals, accommodating complex design features.

The variance estimator for domain d takes the form

$$\hat{\mathbb{V}}\left(\hat{\mu}_d^{HJ}\right) = \hat{J}_d^{-2} \sum_{s=1}^{S} \frac{m_s}{m_s - 1} \sum_{c=1}^{m_s} \left(\tilde{U}_{sc}^{(d)} - \bar{\tilde{U}}_{s\cdot}^{(d)}\right)^2, \tag{1.2}$$

where $s = 1, \ldots, S$ indexes strata, $c = 1, \ldots, m_s$ indexes primary sampling units (PSUs) within stratum s, $\hat{J}_d = \sum_{i \in \mathcal{S}_d} w_i$ is the estimated domain population size, and $\bar{\tilde{U}}_{s\cdot}^{(d)} = m_s^{-1} \sum_{c=1}^{m_s} \tilde{U}_{sc}^{(d)}$

is the estimated stratum mean. The calibration-adjusted score totals at the PSU level are

$$\tilde{U}_{sc}^{(d)} = \sum_{i \in c} \delta_i^{(d)} \cdot w_i \left[\left(y_i - \hat{\mu}_d^{HJ} \right) - \hat{\boldsymbol{\gamma}}' \mathbf{Z}_i \right], \tag{1.3}$$

where $\delta_i^{(d)}$ indicates domain membership, $\mathbf{Z}_i$ contains calibration variables, and $\hat{\boldsymbol{\gamma}}$ is the coefficient from regressing scores on calibration variables.

The estimator in Eq. (1.2) accounts for four design features: (1) unequal probability sampling through the weights $w_i = 1/\pi_i$; (2) clustering at the household (PSU) level, which increases variance relative to independent sampling due to intra-household correlation; (3) stratification by federal state and regional type ($S = 160$ strata), which reduces variance through within-stratum homogeneity; and (4) calibration to known population totals via Deville-Särndal raking (Deville & Särndal, 1992), where the term $\hat{\boldsymbol{\gamma}}' \mathbf{Z}_i$ removes score variability explained by calibration variables.

One design feature cannot be fully accommodated: the MiD's triple-frame design (population registry, landline, and mobile phone sampling) involves frames with different clustering structures, but frame membership is not available in the scientific use file. We therefore apply a single-frame household-clustered variance estimator, which cannot exploit the efficiency gains of a proper multi-frame estimator and thus likely overestimates variance.

Table 1.1 summarizes the design specification. Calibration variables were validated against Destatis population totals, with mean deviations of approximately 1% for federal states, 0.7% for sex, 2.1% for age groups, 3.7% for survey quarter, and 1.4% for weekday.

Element	**Specification**	**Rationale**
PSUs (ids)	H_ID_Reg (Household)	Household members surveyed together
Strata	Federal state × MiD regional type	16 states × 10 regional types
Weights	`P_HOCH` / 4	Pre-calibrated; divided for person-mode structure
Calibration	Federal state, sex, age, quarter, weekday	Informs variance calculation

Table 1.1: Survey design specification for variance estimation.

An alternative variance estimation approach via calibrated bootstrap resampling is presented in Section A.2. Three bootstrap variants are compared: naive (no recalibration), MiD-calibrated (recalibrated to survey-weighted population totals), and census-calibrated (recalibrated to external census totals). The calibrated bootstrap methods yield virtually identical estimates regardless of calibration source ($R^2 > 0.99$, slopes ≈ 1.0), and the naive bootstrap shows no meaningful difference from calibrated approaches. Taylor series linearization produces slightly more conservative estimates than bootstrap methods: bootstrap standard errors are smaller than

their analytical counterparts in 82% of domains, with a median ratio of 0.92 ($R^2 = 0.96$). The close agreement between two methodologically distinct approaches lends credibility to the analytical variance estimates used throughout this paper. We prefer linearization because it jointly accounts for all four design features described above—clustering, stratification, unequal weights, and calibration—whereas the calibrated bootstrap only incorporates calibration through weight recalibration. Additionally, linearization is considerably less costly to compute.

For comparison with the mean squared error (MSE) of model-based predictions, variance estimates are standardized using the coefficient of variation (CV)

$$CV_d^{HJ} = \frac{\sqrt{\hat{\mathbb{V}}\left(\hat{\mu}_d^{HJ}\right)}}{\hat{\mu}_d^{HJ}}, \quad \forall \hat{\mu}_d^{HJ} > 0. \tag{1.4}$$

1.3.3 The Fay-Herriot Model

The FH model is introduced by Fay and Herriot (1979). Detailed accounts can be found in standard SAE literature such as Rao and Molina (2015). The notation used in this paper is based on Morales et al. (2021).

An important feature of the FH model is that it combines the two statistical paradigms of the design-based and the model-based approach. Model-based statistics define a data-generating stochastic process with distributional assumptions, whereas design-based statistics postulate a fixed population, with randomness emerging from the sampling design (Dumelle et al., 2022). The FH model bridges these paradigms by using design-based direct estimates (such as the HJ estimator from Section 1.3.1) as inputs to a model-based framework, thereby leveraging the strengths of both approaches.

The FH model is formulated at the domain level for each domain $d \in \{1, 2, \ldots, D\}$—in purely spatial contexts also referred to as areas—requiring covariates that are available for all domains. In the present multimodal analysis, domains are defined as district $\times$ travel mode combinations. Covariates from Destatis and INKAR are measured at the district level and are constant across the four modes within each district—for example, the unemployment rate in a district applies equally to all four mode-specific domains. Mode-specific variation is captured through sum-to-zero contrast coding of the travel mode factor, allowing both mode-specific intercepts and mode $\times$ covariate interactions in the model.[1] The model is derived by merging two models in a staged process. The first stage comprises the sampling model

$$y_d = \mu_d + e_d, \tag{1.5}$$

[1] An alternative would be a twofold sub-area level model (Torabi & Rao, 2014) with separate random effects for modes and districts. However, with only four mode categories, the mode-level variance component is estimated from very few draws, a setting prone to boundary estimates at zero (cf. Gomes, 2022). Preliminary experiments with the twofold specification confirmed that one variance component collapsed to zero in our data.

where y_d is a direct estimator (here: the HJ estimator in Eq. (1.1), so that $y_d = \hat{\mu}_d^{HJ}$) of the true domain statistic and parameter of interest μ_d. The design-based variances of y_d are assumed to be known and denoted by σ_d^2. In applications, the values of σ_d^2 are estimated and treated as known in the model. Within the FH model framework, the e_d are called sampling errors and express the disturbance of direct estimates y_d around μ_d. The linking model

$$\mu_d = \mathbf{x}_d\boldsymbol{\beta} + u_d \tag{1.6}$$

forms the second stage of the FH model. It linearly links the true domain statistics μ_d to a set of p auxiliary variables contained within a row vector $\mathbf{x}_d \in \mathbb{R}^p$. The strength of the linear relationship is modeled by a set of regression parameters contained in the column vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The i.i.d. u_d are called random effects. They capture variability of the μ_d across domains due to unobserved characteristics that are not explicitly included in the model. Plugging Eq. (1.6) into Eq. (1.5) results in the single formulation known as the FH model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_{D\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{u}_{D\times 1} + \mathbf{e}_{D\times 1} && (1.7)\\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, && (1.8) \end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_D)^\top$, $\mathbf{u} = (u_1, \ldots, u_D)^\top$, $\mathbf{e} = (e_1, \ldots, e_D)^\top$, and $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_D^\top)^\top$. The random components $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_u)$ and $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_e)$ are assumed to be independent, with $\mathbf{V}_u = \text{var}(\mathbf{u}) = \sigma_u^2\mathbf{I}_D$, $\mathbf{V}_e = \text{var}(\mathbf{e}) = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$, and $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{V}_u + \mathbf{V}_e = \text{diag}(\sigma_u^2+\sigma_1^2, \ldots, \sigma_u^2+\sigma_D^2)$. Note that even though $\mathbf{y}$ is modeled, we are interested in predicting the vector of true domain parameters $\boldsymbol{\mu}$. The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and best linear unbiased predictor (BLUP) of $\mathbf{u}$ are obtained with

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{y} \quad \text{and} \quad \tilde{\mathbf{u}} = \mathbf{V}_\mathbf{u}\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right). \tag{1.9}$$

Here, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ assume that the quantity σ_u^2 is known. By plugging in an empirical estimate for σ_u^2, the empirical BLUE (EBLUE) and empirical BLUP (EBLUP) of $\boldsymbol{\beta}$ and $\mathbf{u}$ are obtained with

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\widehat{\mathbf{V}}^{-1}\mathbf{y} \quad \text{and} \quad \hat{\mathbf{u}} = \widehat{\mathbf{V}}_\mathbf{u}\widehat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right). \tag{1.10}$$

The true domain statistic's EBLUP is obtained by plugging $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ into Eq. (1.6)

$$\hat{\mu}_d = \mathbf{x}_d\widehat{\boldsymbol{\beta}} + \hat{u}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_d^2}y_d + \frac{\sigma_d^2}{\hat{\sigma}_u^2 + \sigma_d^2}\mathbf{x}_d\widehat{\boldsymbol{\beta}} = \hat{\gamma}_d y_d + (1-\hat{\gamma}_d)\mathbf{x}_d\widehat{\boldsymbol{\beta}}, \tag{1.11}$$

where the quantities $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\sigma_d^2}$ control the weight put on the direct (y_d) and synthetic ($\mathbf{x}_d\widehat{\boldsymbol{\beta}}$) estimates. If the sampling error variance σ_d^2 of a district d is small relative to the random effects' variance $\hat{\sigma}_u^2$, then the weight given to the design-based direct estimates approaches 1,

$\hat{\gamma}_d \to 1$. Conversely, if the sampling error variance is relatively high, more weight is given to the synthetic estimates that are based on the model-based paradigm, and $\hat{\gamma}_d \to 0$.

1.3.4 The Log-Transformed Fay-Herriot Model

The application of a log-transformation on the data $(\log(y))$ to improve the normality of the data is part of the Box-Cox family of transformations. A discussion of this transformation family can be found in Fox (2015). Illustrative work on transformations with an explicit focus on SAE is conducted by Lee et al. (2023) and Rojas-Perilla et al. (2020). Kreutzmann et al. (2022) provide an application with wealth data. The log-transformed FH model assumes the relationship in Eq. (1.7) on a log scale of the direct estimates $z_d = \log(y_d)$. The model formulation in matrix form changes to

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \tag{1.12}$$

where $\mathbf{z}$ is a column vector of length D containing the log-transformed area-level direct estimator of the target variable indicator. The structure and distributional assumptions made in the standard FH model remain unchanged. The response variable is modeled on the log scale, but the σ_d^2 are usually estimated on the original scale. Using non-transformed σ_d^2 would distort the model. An adjustment of the σ_d^2 to address this problem is proposed by Neves et al. (2014). They transform the σ_d^2 to the log scale using Taylor series approximation. Their transformed variances have the form

$$\mathbb{V}\left[z_d\right] \approx \frac{\mathbb{V}\left[y_d\right]}{\left(\mathbb{E}[y_d]\right)^2}.$$

Since μ_d is based on Eq. (1.12), its BLUP predicts the quantity of interest on the log scale and therefore needs to be back-transformed to the original scale via the exponential function

$$\hat{\mu}_d^* = \exp(\hat{\mu}_d) = \exp\left(\mathbf{x}_d\widehat{\boldsymbol{\beta}} + \hat{u}_d\right).$$

Such a direct back-transformation does not consider bias induced by the nonlinear nature of the logarithm. If maximum likelihood (ML) is used for the estimation of σ_u^2, Slud and Maiti (2006) propose a bias correction of the form

$$\hat{\mu}_d^{*SM} = \exp\left(\mathbf{x}_d\widehat{\boldsymbol{\beta}} + \hat{u}_d + \frac{1}{2}\hat{\sigma}_u^2(1-\hat{\gamma}_d)\right), \tag{1.13}$$

accompanied by an analytical estimator for its MSE. The estimator $\hat{\mu}_d^{*SM}$ is still biased, but its bias vanishes for large D. For the EBLUE and EBLUP in Eq. (1.10), no closed-form expressions are available; this also implies that σ_u^2 remains to be estimated by some non-analytical estimation procedure. Several estimation strategies such as method of moments (MM), ML, and restricted maximum likelihood (REML) are applicable (see e.g., Rao & Molina, 2015).

We focus on the ML method for this study so that the back-transformation by Slud and Maiti (2006) can be applied. For comparison, the MSE is standardized to

$$CV_d^{FH} = \frac{\sqrt{\widehat{MSE}_d\left(\hat{\mu}_d^{*SM}\right)}}{\hat{\mu}_d^{*SM}}, \quad \forall \hat{\mu}_d^{*SM} > 0. \tag{1.14}$$

We later use a measure of relative variability

$$RV_d = 1 - \frac{CV_d^{FH}}{CV_d^{HJ}}, \tag{1.15}$$

to contrast CV_d^{FH} and CV_d^{HJ}.

1.3.5 Variable Selection

There exists a rich literature on selecting explanatory variables in regression contexts (e.g., Hastie et al., 2009; Heinze et al., 2018). Considering the datasets presented in Section 1.2, we face the curse of dimensionality (Bellman, 1961). Combining pre-processed data from Destatis and INKAR results in 1 092 potential area-level covariates (552 and 540, respectively) for 1 590 domain observations; together with the three sum-to-zero mode contrasts (see Section 1.3.3), the candidate pool comprises 1 095 variables. Directly searching for interactions and polynomials among all candidates would create an intractable covariate space—with $p = 1{,}092$ area-level covariates, pairwise interactions alone would add $\binom{p}{2} \approx 596{,}000$ terms.

We address this challenge using a three-phase funnel pipeline that first selects main effects, then expands to higher-order terms and selects among them, and finally selects the best candidate via cross-validation on the FH model. This approach integrates three methodological contributions: (1) the multi-step LASSO procedure of Ren et al. (2022) for variable selection in area-level models, (2) the ranked sparsity LASSO of Peterson and Cavanaugh (2022) for principled selection of interactions and polynomials, and (3) a two-stage selection strategy similar to Wu et al. (2010), who call their approach "screen and clean" in the context of genome-wide association studies. The two-stage structure makes the overall procedure hierarchy-preferring (also called heredity-preferring): by restricting Phase 2 expansion to covariates selected in Phase 1, interactions are only considered among variables that demonstrated marginal predictive value (Bien et al., 2013). This complements the ranked sparsity penalties in Phase 2, which further favor main effects over higher-order terms. Neither mechanism strictly enforces hierarchy—Phase 2 can still drop a main effect while retaining its interactions if the data strongly support this.

Phase 1: Main effect selection via LASSO. The first phase reduces the candidate pool to a manageable set of main effects. Following Ren et al. (2022), we apply three preprocessing steps before regularized regression. First, metric covariates are standardized by centering and

scaling, while proportion covariates (bounded between 0 and 1) remain unscaled. Second, covariates with near-zero variance are removed, as they provide no predictive value. Third, highly correlated variable pairs (Pearson $|r| > 0.7$) are identified and the variable with the largest mean absolute correlation to all others is removed, following `caret::findCorrelation`. The three mode contrast variables (sum-to-zero coding) are protected from filtering to preserve the multimodal domain structure.

After preprocessing, we apply the LASSO (Tibshirani, 1996) to select main effects. The LASSO estimates coefficients by minimizing

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left(\sum_{d=1}^{D} \left(z_d - \sum_{j=1}^{p} x_{dj}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

where the penalty parameter λ controls the strength of the ℓ_1 constraint. The optimal λ is selected via 10-fold cross-validation using `glmnet::cv.glmnet` (Tay et al., 2023). We use $\lambda_{\min}$, which minimizes cross-validated prediction error.

Phase 2: Ranked sparsity LASSO for interactions and polynomials. Standard LASSO treats all candidate terms equally—an assumption termed "covariate equipoise"—which can lead to over-selection of interactions and polynomials relative to main effects (Peterson & Cavanaugh, 2022). When the candidate set includes p main effects and $\binom{p}{2}$ pairwise interactions, the sheer number of interaction terms inflates their collective prior probability of selection, resulting in unnecessarily complex models with many spurious higher-order terms.

To address this, Phase 2 applies the sparsity-ranked LASSO (SRL) via the `sparseR` package (Peterson & Cavanaugh, 2022). The SRL partitions predictors into K groups (e.g., main effects, polynomials, interactions) and assigns differential penalty weights to each group. Formally, the SRL minimizes

$$\hat{\boldsymbol{\beta}}^{\text{SRL}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left(\sum_{d=1}^{D} \left(z_d - \sum_{j=1}^{p} x_{dj}\beta_j \right)^2 + \lambda \sum_{k=1}^{K} \sum_{j \in G_k} w_k |\beta_j| \right),$$

where G_k denotes the set of predictors in group k, $m_k = |G_k|$ is its cardinality, and w_k is the group-specific penalty weight. Setting $w_k = \sqrt{m_k}$ ensures that each group contributes equally to the prior information, counteracting the inflation caused by larger groups (Peterson & Cavanaugh, 2022). This corresponds to fixing the ranked sparsity parameter $\gamma = 0.5$ in the general formulation $w_k = m_k^{\gamma}$.

We further enable cumulative penalties, which increase the penalty for higher-order terms based on the cumulative count of all terms up to that order, requiring stronger evidence for interactions and polynomials to enter the model. The `sparseR` implementation treats polynomial and

interaction penalties separately, with details provided in Section A.3.

Using the Phase 1 covariates as inputs, we construct expanded model matrices under two configurations. The baseline configuration contains squared terms for all continuous covariates and mode $\times$ covariate interactions. The maximal configuration additionally includes cubic polynomials and covariate $\times$ covariate interactions. We test both configurations to empirically evaluate whether covariate synergies (e.g., income $\times$ urbanity) improve prediction. The theoretical expectation is that they will not: since domains are defined by district $\times$ mode, covariate effects that vary across modes are directly predictive of the mode-specific outcome, whereas covariate synergies would need to operate uniformly across all four transport modes to improve prediction—an implausible assumption given the fundamental behavioral differences between walking, cycling, car travel, and public transit. Cross-validation adjudicates between the configurations.

Cross-validation within each configuration produces two candidate variable sets: $\lambda_{\min}$ (minimizing prediction error) and $\lambda_{1\text{se}}$ (within one standard error of the minimum, favoring parsimony). This yields four Phase 2 candidates in total. The specific variable counts are reported in Table 1.3.

Phase 3: Model selection via cross-validation. Following Ren et al. (2022), the final model is selected by comparing candidate variable sets through k-fold cross-validation on the FH model itself, rather than on the LASSO objective. This step serves two purposes. First, it re-evaluates variable sets under the actual FH model structure, which incorporates domain-specific random effects and sampling variances estimated from the direct estimates that the LASSO objective does not account for. Second, because the five candidate sets originate from three different LASSO configurations—Phase 1 base, Phase 2 baseline, and Phase 2 maximal—each operating on a different input matrix, there is no common LASSO-based criterion to rank them against each other. The FH-based cross-validation provides such a common metric.

Five candidate models are compared: (1) the base model using only Phase 1 main effects plus mode contrasts, (2–3) the $\lambda_{\min}$ and $\lambda_{1\text{se}}$ models from the baseline Phase 2 configuration, and (4–5) the $\lambda_{\min}$ and $\lambda_{1\text{se}}$ models from the maximal Phase 2 configuration. Domains are sorted by sample size and distributed across $K = 10$ folds to ensure balanced reliability across folds. For each fold, the FH model is fit on the training domains and predictions of the true domain parameters are made for validation domains using the fixed effects only ($\hat{\mu}_d = \mathbf{x}_d'\hat{\boldsymbol{\beta}}$). The sum of squared errors on reliable domains ($CV_d^{HJ} < 20\%$, cf. Eq. (1.4)) serves as the selection criterion. The candidate with the lowest total sum of squared errors (SSE) across folds is selected as the final model.

1.4 Results

1.4.1 Direct Estimates

Table 1.2 presents summary statistics of the HJ estimates, their CVs, and domain sample sizes by transport mode. The distributions differ markedly: car travel shows the highest MPK with a median of 26.4 km/day, followed by public transit (4.5 km/day), cycling (0.9 km/day), and walking (0.9 km/day). Precision varies substantially across modes. Using a CV threshold of 20% to define acceptable precision, car travel achieves the best coverage with 62.8% of domains meeting this criterion, followed by walking (48.8%), cycling (21.9%), and public transit (11.4%). Overall, 63.7% of all domain estimates exceed the 20% threshold, highlighting the need for model-based approaches. Panel C reports domain sample sizes, defined as the number of surveyed persons who reported at least one trip by that mode. Sample sizes vary substantially across modes: car travel has the largest median (246 persons per domain), while cycling (43) and public transit (34) have considerably smaller samples, explaining their higher CVs.

Panel A: Direct Estimates (km/day)						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Walking	0.00	0.72	0.88	0.92	1.08	3.81
Cycling	0.00	0.54	0.91	1.02	1.34	4.51
Car (MIV)	2.18	21.52	26.41	27.11	31.15	74.37
Public Transit	0.00	2.71	4.46	5.31	6.94	29.78
Panel B: Coefficient of Variation						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Walking	0.03	0.15	0.20	0.24	0.28	1.03
Cycling	0.04	0.21	0.31	0.36	0.44	1.14
Car (MIV)	0.04	0.13	0.17	0.20	0.24	0.94
Public Transit	0.05	0.26	0.38	0.41	0.53	1.07
Panel C: Domain Sample Sizes (persons)						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Walking	0	51	109	219	258	4967
Cycling	0	18	43	102	103	3046
Car (MIV)	2	109	246	418	530	5790
Public Transit	0	15	34	97	87	4419

Table 1.2: Summary statistics of multimodal HJ direct estimates. Panel A: person-kilometers per day. Panel B: coefficient of variation (CV). Panel C: domain sample sizes (persons with at least one trip by that mode). Total domains: 1604 (401 districts $\times$ 4 modes), of which 10 have zero observations for that mode.

Figure 1.1 illustrates the precision challenge using public transit as an example—the mode with the poorest direct estimation precision. Panel (a) shows the spatial distribution of direct estimates, while panel (b) displays the corresponding CVs. Many rural districts have few or no observed transit trips in the sample, resulting in CVs exceeding 60% (shown in red). This pat-

tern motivates the use of model-based estimation, which can borrow strength across domains. Complete spatial distributions for all four modes are provided in Section A.1.

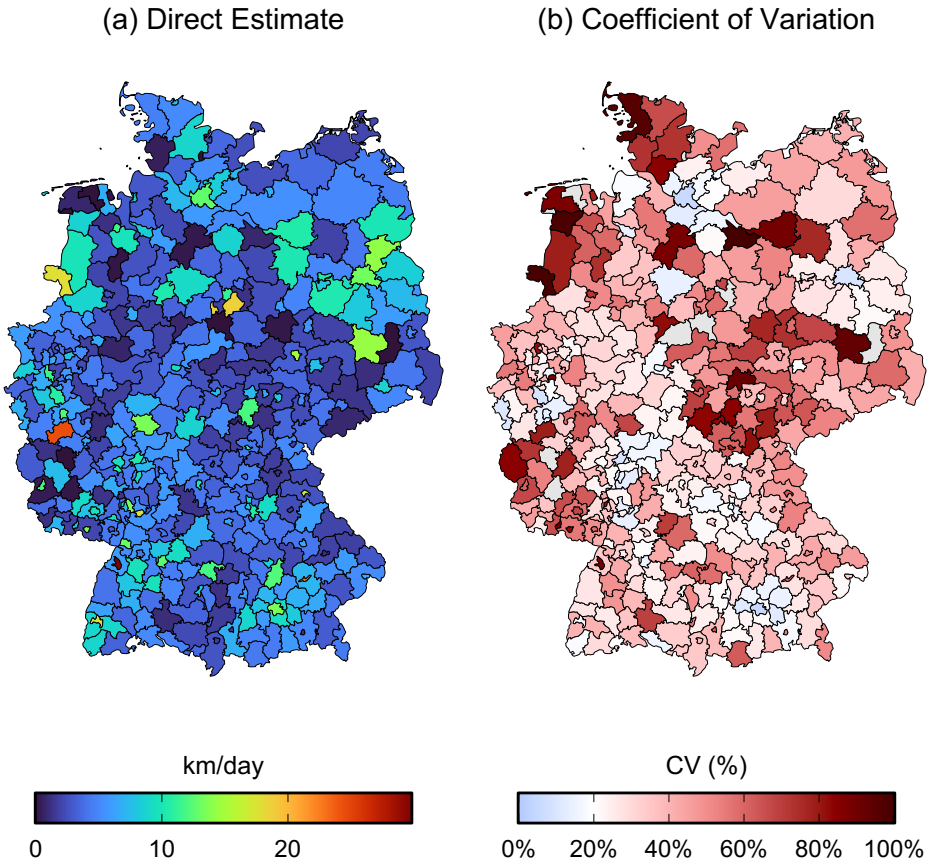


Figure 1.1: Public transit HJ direct estimates across German districts: (a) MPK estimates, (b) CV with the 20% threshold separating reliable (blue/green) from unreliable (yellow/red) estimates.

1.4.2 Variable Selection

Table 1.3 summarizes the results of the three-phase procedure described in Section 1.3.5. Phase 1 LASSO reduces the 1 095 candidate covariates to a manageable set of main effects. Phase 2 applies ranked sparsity LASSO to select interactions and polynomial terms under two configurations (baseline and maximal), producing four candidate variable sets at different regularization levels. In Phase 3, cross-validation selects the baseline $\lambda_{\min}$ model as the preferred FH specification by minimizing prediction error across reliable domains.

Table 1.4 compares the five candidate variable sets. The baseline $\lambda_{\min}$ model achieves the lowest SSE on reliable domains, outperforming the maximal configuration despite its additional covariate $\times$ covariate interactions. The winning model includes main effects from census demographics and INKAR regional indicators, augmented by mode $\times$ covariate interactions and squared terms for continuous covariates. This composition reflects the multimodal domain

Phase	Description	Input	Output
1	LASSO main effects	1,095 candidates	49 covariates (R^2 = 0.86)
2a	sparseR baseline	233 terms (poly^2, mode$\times$covar)	$\lambda_{\min}$: 122 vars, λ_{1se}: 52 vars
2b	sparseR maximal	1,314 terms (+poly^3, +covar$\times$covar)	$\lambda_{\min}$: 107 vars, λ_{1se}: 44 vars
3	FH cross-validation	5 candidate sets (K=10)	Winner: baseline $\lambda_{\min}$

Table 1.3: Three-phase variable selection pipeline. Phase 1 reduces candidates via LASSO on main effects. Phase 2 applies ranked sparsity LASSO (sparseR) with two configurations: (a) baseline with quadratic polynomials and mode$\times$covariate interactions; (b) maximal adding cubic polynomials and covariate$\times$covariate interactions. Phase 3 selects the final model via 10-fold cross-validation on the Fay-Herriot model.

structure, allowing mode-specific covariate effects where supported by the data.

Config	Candidate	Variables	Main	Poly	Inter	SSE
–	Base	49	49	0	0	21,632
Baseline	λ_{1se}	52	27	6	19	7,955
Baseline	$\lambda_{\min}$	122	37	22	63	**7,208**
Maximal	λ_{1se}	44	30	8	6	9,185
Maximal	$\lambda_{\min}$	107	40	28	39	7,590

Table 1.4: Cross-validation comparison of five candidate variable sets. SSE calculated on reliable domains (CV $< 20\%$). Baseline configuration uses quadratic polynomials and mode$\times$covariate interactions only. Maximal configuration adds cubic polynomials and covariate$\times$covariate interactions. Winner (baseline $\lambda_{\min}$) shown in bold.

A notable aspect of the variable selection approach is that it is data-driven rather than guided by pre-existing theoretical knowledge. Some selected covariates intuitively relate to mobility patterns (e.g., transit quality indicators, car ownership), while others capture indirect relationships through regional socioeconomic characteristics. The primary objective is prediction of unobserved domain values rather than causal identification. Shmueli (2010) provides a comprehensive discussion of the distinction between explanatory and predictive modeling in statistics. The complete list of variables in the final model is provided in Section A.4.

1.4.3 Model Estimates

The bias-corrected back-transformed EBLUP estimates $\hat{\mu}_d^{*SM}$ (Eq. (1.13)) from the log-transformed FH model reveal distinct spatial patterns in German district-level mobility. Figure 1.2 maps these estimates by transport mode, highlighting the geographic structure of daily travel behavior.

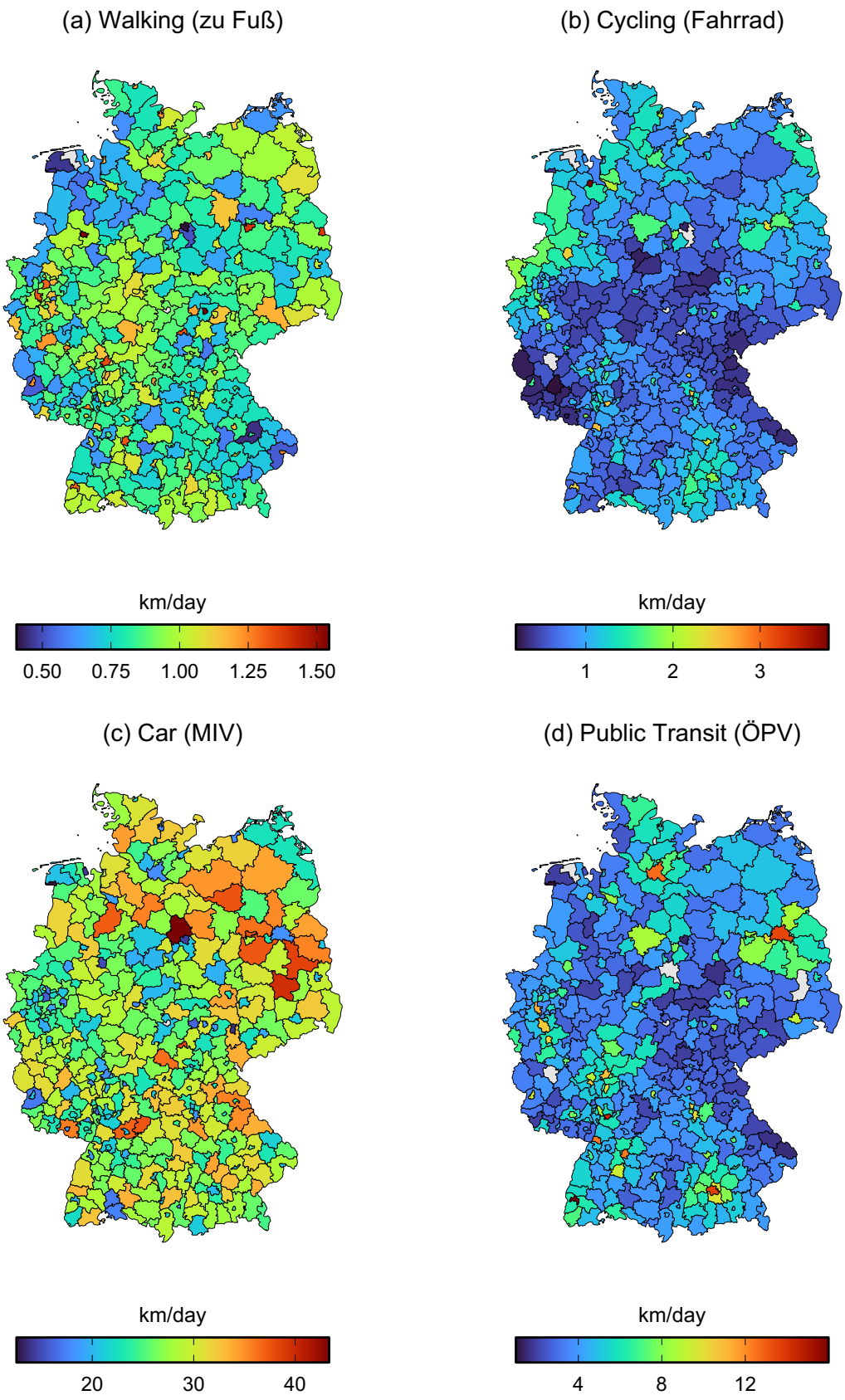


Figure 1.2: Spatial distributions of FH MPK estimates by transport mode across German districts: (a) walking (*zu Fuß*), (b) cycling (*Fahrrad*), (c) car (*MIV*), and (d) public transit (*ÖPV*).

The spatial patterns differ markedly across modes and reflect underlying regional characteristics. Car travel displays a pronounced east–west gradient, with the highest values in rural north-eastern Germany—particularly in sparsely populated districts of Mecklenburg-Vorpommern and Brandenburg—where longer distances to employment and services necessitate greater car dependence. The densely populated Ruhr area and major metropolitan regions show lower car distances, consistent with shorter commutes and greater availability of alternative modes. Public transit usage concentrates in metropolitan areas with established rail and bus networks: Berlin, Hamburg, Munich, Frankfurt, and the Rhine-Ruhr conurbation stand out with elevated values. In contrast, many rural districts—especially in eastern Germany—show minimal public transit usage, reflecting limited service availability. Walking and cycling estimates are relatively homogeneous across Germany, with median values around 0.9 km/day for both modes. Cycling shows slightly higher values in northwestern Germany near the Dutch border—where the Netherlands' renowned cycling culture appears to spill over—as well as in northern coastal regions and university cities, where flat terrain and established cycling infrastructure encourage bicycle use.

Table 1.5 quantifies these patterns. Panel A confirms the modal hierarchy: car travel dominates daily mobility with a median of 26.4 km/day, followed by public transit (4.2 km/day), cycling (0.9 km/day), and walking (0.9 km/day). The near-identical values for walking and cycling are not coincidental: cycling trips are 2.4 times longer on average, but walking participation is 2.3 times higher (27% vs. 12% of the population on a given day), and these factors nearly cancel out. This distribution reflects car dependence in German daily travel, where private vehicles account for the majority of passenger-kilometers despite the availability of alternatives in urban areas. Complete district-level results across all four transport modes are available in an interactive Shiny application at https://github.com/muehlbauer-m/sae-mobility-germany.

Compared to the direct estimates in Figure A.1, the FH estimates show less severe variation between neighboring districts and fewer extreme values. This spatial smoothing effect is most apparent for modes with high direct estimate variance, where isolated extreme values in the direct estimates are moderated toward regional patterns. Figure 1.4 illustrates this shrinkage effect: regressing FH estimates on direct estimates yields slopes $\hat{\beta}_1 < 1$, indicating compression toward the mean, i.e., the FH estimates are less dispersed than the direct estimates. This compression is strongest for walking ($\hat{\beta}_1 = 0.37$), followed by public transit ($\hat{\beta}_1 = 0.40$), while cycling ($\hat{\beta}_1 = 0.50$) and car travel ($\hat{\beta}_1 = 0.48$) show the least compression.

Panel C of Table 1.5 reports the distribution of $\hat{\gamma}_d$ by mode, and Figure 1.3 visualizes these distributions. Public transit and cycling show strongly right-skewed distributions concentrated near zero, indicating heavy reliance on the synthetic predictor. Car travel shows a more uniform distribution, reflecting greater direct estimate reliability. Walking falls between these extremes. The relationship between the shrinkage factor $\hat{\gamma}_d$ and the apparent compression in Figure 1.4

Panel A: FH EBLUP Estimates (km/day)						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Cycling	0.20	0.63	0.88	0.95	1.17	3.78
Walking	0.41	0.76	0.87	0.88	0.99	1.54
Car	12.59	22.67	26.45	26.27	29.88	43.29
Public Transit	1.03	3.14	4.22	4.73	5.65	15.97
Panel B: Coefficient of Variation						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Cycling	0.03	0.16	0.20	0.19	0.22	0.34
Walking	0.04	0.13	0.16	0.16	0.19	0.32
Car	0.04	0.12	0.14	0.15	0.17	0.33
Public Transit	0.05	0.19	0.22	0.21	0.24	0.37
Panel C: Shrinkage Factor ($\hat{\gamma}_d$)						
Mode	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Cycling	0.03	0.16	0.28	0.32	0.46	0.97
Walking	0.03	0.32	0.48	0.48	0.63	0.98
Car	0.04	0.40	0.57	0.54	0.70	0.96
Public Transit	0.03	0.12	0.21	0.26	0.36	0.93

Table 1.5: Summary statistics of multimodal FH EBLUP estimates. Panel A shows bias-corrected back-transformed estimates (km/day) by transport mode. Panel B shows the coefficient of variation. Panel C shows the shrinkage factor $\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \sigma_d^2)$, where higher values indicate more weight on the direct estimate. Total domains: 1590 (400 districts $\times$ 4 modes, after exclusions).

is not straightforward. Car travel has the highest mean $\hat{\gamma}_d$ (0.54) and comparatively little compression ($\hat{\beta}_1 = 0.48$), consistent with a large weight on direct estimates. Walking, however, shows a similarly high mean $\hat{\gamma}_d$ (0.48) but the strongest compression ($\hat{\beta}_1 = 0.37$). This pattern arises because $\hat{\gamma}_d$ governs how much weight the direct estimate receives, but the scatter plot slope also depends on how far the synthetic predictor $\mathbf{x}_d'\hat{\boldsymbol{\beta}}$ lies from the direct estimate. Since the synthetic predictor is fitted to the bulk of the data, it predicts values near the center of the distribution. For walking, where the direct estimates are highly dispersed, the gap between extreme direct estimates and the central synthetic predictor is large, so even a moderate synthetic weight of $(1 - \hat{\gamma}_d) = 0.52$ can produce strong compression.

The fill color in Figure 1.4 encodes the CV of the direct estimate. Domains with low CV (dark fill) generally cluster close to the bisectrix across all four modes, confirming that the model agrees well with direct estimates where the latter are precise. Domains that are both reliable and far from the bisectrix—which would indicate local model misfit—are largely absent.

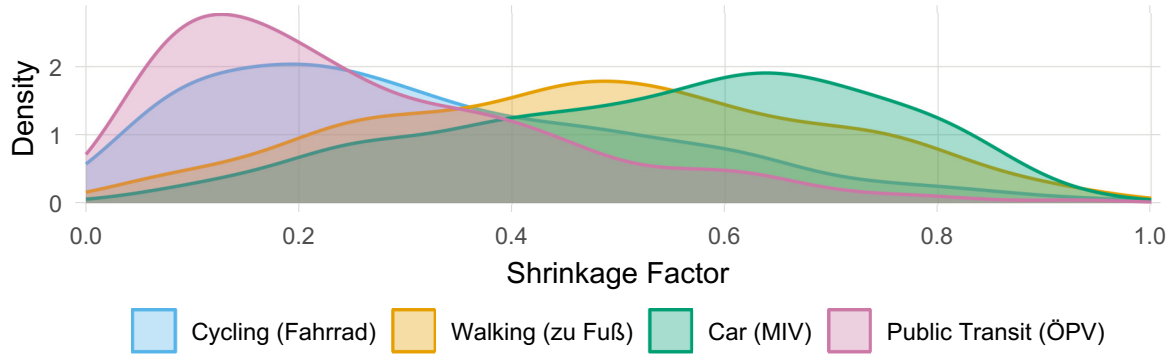


Figure 1.3: Distribution of the shrinkage factor $\hat{\gamma}_d$ by transport mode. Higher values indicate greater weight on the direct estimate; lower values indicate greater weight on the synthetic predictor $\mathbf{x}_d'\hat{\boldsymbol{\beta}}$.

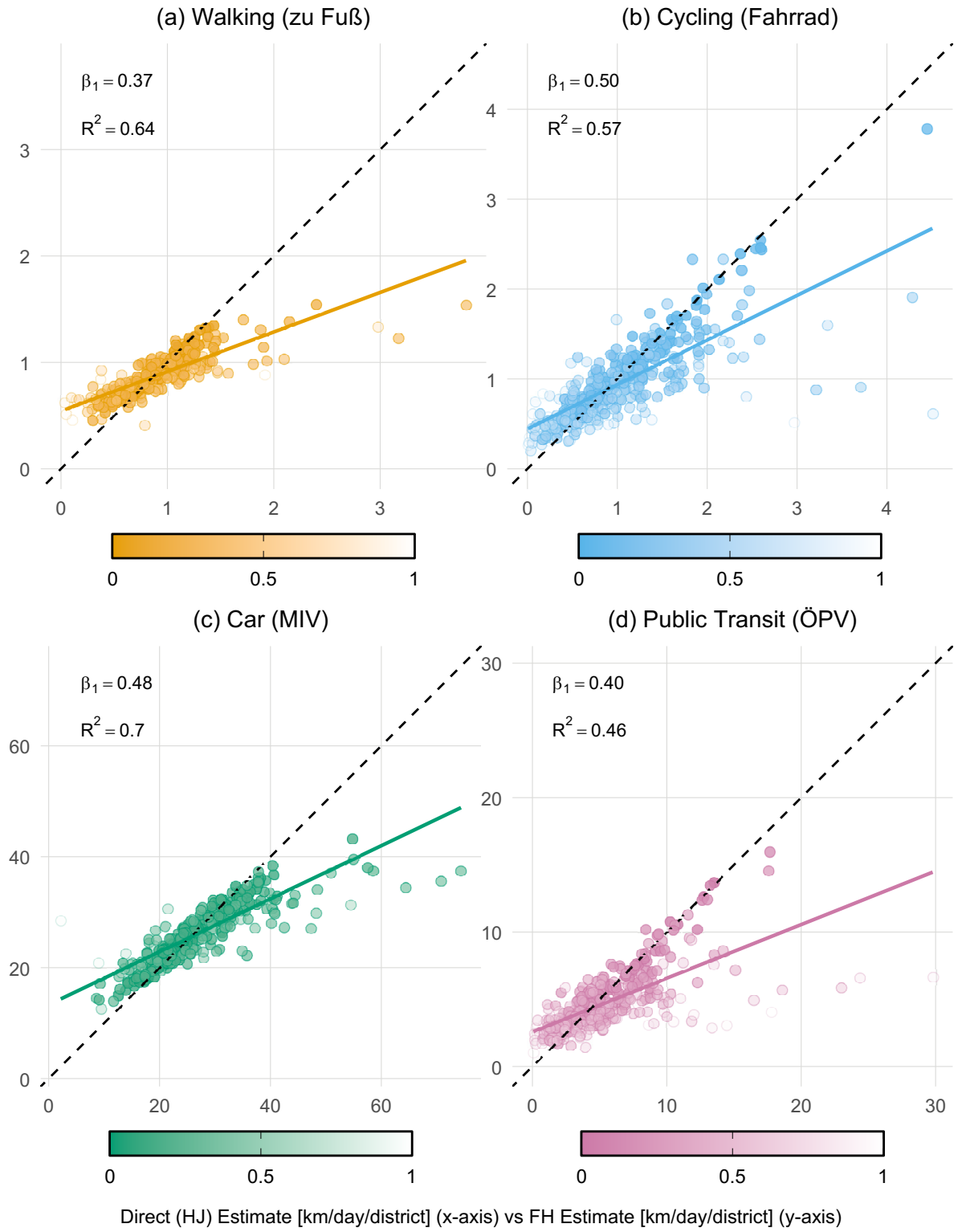


Figure 1.4: Scatter plot of FH versus HJ estimates by transport mode. Dashed line indicates perfect agreement; colored lines show fitted linear regression slopes. Fill color indicates the CV of the direct estimate: dark = low CV (reliable), light = high CV (unreliable).

The model-based approach yields substantial precision improvements. Panel B of Table 1.5 shows that FH CVs are substantially lower than their HJ counterparts: whereas 63.7% of direct

estimates exceed the 20% CV threshold (Section 1.4.1), only 37.0% of FH estimates do so. By mode, the share of domains achieving acceptable precision (CV $< 20\%$) increases from 62.8% to 88.2% for car travel, from 48.8% to 78.4% for walking, from 21.9% to 51.3% for cycling, and from 11.4% to 33.4% for public transit. Figure 1.5 displays the CV distributions by method and mode.

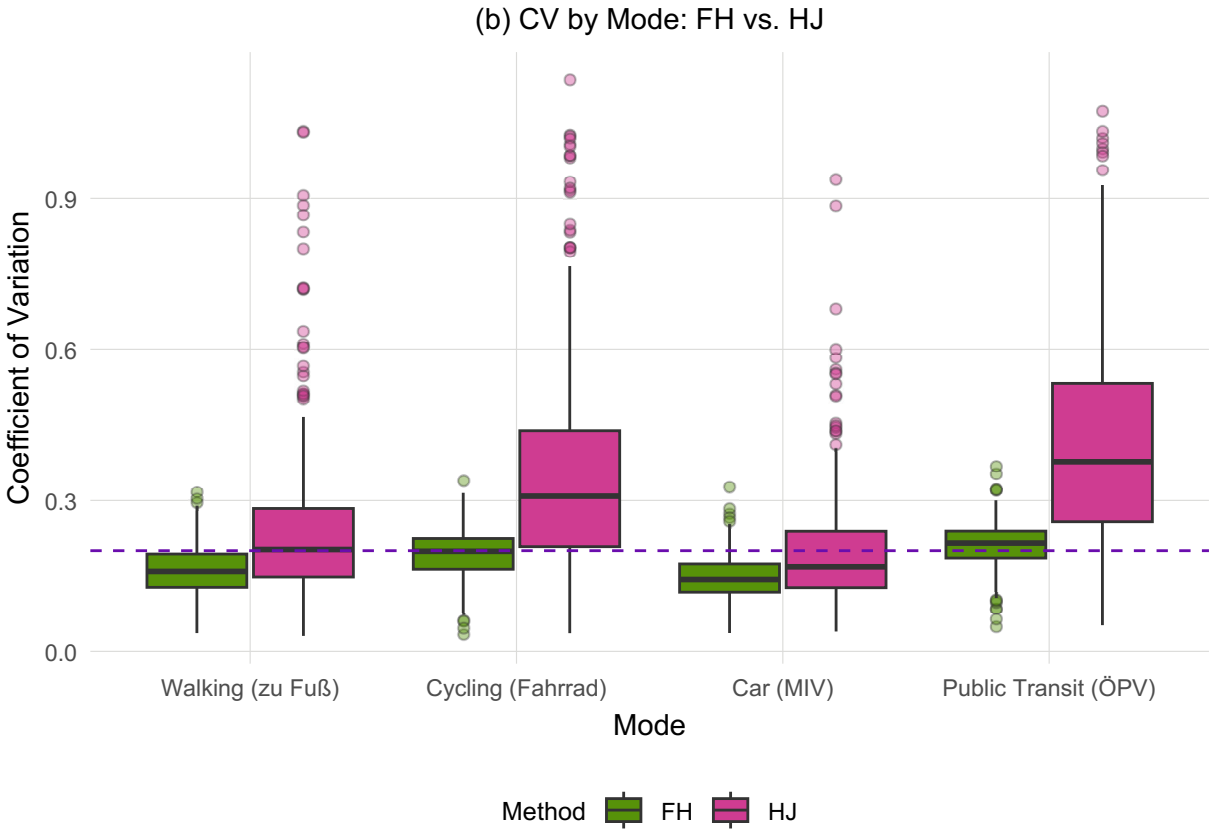


Figure 1.5: Boxplots of CV distributions by method and mode, with dashed line at the 20% threshold.

Table 1.6 quantifies the precision improvement using the relative variability reduction RV_d (Eq. (1.15)). Positive values indicate that the FH model improves precision over direct estimation. The median RV ranges from 0.16 (car) to 0.44 (public transit), indicating that the FH model typically reduces the CV by 16–44%. The improvement is largest for public transit, where direct estimates had the poorest precision, and smallest for car travel, which already had relatively reliable direct estimates. A total of 158 of 1 590 domains (9.9%) show negative RV values. These occur predominantly in domains with already-precise direct estimates (high $\hat{\gamma}_d$), where the MSE of the FH estimator—which accounts for uncertainty in estimating σ_u^2 and $\boldsymbol{\beta}$—can exceed the sampling variance of the direct estimator. Nevertheless, these deteriorations are small in magnitude.

Figure 1.6 maps the efficiency gains by mode. Green colors indicate precision improvement (positive RV), while pink/red would indicate deterioration. The maps confirm that precision

Mode	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cycling	-0.47	0.22	0.37	0.36	0.52	0.78
Walking	-0.28	0.09	0.21	0.22	0.35	0.77
Car	-0.41	0.04	0.16	0.17	0.30	0.72
Public Transit	-0.23	0.26	0.44	0.41	0.57	0.76

Table 1.6: Summary statistics of the relative variability reduction $RV_d = (CV_d^{HJ} - CV_d^{FH})/CV_d^{HJ}$ by transport mode. Positive values indicate precision improvement from the FH model. Domains with negative RV: 158 out of 1590.

gains are widespread across Germany for all modes, with particularly strong improvements in rural eastern districts where direct estimates were least reliable due to small sample sizes. The few domains with negative efficiency gains are scattered and show only relatively low deterioration.

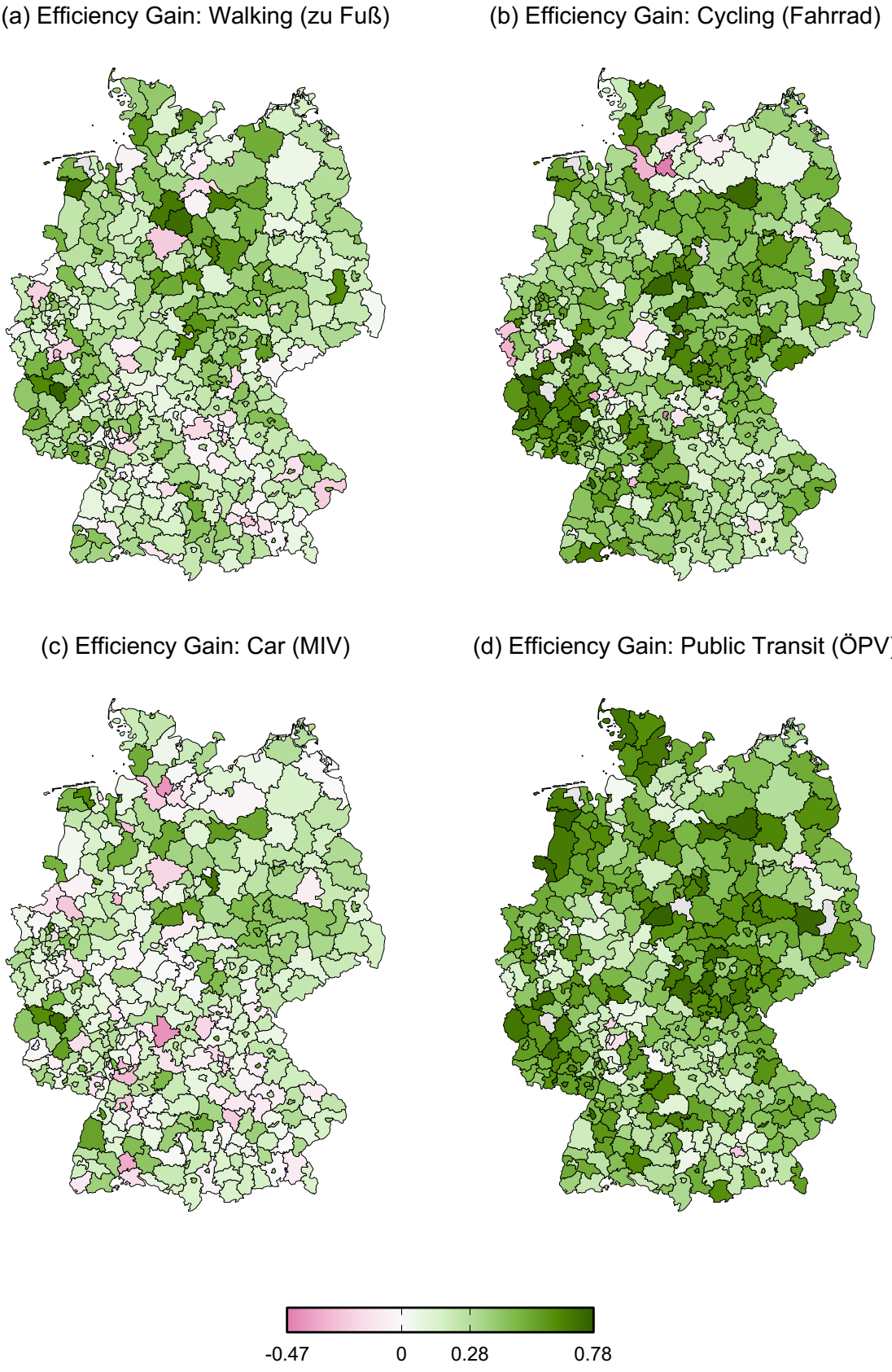


Figure 1.6: Spatial distributions of efficiency gain RV_d by transport mode: (a) walking, (b) cycling, (c) car, and (d) public transit. Green indicates precision improvement from the FH model; pink/red would indicate deterioration.

1.4.4 Model Diagnostics

Trusting the model-based MPK estimates requires that the assumptions in Section 1.3.3 and Section 1.3.4 are plausible. Key indicators of model quality include the R^2 measure and the fit to distributional (normality) assumptions. Lahiri and Suntornchost (2015) developed an adjusted R^2 specifically for FH models.

Before comparing variable sets, we motivate the log transformation by comparing base FH models with and without transformation. Both models use only Phase 1 LASSO-selected covariates (49 for log, 37 for non-log) to ensure a fair comparison. Figure 1.7 visualizes this comparison via kernel density estimations (KDEs) of the standardized residuals and random effects. Table 1.7 shows that the log-transformed model achieves substantially improved normality: kurtosis of the standardized random effects drops from 17.40 to 3.25 (target: 3.0), and the Shapiro-Wilk p-values increase by many orders of magnitude (e.g., from 2.17×10^{-36} to 2.04×10^{-3} for random effects). These results strongly favor the log transformation, which we adopt for all subsequent models.

	Base (log-transformed) (49 vars)	**Standard** (37 vars)
Variables	49	37
R^2 (adjusted)	85.5%	84.3%
$\hat{\sigma}_u^2$ (domain variance)	0.1014	0.3469
Skewness (Std. Residuals)	-1.0425	-0.2721
Skewness (Std. Random Effects)	-0.1493	-0.4737
Kurtosis (Std. Residuals)	6.28	11.40
Kurtosis (Std. Random Effects)	3.25	17.40
Shapiro-Wilk p (Residuals)	7.44e-25	8.82e-33
Shapiro-Wilk p (Random Effects)	2.04e-03	2.17e-36

Table 1.7: Comparison of distributional statistics between log-transformed and standard FH models. Both models use Phase 1 LASSO-selected variables.

The final baseline $\lambda_{\min}$ model shows acceptable normality diagnostics. The standardized random effects exhibit slight leptokurtosis (3.83 vs. target 3.0), while the standardized residuals show moderate leptokurtosis (4.43). The Shapiro-Wilk test (Shapiro & Wilk, 1965) rejects normality at the 5% level ($p < 0.01$). However, this rejection must be contextualized: with 1 590 domains, even minor deviations from normality cause rejection. Statistically significant deviations do not necessarily imply substantively significant deviations[2]. Figure 1.8 shows KDEs and quantile-quantile plots (QQ-Plots) for the final baseline $\lambda_{\min}$ model. The QQ-Plots reveal that the distributions closely follow the theoretical normal line across most of their range, with departures only in the tails. These tail deviations are typical in empirical applications.

[2]The Shapiro-Wilk test is designed as a test against normality, not for it. Moreover, the test is highly sensitive to sample size: with large n, even trivial departures from normality lead to rejection. High p-values can only be regarded as weak evidence for normality. Greenland et al. (2016) provide an extensive discussion.

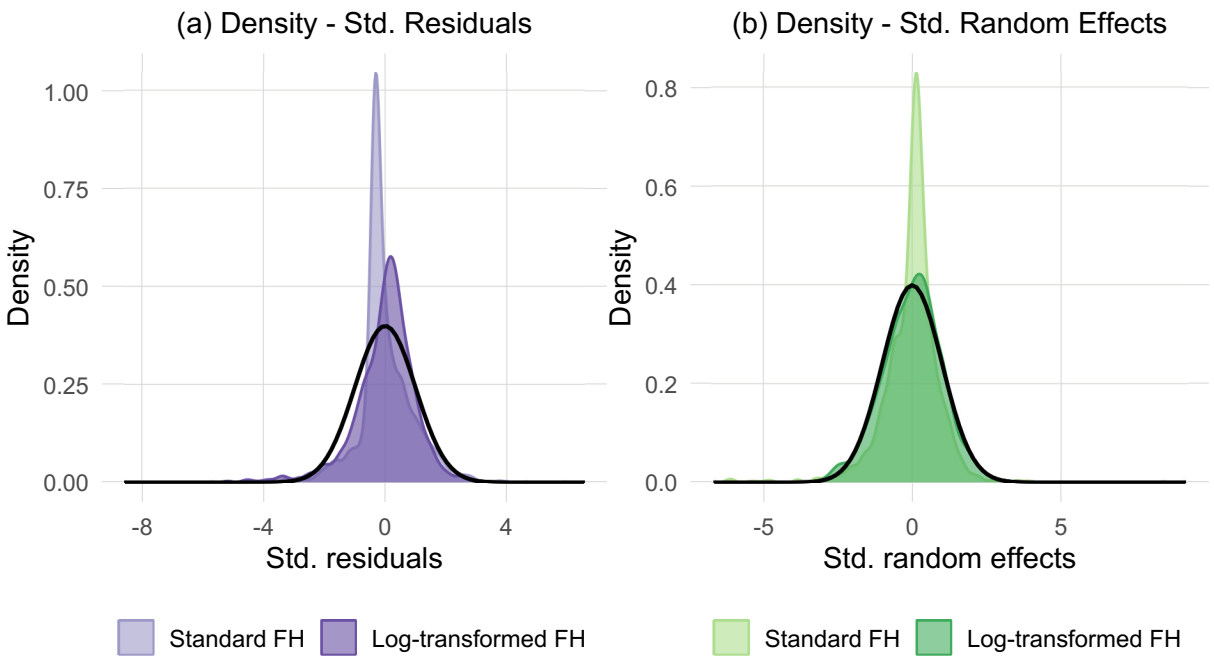


Figure 1.7: KDEs of the (a) std. residuals and (b) std. random effects of the standard FH model compared to the log-transformed FH model. Both models use Phase 1 LASSO-selected covariates.

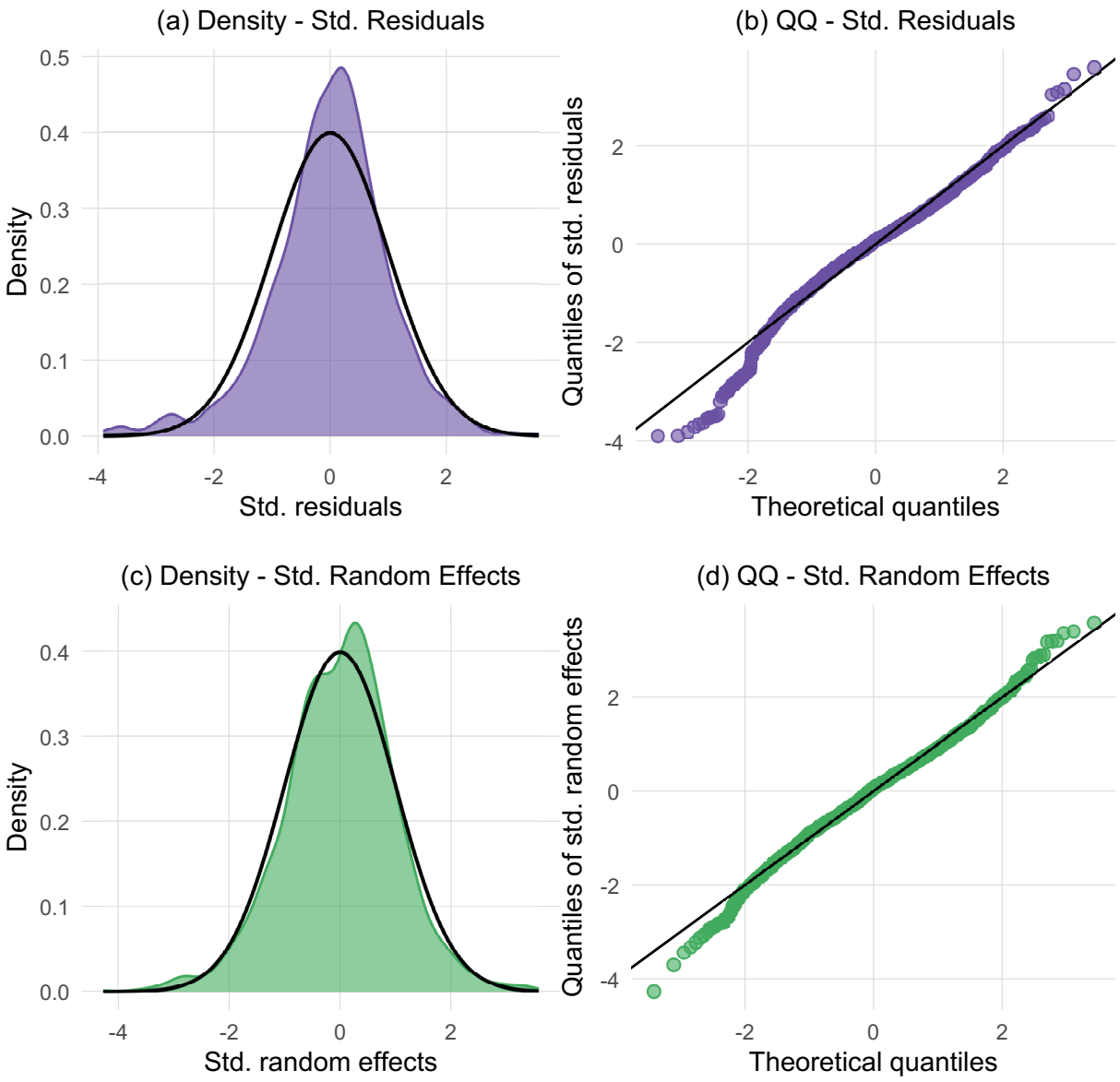


Figure 1.8: Normality diagnostics for the final baseline λ_{min} model: (a), (c) KDEs and (b), (d) QQ-Plots of the std. residuals and std. random effects.

1.5 Discussion

A fundamental consideration concerns the nature of statistical modeling. The variable selection approach employed here—combining LASSO regularization with cross-validation on the FH model—is inherently data-driven rather than theory-guided. This is consistent with the goals of SAE, which prioritizes optimal prediction for small domains over causal identification. Shmueli (2010) distinguishes such predictive modeling from explanatory modeling aimed at testing causal hypotheses. The selected covariates correlate with district-level mobility patterns but do not necessarily cause them. For instance, income-related variables may serve as proxies for car ownership and travel demand rather than direct determinants. Policy conclusions drawn from these results require careful interpretation: the model identifies statistical associations useful for prediction, not causal mechanisms that would inform intervention design.

Variance estimation for the direct estimates presents inherent challenges in complex survey contexts. While the Taylor series linearization approach employed here shows close agreement with calibrated bootstrap estimates ($R^2 = 0.96$; see Section A.2), the MiD documentation does not provide sufficient detail about the sampling design to construct proper replicate weights. Resampling weights calculated as part of official MiD data processing—as discussed by Rao and Wu (1988) and Rao et al. (1992)—would enable more precise variance estimation and facilitate broader analytical use of the survey data. Additionally, the stratification variable used in our analysis (MiD regional type) was reconstructed from publicly available regional classifications rather than obtained from the original survey design, introducing potential approximation error. Future releases of the MiD survey would benefit from including replicate weights and complete design documentation.

Two additional limitations should be acknowledged. First, the survey weights in MiD 2017 are calibrated to known population totals at the national level and for designated in-depth sampling areas, but not at the district level. This may introduce bias in the HJ estimates for districts whose demographic composition deviates substantially from national patterns. Second, the MiD 2017 survey covers the period from May 2016 to September 2017, capturing pre-pandemic mobility patterns that may not fully reflect current travel behavior. However, the auxiliary data from INKAR and Destatis are also from 2017, ensuring temporal alignment between survey outcomes and model covariates.

These considerations point to several directions for future research. From a methodological perspective, Burgard et al. (2019, 2021) develop a regularized FH model integrating elastic net penalties directly into coefficient estimation, framed as robustness to covariate measurement error. However, the method lacks software implementation, has not been extended to log-transformed models, and provides only conservative MSE bounds. More broadly, methods

for high-dimensional inference in linear mixed models exist (e.g., Li et al., 2022), but extending these to the FH setting with heterogeneous sampling variances and response transformations remains an open problem. From an applied perspective, data from the MiD 2023 survey are now available and would enable validation of the methodology developed here as well as temporal comparisons of German mobility patterns. With adequate auxiliary data, the methods presented can be applied at higher levels of geographic disaggregation, such as sub-district estimation or grid-based approaches (see Masaki et al., 2022, as an example). Finally, the mode-specific estimates produced by this analysis could inform targeted policy applications, such as identifying gaps in public transit coverage or prioritizing cycling infrastructure investments based on district-level cycling demand.

1.6 Conclusion

This study develops a multimodal SAE framework for estimating district-level MPK across four transport modes—walking, cycling, car travel, and public transit—from German national survey data. The methodological contribution is twofold: first, integrating transport modes as a domain dimension within the FH model rather than fitting separate models per mode; second, proposing a three-phase variable selection pipeline that combines standard LASSO for main effect screening, ranked sparsity LASSO for interactions and polynomials, and cross-validation on the FH model itself to select the final specification.

The empirical results demonstrate substantial precision gains from model-based estimation. Direct HJ estimates achieve acceptable precision (CV $< 20\%$) for only 11% of public transit domains, 22% of cycling domains, 49% of walking domains, and 63% of car travel domains. The FH model raises these shares to 33%, 51%, 78%, and 88%, respectively, with the largest improvements for public transit and cycling where survey coverage is sparsest. The final model achieves an FH R^2 of 93% (adjusted R^2: 89%) and shows acceptable normality diagnostics (random effect kurtosis: 3.83; target: 3.0).

The estimates fill a gap between national statistics, which mask regional heterogeneity, and local traffic counts, which capture flows but not person-level travel behavior. For instance, district-level MPK by mode can reveal where car dependence is highest, inform modal shift targets, or feed into spatially disaggregated emissions models. The spatial patterns confirm expected relationships: car dominance in rural northeastern Germany, public transit concentration in metropolitan cores, and elevated cycling in flat northern regions and university cities.

Application of this framework to the recently released MiD 2023 data would enable temporal comparisons and assessment of post-pandemic mobility shifts.

Declarations

Funding and/or Conflicts of interest/Competing interests

All authors declare no conflicts of interest.

Code Availability Statement

All analysis code is publicly available on GitHub at https://github.com/muehlbauer-m/sae-mobility-germany. The repository contains R scripts for data preparation, direct estimation, variable selection, Fay-Herriot model fitting, and visualization.

Data Availability Statement

The MiD 2017 dataset

The German Federal Ministry for Digital and Transport is allowed to pass on the anonymized datasets of the MiD sample under the established data protection requirements to the following user groups:

a) Commissioners of regional in-depth samples for the purpose of traffic policy, planning, or science,

b) Universities or other institutions with the task of independent scientific research for the execution of research projects and activities,

c) Other third parties who need the data for research purposes within the framework of a project commissioned publicly or a publicly funded project.

Users satisfying one of these conditions can obtain the MiD 2017 dataset through the official channel at: https://daten.clearingstelle-verkehr.de/279/.

Official Statistics from the German Statistical System

Administrative statistics were retrieved from GENESIS (2025) and RDB (2025) under the Data Licence Germany – Attribution – Version 2.0 (dl-de/by-2-0); own calculations. Both databases are publicly accessible at https://www-genesis.destatis.de/ and https://www.regionalstatistik.de/, respectively.

INKAR Regional Indicators

Regional indicator data were retrieved from the INKAR database of the BBSR (2025) under the Data Licence Germany – Attribution – Version 2.0 (dl-de/by-2-0); source: BBSR Bonn 2025; own calculations. The database is publicly accessible at https://www.inkar.de/.

A Appendix

A.1 Direct Estimate Maps by Transport Mode

Figure A.1 shows the spatial distributions of HJ direct estimates for all four transport modes. Car travel dominates in rural and peripheral districts, particularly in northeastern Germany, where distances to services and employment are greater. Urban areas show relatively lower car usage but higher public transit values. Walking and cycling estimates are more spatially homogeneous, though cycling shows higher values in northern coastal regions and university cities.

Figure A.2 displays the corresponding CVs. Car travel shows the best precision across most of Germany, with only scattered districts exceeding the 20% threshold. Walking and cycling display intermediate patterns, with lower precision in smaller districts. Public transit has the poorest precision, with many rural districts showing CVs exceeding 60%.

A.2 Bootstrap Variance Estimation

As an alternative to Taylor series linearization, variance estimates can also be obtained via bootstrap resampling. The bootstrap approach is computationally expensive, but provides a useful point of comparison for the analytical results. Three bootstrap variants are considered: (1) *naive bootstrap*, which uses sampling weights without recalibration; (2) *MiD-calibrated bootstrap*, which recalibrates each bootstrap sample to match MiD-weighted population totals; and (3) *census-calibrated bootstrap*, which recalibrates to external census totals from Destatis. The calibrated bootstrap procedure is depicted in Algorithm 1, explicitly described in Alfons and Templ (2012), and implemented in the `emdi::direct` function (Kreutzmann et al., 2019).

Algorithm 1 Calibrated bootstrap variance estimation

1: Draw B independent bootstrap samples $\mathcal{S}^{*,1}, \ldots, \mathcal{S}^{*,\,B}$ from $\mathcal{S}$.

2: Calibrate the sample weights for each bootstrap sample $\mathcal{S}^{*,\,b}$, $b \in \{1, \ldots, B\}$, so that estimated population totals $\hat{\tau}_d^{HT}(.)$ of a set $\mathcal{A}$ of calibration variables $x_{i,\,a}^{cal}$, $a \in \{1, \ldots, A\}$ are achieved. This ensures $\sum_{i \in \mathcal{S}_d} x_{i,\,a}^{cal} \cdot w_i = \sum_{i \in \mathcal{S}_d^{*,\,b}} x_{i,\,a}^{*,\,b,\,cal} \cdot w_i^{*,\,b,\,cal}$, $\forall a \in \mathcal{A}$ holds.

3: Compute the bootstrap replicate estimates $\hat{\mu}_d^{*,\,b,\,HJ} = \hat{\mu}_d^{*,\,b,\,HJ}\left(\mathcal{S}^{*,\,b}\right)$, for each bootstrap sample $b \in \{1, \ldots, B\}$ taking the calibrated sample weights $w_i^{*,\,b,\,cal}$ from the respective bootstrap samples into account.

4: Estimate the variance $\hat{\mathbb{V}}\left(\hat{\mu}_d^{HJ}\right)$ by the variance of the B bootstrap replicate estimates

$$\hat{\mathbb{V}}^{CB}\left(\hat{\mu}_d^{HJ}\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_d^{*,\,b,\,HJ} - \frac{1}{B} \sum_{s=1}^{B} \hat{\mu}_d^{*,\,s,\,HJ} \right)^2 .$$

Source: Adapted from Alfons and Templ (2012).

Table A.1 lists the calibration variables used in the bootstrap procedure. To validate the quality of the MiD `P_HOCH` weights, we compared the MiD-weighted population totals against external census population totals obtained from Destatis (2025). The comparison showed close agreement: mean deviations were approximately 1% for federal states, 0.7% for sex, 2.1% for age groups, 3.7% for survey quarter, and 1.4% for weekday. This confirms that the pre-calibrated MiD weights accurately reproduce known population benchmarks, justifying their use as the target totals in the bootstrap re-calibration step.

	Variable	**Type**	**Description**
1	BLAND	Factor	German federal state (16 categories)
2	HP_SEX	Factor	Sex (2 categories)
3	AGE_GROUP	Factor	Age group (5 categories)
4	ST_WOTAG	Factor	Weekday of survey date (7 categories)
5	ST_QUARTAL	Factor	Quarter of survey date (4 categories)

Table A.1: MiD variables used for the calibrated bootstrap variance estimation. Population totals for these variables were validated against external administrative statistics from Destatis.

Figure A.3 compares variance estimates across the four methods ($B = 500$ bootstrap replications). Panel (a) shows Taylor series linearization versus MiD-calibrated bootstrap: the two approaches agree closely ($R^2 = 0.96$), and the points lie predominantly below the identity line—the bootstrap standard error is smaller than its analytical counterpart in 82% of domains, with a median ratio of 0.92—indicating that Taylor series linearization produces slightly larger standard errors than bootstrap methods. The difference reflects the fact that linearization jointly accounts for clustering, stratification, unequal weights, and calibration, whereas the bootstrap only incorporates calibration through weight recalibration. Panel (b) compares census-calibrated against MiD-calibrated bootstrap, revealing near-perfect agreement ($\beta_1 \approx 1.00$, $R^2 > 0.99$). Panel (c) shows that naive bootstrap estimates are virtually identical to census-calibrated estimates ($\beta_1 \approx 1.00$, $R^2 > 0.99$). These results demonstrate that the choice of calibration targets—whether MiD-weighted totals, external census totals, or no recalibration—has negligible impact on bootstrap variance estimates. The close agreement between two methodologically distinct approaches lends credibility to the analytical variance estimates used in the main analysis.

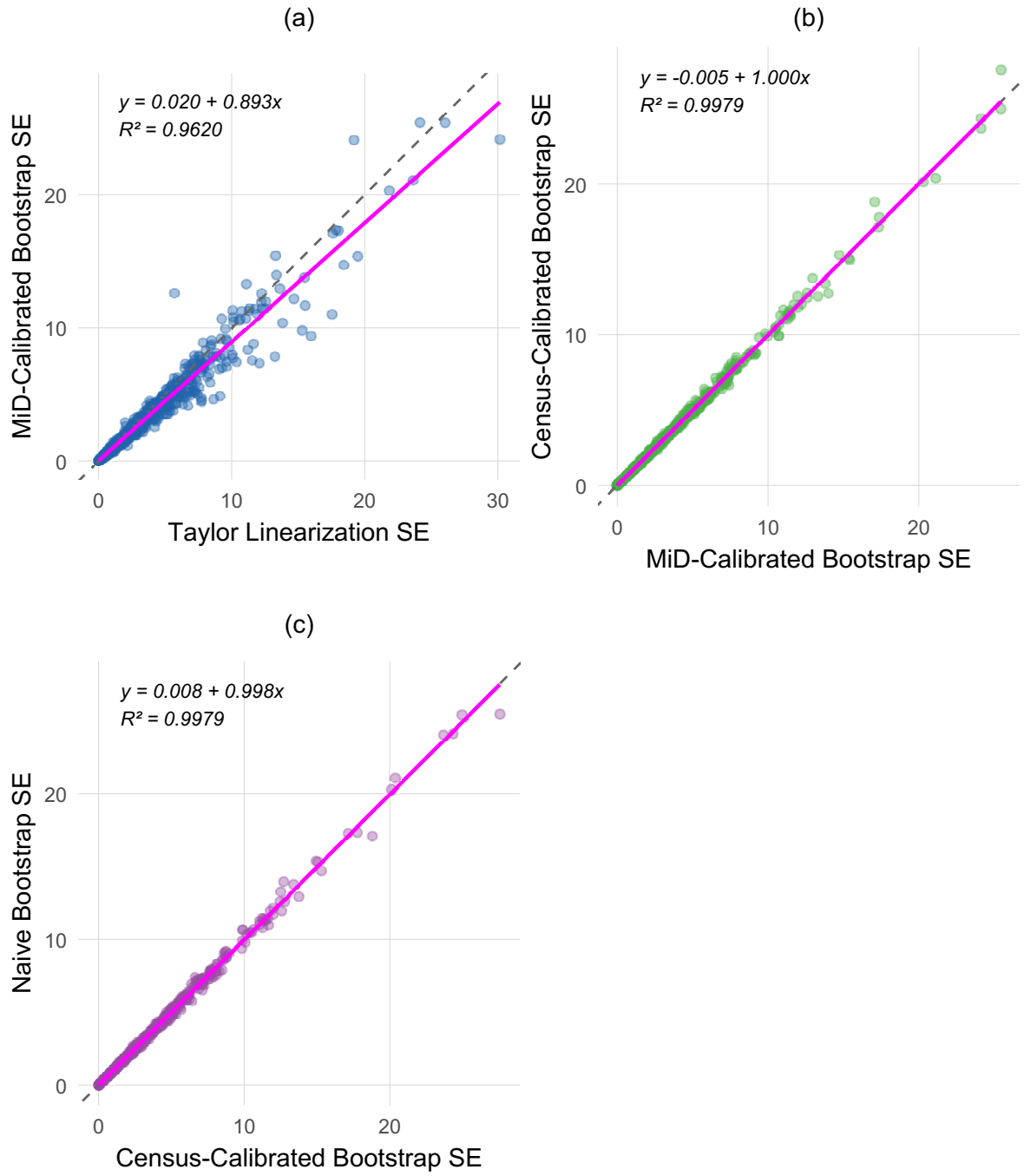


Figure A.3: Comparison of variance estimation methods. Panel (a): Taylor series linearization versus MiD-calibrated bootstrap. Panel (b): Census-calibrated versus MiD-calibrated bootstrap. Panel (c): Naive versus census-calibrated bootstrap. Dashed grey lines indicate perfect agreement; solid magenta lines show fitted regression. All bootstrap methods produce virtually identical estimates regardless of calibration approach.

A.3 Cumulative Penalty Weights in Sparsity-Ranked LASSO

The `sparseR` package (Peterson & Cavanaugh, 2022) implements cumulative penalties that increase the penalty for higher-order terms based on the cumulative count of all terms up to

that order. The implementation treats polynomial and interaction penalties as two separate dimensions.

For polynomials, the penalty weight for degree r is

$$w_r^{\text{poly}} = \frac{\left(\sum_{\ell=1}^{r} m_\ell^{\text{poly}}\right)^\gamma}{\left(m_1^{\text{poly}}\right)^\gamma},$$

where m_r^{poly} is the number of terms of polynomial degree r. The normalization ensures that main effects (degree 1) receive a baseline penalty of 1, while higher-degree terms receive progressively larger penalties.

For interactions, the penalty weight for interaction order s (where $s = 0$ denotes main effects and $s = 1$ denotes pairwise interactions) is

$$w_s^{\text{int}} = \left(\sum_{\ell=0}^{s} m_\ell^{\text{int}}\right)^\gamma,$$

where m_s^{int} is the number of terms of interaction order s. Unlike polynomial penalties, interaction penalties are not normalized.

The final penalty for each predictor is the product $w_j = w_r^{\text{poly}} \times w_s^{\text{int}}$, combining both polynomial order and interaction order effects. With $\gamma = 0.5$, this yields square-root scaling: main effects receive lower penalties than quadratic terms, which in turn receive lower penalties than interactions or cubic terms.

A.4 Final Variable Set

Table A.2 lists all variables in the winning baseline $\lambda_{\min}$ FH model, grouped by type: main effects, polynomial (squared) terms, and mode $\times$ covariate interactions.

Table A.2: Final variable set from the winning baseline $\lambda_{\min}$ model (122 variables). Variables are grouped by type: main effects, polynomial (squared) terms, and mode $\times$ covariate interactions.

Variable	Type	Source	Description
Main effects (37 variables)			
Car (MIV)	Main	MiD	Car (MIV) contrast, mode_C3, OEPV = -1
Cycling (Rad)	Main	MiD	Cycling (Rad) contrast, mode_C2, OEPV = -1
Walking (Fuss)	Main	MiD	Walking (Fuss) contrast, mode_C1, OEPV = -1

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
anteil erwerbstaetige verarbeitendes gewerbe an industrie	Main	INKAR	Share employed in manufacturing vs. industry
anteil minijobs frauen	Main	INKAR	Share of mini-jobs (women)
arbeitsmarktpolitische hilfen der ba fuer arbeit kurzfristig	Main	INKAR	Short-term labor market policy support (Federal Employment Agency)
asylbewerber	Main	INKAR	Asylum seekers
beschaeftigte in kreativbranchen	Main	INKAR	Employees in creative industries
beschaeftigte mit anforderungsniveau helfer	Main	INKAR	Employees with helper skill level
ehescheidungen	Main	INKAR	Divorces
einwohner 85 jahre und aelter	Main	INKAR	Inhabitants aged 85+
einzelhandelsrelevante kaufkraft	Main	INKAR	Retail-relevant purchasing power
elterngeldbezieher vaeter im ersten lebensjahr	Main	INKAR	Parental allowance recipients (fathers) in first year
erreichbarkeit von ic ec ice bahnhoefen	Main	INKAR	IC/EC/ICE train station accessibility
erreichbarkeit von oberzentren	Main	INKAR	Accessibility of major centers
krankenhaeuser der umfassenden notfallversorgung	Main	INKAR	Hospitals with comprehensive emergency care
kurzarbeiter	Main	INKAR	Short-time workers
landgerichte	Main	INKAR	Regional courts
leistungen fuer wohngeld	Main	INKAR	Housing benefit payments
oberlandesgerichte	Main	INKAR	Higher regional courts
offene stellen mit anforderungsniveau helfer	Main	INKAR	Job vacancies requiring helper skill level
pkw gas	Main	INKAR	Gas-powered cars
pkw hybrid insgesamt	Main	INKAR	Hybrid cars (total)
schlafgelegenheiten in beherbergungsbetrieben	Main	INKAR	Beds in accommodation establishments
sdg 11 pkw dichte	Main	INKAR	(SDG 11) Car density
sdg 5 frauenanteil im stadtrat bzw kreistag	Main	INKAR	(SDG 5) Share of women in council
strassenverkehrsunfaelle	Main	INKAR	Road traffic accidents
teilzeitbeschaeftigte	Main	INKAR	Part-time employees

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
unternehmensinsolvenzen	Main	INKAR	Business insolvencies
verbraucherinsolvenzverfahren	Main	INKAR	Consumer insolvency proceedings
wasserflaeche	Main	INKAR	Water area
weibliche berufseinstiegswanderer	Main	INKAR	Female career entry migrants
wohngeldhaushalte mietzuschuss	Main	INKAR	Housing benefit households (rent subsidy)
bodenflaeche fliessgewaesser	Main	Destatis	Land area flowing water
bodenflaeche gewaesser	Main	Destatis	Land area Gewaesser
bodenflaeche wald	Main	Destatis	Land area forest
wohnflaeche in wohngebaeuden 1 wohnung	Main	Destatis	Wohnflaeche In Wohngebaeuden 1 dwelling
Poly effects (22 variables)			
anteil minijobs frauen²	Poly	INKAR	Share of mini-jobs (women)
asylbewerber²	Poly	INKAR	Asylum seekers
beschaeftigte im handwerk²	Poly	INKAR	Employees in skilled trades
beschaeftigte mit anforderungsniveau helfer²	Poly	INKAR	Employees with helper skill level
bibliotheken²	Poly	INKAR	Libraries
ehescheidungen²	Poly	INKAR	Divorces
einzelhandelsrelevante kaufkraft²	Poly	INKAR	Retail-relevant purchasing power
elterngeldbezieher vaeter im ersten lebensjahr²	Poly	INKAR	Parental allowance recipients (fathers) in first year
gender index²	Poly	INKAR	Gender index
krankenhaeuser der umfassenden notfallversorgung²	Poly	INKAR	Hospitals with comprehensive emergency care
krankenhaeuser mit nofallversorgung kinder²	Poly	INKAR	Hospitals with pediatric emergency care
ladepunkte je 100 elektrofahrzeuge bev²	Poly	INKAR	Charging points per 100 electric vehicles (BEV)
offene stellen mit anforderungsniveau helfer²	Poly	INKAR	Job vacancies requiring helper skill level
polizeidienststellen teilweise besetzt²	Poly	INKAR	Police stations – partially staffed
sdg 11 pkw dichte²	Poly	INKAR	(SDG 11) Car density
sdg 5 frauenanteil im stadtrat bzw kreistag²	Poly	INKAR	(SDG 5) Share of women in council

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
stimmenanteile spd²	Poly	INKAR	Vote share SPD
unternehmensinsolvenzen²	Poly	INKAR	Business insolvencies
verbraucherinsolvenzverfahren²	Poly	INKAR	Consumer insolvency proceedings
wasserflaeche²	Poly	INKAR	Water area
weibliche berufseinstiegswanderer²	Poly	INKAR	Female career entry migrants
wohnflaeche²	Poly	INKAR	Living space
Inter effects (63 variables)			
mode C1 × census bodenflaeche fliessgewaesser	Inter	MiD × Destatis	Walking (Fuss) × bodenflaeche fliessgewaesser
mode C1 × census wohnflaeche in wohngebaeuden 1 wohnung	Inter	MiD × Destatis	Walking (Fuss) × wohnflaeche in wohngebaeuden 1 wohnung
mode C1 × inkar anteil erwerbstaetige verarbeitendes gewerbe an industrie	Inter	MiD × INKAR	Walking (Fuss) × anteil erwerbstaetige verarbeitendes gewerbe an industrie
mode C1 × inkar arbeitsmarktpolitische hilfen der ba fuer arbeit kurzfristig	Inter	MiD × INKAR	Walking (Fuss) × arbeitsmarktpolitische hilfen der ba fuer arbeit kurzfristig
mode C1 × inkar asylbewerber	Inter	MiD × INKAR	Walking (Fuss) × asylbewerber
mode C1 × inkar beschaeftigte im handwerk	Inter	MiD × INKAR	Walking (Fuss) × beschaeftigte im handwerk
mode C1 × inkar beschaeftigte mit anforderungsniveau helfer	Inter	MiD × INKAR	Walking (Fuss) × beschaeftigte mit anforderungsniveau helfer
mode C1 × inkar bibliotheken	Inter	MiD × INKAR	Walking (Fuss) × bibliotheken
mode C1 × inkar einwohner 85 jahre und aelter	Inter	MiD × INKAR	Walking (Fuss) × einwohner 85 jahre und aelter
mode C1 × inkar erreichbarkeit von ic ec ice bahnhoefen	Inter	MiD × INKAR	Walking (Fuss) × erreichbarkeit von ic ec ice bahnhoefen
mode C1 × inkar hochschulfoerderung kurzfristig	Inter	MiD × INKAR	Walking (Fuss) × hochschulfoerderung kurzfristig
mode C1 × inkar kurzarbeiter	Inter	MiD × INKAR	Walking (Fuss) × kurzarbeiter

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
mode C1 × inkar ladepunkte je 100 elektrofahrzeuge bev	Inter	MiD × INKAR	Walking (Fuss) × ladepunkte je 100 elektrofahrzeuge bev
mode C1 × inkar oberlandesgerichte	Inter	MiD × INKAR	Walking (Fuss) × oberlandesgerichte
mode C1 × inkar pkw gas	Inter	MiD × INKAR	Walking (Fuss) × pkw gas
mode C1 × inkar strassenverkehrsunfaelle	Inter	MiD × INKAR	Walking (Fuss) × strassenverkehrsunfaelle
mode C1 × inkar weibliche berufseinstiegswanderer	Inter	MiD × INKAR	Walking (Fuss) × weibliche berufseinstiegswanderer
mode C1 × inkar weibliche schulabgaenger mit hauptschulabschluss	Inter	MiD × INKAR	Walking (Fuss) × weibliche schulabgaenger mit hauptschulabschluss
mode C2 × census bodenflaeche fliessgewaesser	Inter	MiD × Destatis	Cycling (Rad) × bodenflaeche fliessgewaesser
mode C2 × census bodenflaeche wald	Inter	MiD × Destatis	Cycling (Rad) × bodenflaeche wald
mode C2 × inkar anteil erwerbstaetige verarbeitendes gewerbe an industrie	Inter	MiD × INKAR	Cycling (Rad) × anteil erwerbstaetige verarbeitendes gewerbe an industrie
mode C2 × inkar anteil minijobs frauen	Inter	MiD × INKAR	Cycling (Rad) × anteil minijobs frauen
mode C2 × inkar bibliotheken	Inter	MiD × INKAR	Cycling (Rad) × bibliotheken
mode C2 × inkar einwohner 85 jahre und aelter	Inter	MiD × INKAR	Cycling (Rad) × einwohner 85 jahre und aelter
mode C2 × inkar einzelhandelsrelevante kaufkraft	Inter	MiD × INKAR	Cycling (Rad) × einzelhandelsrelevante kaufkraft
mode C2 × inkar elterngeldbezieher vaeter im ersten lebensjahr	Inter	MiD × INKAR	Cycling (Rad) × elterngeldbezieher vaeter im ersten lebensjahr
mode C2 × inkar erreichbarkeit von ic ec ice bahnhoefen	Inter	MiD × INKAR	Cycling (Rad) × erreichbarkeit von ic ec ice bahnhoefen
mode C2 × inkar erreichbarkeit von oberzentren	Inter	MiD × INKAR	Cycling (Rad) × erreichbarkeit von oberzentren
mode C2 × inkar kurzarbeiter	Inter	MiD × INKAR	Cycling (Rad) × kurzarbeiter

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
mode C2 × inkar pkw gas	Inter	MiD × INKAR	Cycling (Rad) × pkw gas
mode C2 × inkar polizeidienststellen teilweise besetzt	Inter	MiD × INKAR	Cycling (Rad) × polizeidienststellen teilweise besetzt
mode C2 × inkar schlafgelegenheiten in beherbergungsbetrieben	Inter	MiD × INKAR	Cycling (Rad) × schlafgelegenheiten in beherbergungsbetrieben
mode C2 × inkar stimmenanteile spd	Inter	MiD × INKAR	Cycling (Rad) × stimmenanteile spd
mode C2 × inkar strassenverkehrsunfaelle	Inter	MiD × INKAR	Cycling (Rad) × strassenverkehrsunfaelle
mode C2 × inkar wasserflaeche	Inter	MiD × INKAR	Cycling (Rad) × wasserflaeche
mode C2 × inkar weibliche berufseinstiegswanderer	Inter	MiD × INKAR	Cycling (Rad) × weibliche berufseinstiegswanderer
mode C2 × inkar wohnflaeche	Inter	MiD × INKAR	Cycling (Rad) × wohnflaeche
mode C3 × census wohnflaeche in wohngebaeuden 1 wohnung	Inter	MiD × Destatis	Car (MIV) × wohnflaeche in wohngebaeuden 1 wohnung
mode C3 × inkar anteil erwerbstaetige verarbeitendes gewerbe an industrie	Inter	MiD × INKAR	Car (MIV) × anteil erwerbstaetige verarbeitendes gewerbe an industrie
mode C3 × inkar beschaeftigte in kreativbranchen	Inter	MiD × INKAR	Car (MIV) × beschaeftigte in kreativbranchen
mode C3 × inkar beschaeftigte mit anforderungsniveau helfer	Inter	MiD × INKAR	Car (MIV) × beschaeftigte mit anforderungsniveau helfer
mode C3 × inkar ehescheidungen	Inter	MiD × INKAR	Car (MIV) × ehescheidungen
mode C3 × inkar einzelhandelsrelevante kaufkraft	Inter	MiD × INKAR	Car (MIV) × einzelhandelsrelevante kaufkraft
mode C3 × inkar erreichbarkeit von ic ec ice bahnhoefen	Inter	MiD × INKAR	Car (MIV) × erreichbarkeit von ic ec ice bahnhoefen
mode C3 × inkar erreichbarkeit von oberzentren	Inter	MiD × INKAR	Car (MIV) × erreichbarkeit von oberzentren

Continued on next page

Table A.2 – continued from previous page

Variable	Type	Source	Description
mode C3 × inkar gender index	Inter	MiD × INKAR	Car (MIV) × gender index
mode C3 × inkar hochschulfoerderung kurzfristig	Inter	MiD × INKAR	Car (MIV) × hochschulfoerderung kurzfristig
mode C3 × inkar krankenhaeuser der erweiterten notfallversorgung	Inter	MiD × INKAR	Car (MIV) × krankenhaeuser der erweiterten notfallversorgung
mode C3 × inkar kurzarbeiter	Inter	MiD × INKAR	Car (MIV) × kurzarbeiter
mode C3 × inkar ladepunkte je 100 elektrofahrzeuge bev	Inter	MiD × INKAR	Car (MIV) × ladepunkte je 100 elektrofahrzeuge bev
mode C3 × inkar landgerichte	Inter	MiD × INKAR	Car (MIV) × landgerichte
mode C3 × inkar leistungen fuer wohngeld	Inter	MiD × INKAR	Car (MIV) × leistungen fuer wohngeld
mode C3 × inkar oberlandesgerichte	Inter	MiD × INKAR	Car (MIV) × oberlandesgerichte
mode C3 × inkar offene stellen mit anforderungsniveau helfer	Inter	MiD × INKAR	Car (MIV) × offene stellen mit anforderungsniveau helfer
mode C3 × inkar pkw gas	Inter	MiD × INKAR	Car (MIV) × pkw gas
mode C3 × inkar pkw hybrid insgesamt	Inter	MiD × INKAR	Car (MIV) × pkw hybrid insgesamt
mode C3 × inkar sdg 11 pkw dichte	Inter	MiD × INKAR	Car (MIV) × sdg 11 pkw dichte
mode C3 × inkar sdg 5 frauenanteil im stadtrat bzw kreistag	Inter	MiD × INKAR	Car (MIV) × sdg 5 frauenanteil im stadtrat bzw kreistag
mode C3 × inkar strassenverkehrsunfaelle	Inter	MiD × INKAR	Car (MIV) × strassenverkehrsunfaelle
mode C3 × inkar teilzeitbeschaeftigte	Inter	MiD × INKAR	Car (MIV) × teilzeitbeschaeftigte
mode C3 × inkar verbraucherinsolvenzverfahren	Inter	MiD × INKAR	Car (MIV) × verbraucherinsolvenzverfahren
mode C3 × inkar weibliche berufseinstiegswanderer	Inter	MiD × INKAR	Car (MIV) × weibliche berufseinstiegswanderer

Continued on next page

Table A.2 – continued from previous page

Variable	**Type**	**Source**	**Description**
mode C3 × inkar wohngeld-haushalte mietzuschuss	Inter	MiD × INKAR	Car (MIV) × wohngeldhaushalte mietzuschuss

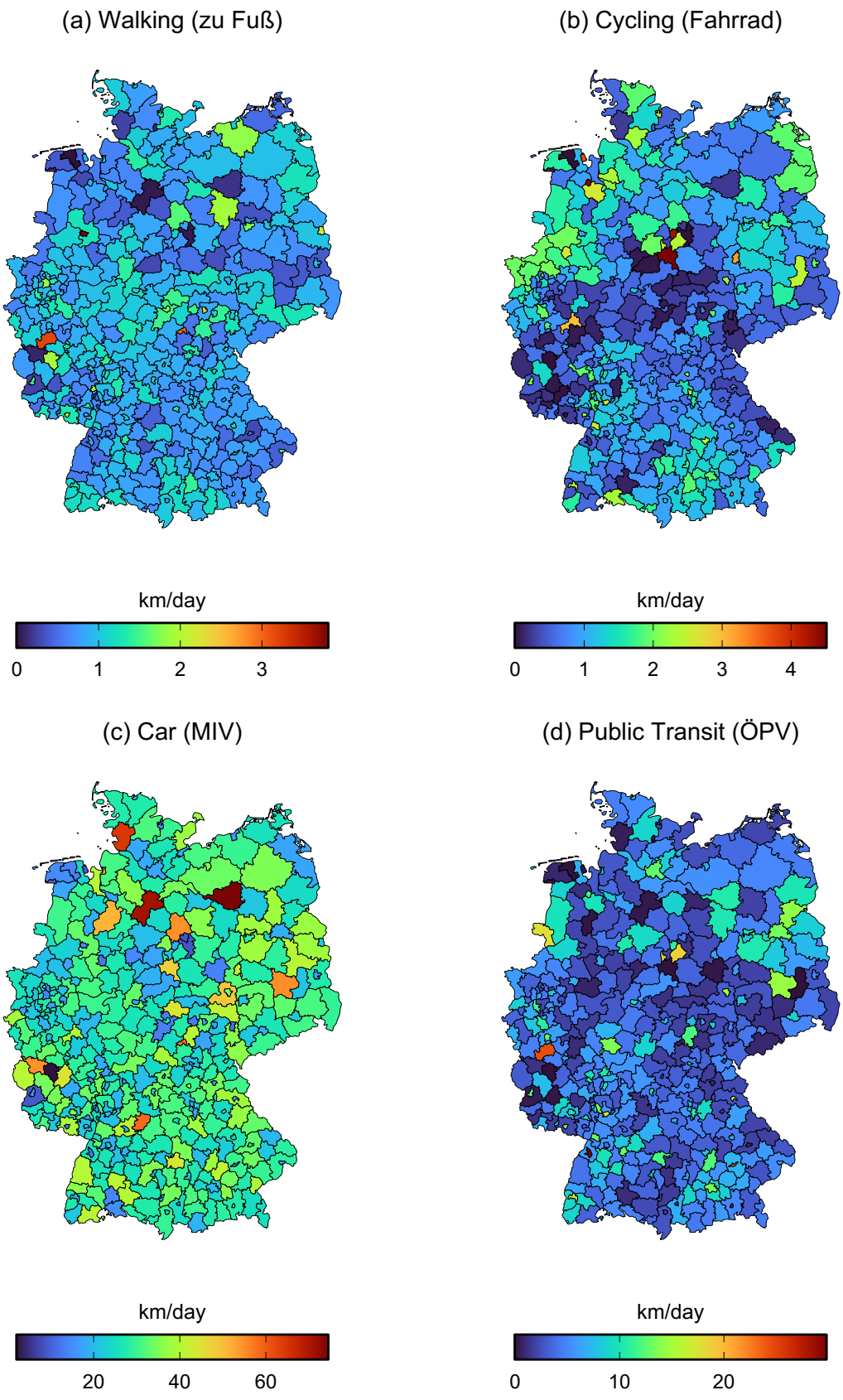


Figure A.1: Spatial distributions of HJ MPK direct estimates by transport mode across German districts: (a) walking (*zu Fuß*), (b) cycling (*Fahrrad*), (c) car (*MIV*), and (d) public transit (*ÖPV*).

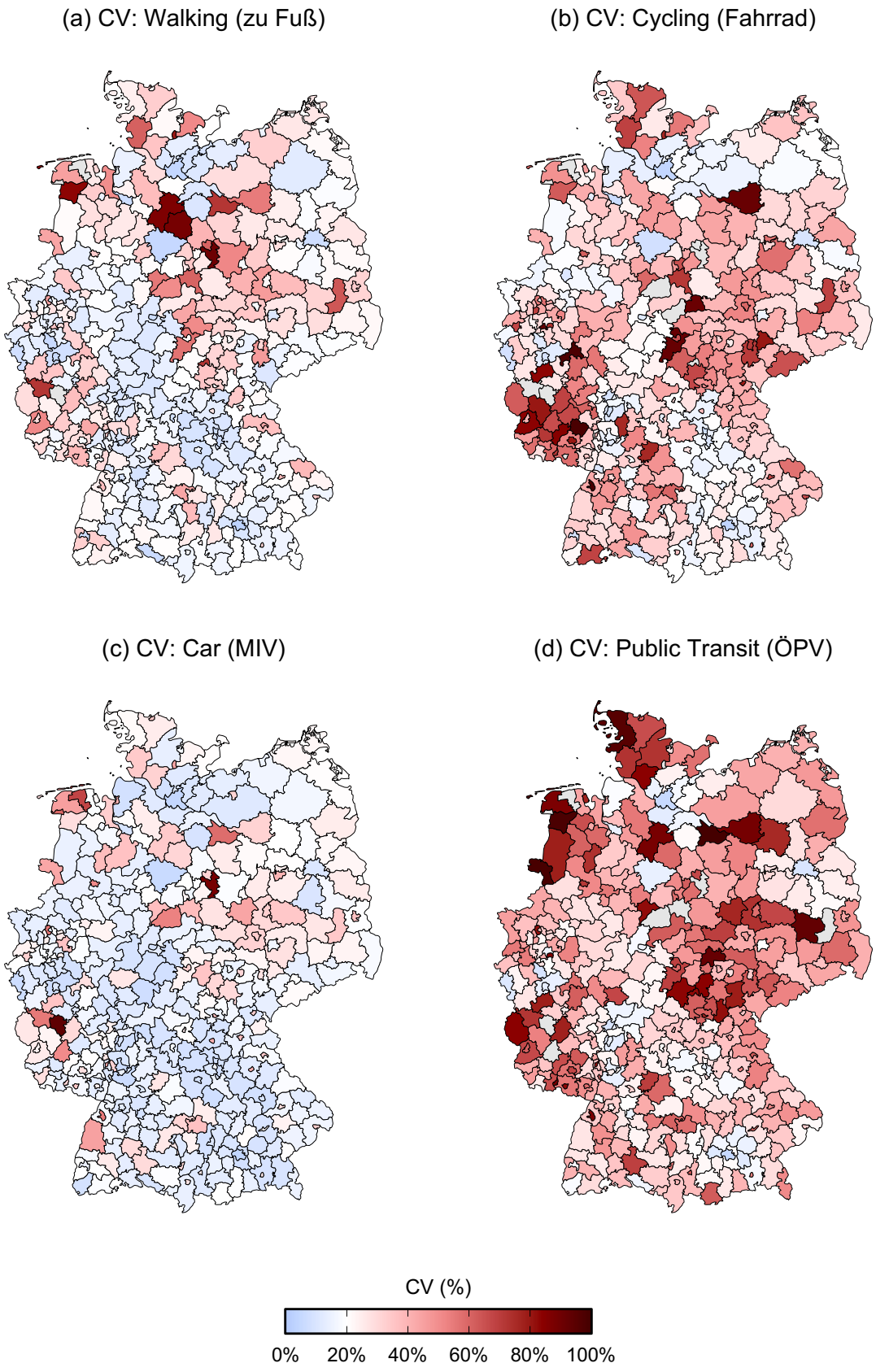


Figure A.2: Spatial distributions of CV for HJ direct estimates by transport mode across German districts: (a) walking (*zu Fuß*), (b) cycling (*Fahrrad*), (c) car (*MIV*), and (d) public transit (*ÖPV*). The 20% threshold separates reliable (blue/green) from unreliable (yellow/red) estimates.

Chapter 2

Density Estimation from Aggregated Data with Integrated Auxiliary Information: Estimating Population Densities with Geospatial Data

Abstract

Density estimation for geospatial data ideally relies on precise geocoordinates, typically defined by longitude and latitude. However, such detailed information is often unavailable due to confidentiality constraints. As a result, analysts frequently work with spatially aggregated data, commonly visualized through choropleth maps. Approaches that reverse the aggregation process using measurement error models in the context of kernel density estimation have been proposed in the literature. From a methodological perspective, we extend this line of work by incorporating auxiliary information to improve the precision of density estimates derived from aggregated data. Our approach employs a correlation-based weighting scheme to combine the auxiliary density with the estimate obtained from aggregated data. We evaluate the method through a series of model-based simulation scenarios reflecting varying conditions of auxiliary data quality. From an applied perspective, we demonstrate the utility of our method in two real-world case studies: (1) estimating population densities from the 2022 German Census in Bavaria, using satellite imagery of nighttime light emissions as auxiliary data; and (2) analyzing brown hare hunting bag data in the German state of Lower Saxony. Overall, our results show that integrating auxiliary information into the estimation process leads to more precise density estimates.

Keywords: Choropleth maps, Geo-referenced data, Kernel density estimation, Regional aggregates

2.1 Introduction

Understanding the spatial distribution of phenomena is relevant across various fields such as public health (see e.g., Yang et al. (2022), who explore clusters of gastrointestinal tumors), urban planning (see e.g., Brandão et al. (2018) who estimate the distribution of building renova-

tions in Lisbon) and environmental monitoring, where Lin et al. (2010) provides an example of identifying soil pollution hotspots in the Changhua County of central Taiwan. A widely applied non-parametric method for estimating spatial distributions is kernel density estimation (KDE) (see Silverman (2018) for a general introduction and Rosenblatt (1956) and Parzen (1962) for the foundational work on the method), which traditionally requires precise data points, such as exact 2D geocoordinates comprised of longitude and latitude values for bivariate spatial analysis. Such precise data are not always available to analysts due to reasons such as the data collection method or post hoc spatial aggregation implemented to address confidentiality concerns. For example, vaccination data for common diseases such as tick-borne encephalitis (TBE) and influenza are only published at the level of the German Nomenclature of Territorial Units for Statistics (NUTS)-3 (Rieck et al., 2020). Similarly, the European brown hare (colloquially also referred to as "jackrabbits") hunting bag data used in this paper is also only recorded at the NUTS-3 district level. The Groß–Rendtel-Schmid–Schmon–Tzavidis (GRSST) estimator (Groß et al., 2017), which is detailed in the next section, is an approach designed for data scenarios characterized by these limitations. The GRSST estimator builds upon the core idea of iteratively enriching a flat pilot KDE estimate with aggregated information, described in Groß et al. (2017), and extends it to the bivariate context of spatial data. Subsequent literature includes Groß et al. (2020), which demonstrates how the GRSST can be used to transform area aggregates between non-hierarchical area systems. In Rendtel and Ruhanen (2018), the algorithm is applied to improve the estimation of local childcare needs in Berlin. Later, Rendtel et al. (2021) utilize the concept to enhance the visualization of COVID-19 incidence clusters.

An inherent characteristic of the GRSST estimator, already identified in the original work by Groß et al. (2017), is its dependence on the geographical size of the input data areas. Large areas tend to produce largely homogeneous density plateaus on the insides away from their borders, as there is no input information to support a more accurate estimate within those areas, apart from the aggregated information. This implies that, ceteris paribus, the quality of the estimate decreases with fewer and larger input areas. Depending on the context and the underlying distribution, the coarseness of the input areas might be negligible, but when looking e.g. at the NUTS-3 district level and the distribution of the human population, accepting a lack of information on the distribution inside of the districts intuitively comes with density estimates of decreased quality. Motivated by the spirit of small area estimation (SAE) (see Rao and Molina (2015) for a comprehensive overview), we try to address this shortcoming through the use of auxiliary data, because oftentimes, analysts may find themselves with auxiliary data that carries some informational value regarding the actual density of interest. For instance, population density tends to correlate with satellite imagery of nighttime light emissions (Sutton et al., 2001). Similarly, in the context of ecological data where the distribution of a certain animal species is only recorded at the level of large habitat areas, detailed vegetation maps could provide valuable auxiliary information about potentially preferred microhabitats within

those broader zones.

At this point, the natural question arises on how to exploit such auxiliary information in order to improve the GRSST estimate. To achieve this, we propose an additional step to the GRSST algorithm: forming a convex combination of an auxiliary density, derived from auxiliary data, and the standard GRSST density estimate. The weighting for this combination is based on the correlation calculated between the auxiliary density, aggregated to the spatial level of the GRSST input data, and the GRSST input data itself. We refer to this augmented approach as augmented GRSST (AGRSST). A special case of the AGRSST arises when the auxiliary density is weighted with 1; we shall call this special case AGRSST1. This version essentially relies fully on the auxiliary density and benchmarks it against the GRSST input data. In this paper, we first recall the GRSST algorithm and then introduce our extension, AGRSST. Subsequently, we investigate the performance of the GRSST, AGRSST, and AGRSST1 in a model-based simulation study and later in an evaluation study based on real-world German population data from the 2022 Census. We also apply the AGRSST to brown hare hunting bag data in the German state of Lower Saxony, using remote sensing normalized difference vegetation index (NDVI) data to construct an auxiliary density. Our results indicate that the AGRSST can be a valuable addition to the exclusive use of the GRSST; however, its strong reliance, especially in the AGRSST1 case, necessitates a sound theoretical justification for the auxiliary density. The structure of the paper is as follows: In Section 2.2, we revisit the GRSST and propose the additional AGRSST step. In Section 2.3, we evaluate our method in a simulation study. Section 2.4 presents an evaluation study that applies the AGRSST in the context of population density estimation. In Section 2.5, we apply our method to a real-world scenario, estimating the density of the brown hare hunting bag. Finally, in Section 2.6, we conclude our paper with a discussion and final remarks, pointing towards possible future research directions.

2.2 Methodology

2.2.1 The Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator

The GRSST estimator was first introduced in Groß et al. (2017) to estimate bivariate densities based on rounded geocoordinate data—where precise geocoordinates are rounded to the center points of the rectangles in which they fall—of ethnic minorities and aged people in Berlin. The estimator is based on a measurement error model which allows the derivation of an iterative procedure reminiscent of the stochastic expectation–maximization (SEM) algorithm (Celeux et al., 1996) combined with KDE in each iteration. The following formulation of the GRSST estimator is slightly more general, describing the rounding process as the association of a geocoordinate with an area. It is important to note that this formulation is not novel but rather a slight variation of the framework already laid out in Groß et al. (2020). The GRSST algorithm is implemented in the `R` package `Kernelheaping` (Groß and Fuchs (2022), see Gril et al.

(2025) for a detailed tutorial).

We assume a geographical map as the sample space $\Omega = \{(s^o, s^a) \in \mathbb{R}^2\}$ for the random variable Z of the latent (unobserved) geocoordinates $z_i = (s_i^o, s_i^a)$, where s_i^o and s_i^a are geographical longitude and latitude values and the index $i \in \{1, 2, \ldots, n\}$ refers to the geocoordinates that constitute the total sample size n. We employ $\boldsymbol{s} = (s^o, s^a)$ to denote a geocoordinate location independently of the underlying random variable. The n geocoordinates $z_1, z_2, \ldots, z_n$ are assumed to be independent and identically distributed draws from a continuous probability density function (PDF) $P_Z(\boldsymbol{s})$, i.e.,

$$z_i \overset{iid}{\sim} P_Z(\boldsymbol{s}), \quad i = 1, 2, \ldots, n.$$

Many applications of the GRSST estimator assume that the entire population is sampled, so $n = N$. We also postulate a random variable $X(Z)$, which maps Z to a set of disjoint sets of geocoordinates, so that $X = f(Z) : \Omega \to \{A_1, \ldots, A_d, \ldots, A_D\}$, where $\Omega = \cup_{d=1}^{D} A_d$. Here, A_d denotes the set of all possible geocoordinates that fall inside of the d-th geographical area, and x represents a particular outcome of the random variable X, where this outcome is one of the defined geographical areas A_d. For example, if Ω represents the map of Germany and the D geographical areas are the German NUTS-1 regions, then, since the NUTS-1 level consists of the 16 federal states, $D = 16$. In this case, one specific A_d would be the set containing all geocoordinates within the state of Bavaria. Consequently, the observed areas $x_1, x_2, \ldots, x_n$ each follow a categorical distribution over the D areas, determined by the location of the corresponding latent geocoordinate:

$$x_i \sim P_X(\boldsymbol{s}), \quad i = 1, 2, \ldots, n, \tag{2.1}$$

which can be evaluated at the geocoordinate level using the map $f(Z)$. The conditional probability distribution of the area X, given a specific latent geocoordinate Z located at $\boldsymbol{s}$ and denoted as $P_{X|Z}(\boldsymbol{s})$, follows a Dirac distribution:

$$P_{X|Z}(\boldsymbol{s}) = \begin{cases} 1 & \text{if } \boldsymbol{s} \in A_d, \text{ with } x = A_d \\ 0 & \text{otherwise.} \end{cases}$$

Applying Bayes' theorem, we obtain

$$P_{Z|X}(\boldsymbol{s}) = \frac{P_{X|Z}(\boldsymbol{s})\, P_Z(\boldsymbol{s})}{P_X(\boldsymbol{s})}. \tag{2.2}$$

This is a measurement error model for the density of the exact, unknown geocoordinate z_i,

given the observed area x_i in which it is located. We draw attention to the fact that $P_{X|Z}(\boldsymbol{s}) = \mathbb{1}_{\boldsymbol{s}\in A_d}(\boldsymbol{s})$, which allows us to rewrite Eq. (2.2) as

$$P_{Z|X}(\boldsymbol{s}) \propto P_{X|Z}(\boldsymbol{s})\, P_Z(\boldsymbol{s}) = \mathbb{1}_{\boldsymbol{s}\in A_d}(\boldsymbol{s}) P_Z(\boldsymbol{s}).$$

Here, we can see that conditioning on the known area membership x_i essentially limits the possible draws of a geocoordinate z_i to the area A_i in which it is known to be located.

The core idea of the GRSST estimator (formally outlined in Algorithm 2) is to begin with a pilot estimate $\hat{P}_Z^{(0)}(\boldsymbol{s})$, incorporate the information contained in x_i by sampling according to Eq. (2.2) for all x_i, to update to a new $\hat{P}_Z^{(l)}(\boldsymbol{s})$ (the "S-Step"), and minimize the root mean integrated squared error (RMISE) using KDE (the "M-Step"). This process is repeated until convergence. Note that the superscript (l) denotes the iteration number, which runs from 0 to $M = L + B$, where B is the number of burn-in iterations and L the number of additional iterations. The S-Step refers to the stochastic step of the SEM algorithm (Celeux, 1985; Celeux et al., 1996), which replaces the deterministic E-Step of the standard expectation-maximization (EM) algorithm (Dempster et al., 1977). Following the SEM framework, the first B burn-in iterations are discarded and the final estimate is obtained by averaging the subsequent L density estimates. While the algorithm shares structural similarities with the EM family, it is important to emphasize several key differences, which we discuss below.

Algorithm 2 Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator

1: Start with a pilot estimate of $P_Z(\boldsymbol{s})$, $\hat{P}_Z^{(0)}(\boldsymbol{s})$, by estimating a KDE with a large 2×2 diagonal bandwidth matrix $\mathbf{H}$. In the `Kernelheaping` package, the default pilot bandwidth is $\mathbf{H} = \operatorname{diag}\left(\left(\Delta s^o/\sqrt{D}\right)^2, \left(\Delta s^a/\sqrt{D}\right)^2\right)$, where Δs^o and Δs^a are the ranges of the longitude and latitude coordinates of the study region.

2: Evaluate $\hat{P}_Z^{(l-1)}(\boldsymbol{s})$ from the last iteration on a fine grid $\mathbf{G} \subset \Omega$ of J geocoordinates (grid points) denoted $\boldsymbol{g}_j$, with index $j \in \{1, 2, \ldots, J\}$. The default in `Kernelheaping` is a 200×200 grid, i.e., $J = 40\,000$.

3: S-Step: For all i, sample from $\mathbf{G}$ using the corresponding value of $P_{X|Z}(x_i|\boldsymbol{s})\, \hat{P}_Z^{(l-1)}(\boldsymbol{s}) + c$ at each geocoordinate as a sampling weight, to obtain samples from $\hat{P}_{Z|X}^{(l)}(\boldsymbol{s})$. The realizations of x_i are given by the data. The constant c acts as an additional smoothing parameter and defaults to $1 \cdot 10^{-10}$ in the `Kernelheaping` package.

4: M-Step: Estimate a new bandwidth $\mathbf{H}$ using the samples from step 3, and use the samples and the new $\mathbf{H}$ to compute the new KDE estimate of the current iteration, $\hat{P}_Z^{(l)}(\boldsymbol{s})$.

5: Repeat steps 2–4 for B (burn-in iterations) + L (additional iterations) times.

6: Discard the B burn-in density estimates and obtain the final density estimate of $P_Z(\boldsymbol{s})$ by averaging the remaining L density estimates $\hat{P}_Z(\boldsymbol{s})$ on the evaluation grid $\mathbf{G}$. The final estimate is called $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$.

It is crucial to observe that the full set of geocoordinates drawn in step 3 of Algorithm 2 is—depending on the granularity of **G**—approximately equivalent to draws from $\hat{P}_Z^{(l)}(\boldsymbol{s})$. For each individual draw, x_i is nonstochastic; the scaling by the marginal likelihood $P_X(\boldsymbol{s})$ in the denominator of Eq. (2.2) is accounted for by the relative empirical frequencies of the x_i in the sample. This is especially true in the case where $n = N$.

As already touched upon above, we can observe familiar structures from the EM algorithm in the GRSST algorithm, but there are significant differences, which we will highlight after a brief introduction to the core components of the EM family of algorithms. The EM algorithm assumes a fully specified parametric model $P_{Z,X}(z, x \mid \boldsymbol{\theta})$, with a latent variable Z, an observed variable X, and a vector of parameters $\boldsymbol{\theta}$. Given the parameters $\boldsymbol{\theta}^{(l)}$ of the current iteration l, the E-Step typically constructs a function

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)}\right) = \int_{\Omega_Z} \log\left[P_{X,Z|\boldsymbol{\theta}}(x, z)\right] P_{Z|X,\boldsymbol{\theta}^{(l)}}(z \mid x) dz = \mathbb{E}_{Z|X,\boldsymbol{\theta}^{(l)}}\left[\log\left[P_{X,Z|\boldsymbol{\theta}}(x, z)\right]\right],$$

which is a lower bound of the log-likelihood $\log\left[P_{X|\boldsymbol{\theta}}(x)\right]$. The M-Step maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)})$ with respect to $\boldsymbol{\theta}$ and thereby provides a new estimate $\boldsymbol{\theta}^{(l+1)}$ for the E-Step of the next iteration. The resulting sequence $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(M)}$ monotonically increases towards local maxima or the global maximum of $\log\left[P_{X|\boldsymbol{\theta}}(x)\right]$. Convergence properties are studied in more detail by Wu (1983). Note that $P_{Z|X,\boldsymbol{\theta}^{(l)}}(z \mid x)$ can be interpreted as a weight for each data point, describing the association of each value of the observed value x with each possible value z given the current parameter estimate $\boldsymbol{\theta}^{(l)}$. These weights enable the appropriate consideration of the current estimate of the distribution of the latent variable z in the subsequent M-Step. For some scenarios, the integration over z in the E-Step is prohibitively difficult or impossible. The Monte Carlo EM algorithm, introduced by Wei and Tanner (1990), approximates $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)}\right)$ by

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)}\right) \approx \frac{1}{T} \sum_{t=1}^{T} \log\left[P_{X,Z|\boldsymbol{\theta}}(x, \tilde{z}_t)\right], \tag{2.3}$$

where z_t denotes samples drawn from the current estimate of $P_{Z|X,\boldsymbol{\theta}^{(l)}}(z \mid x)$. So, instead of weighting with the current value of $P_{Z|X,\boldsymbol{\theta}^{(l)}}(z \mid x)$, T many pseudo-samples of z, denoted $\tilde{z}_t$, are drawn for each observed x, and the average joint log-likelihood is computed. A special case of the Monte-Carlo EM (MCEM) algorithm for a finite mixture model arises for $T = 1$ and is called the SEM algorithm (Celeux, 1985). The M-Step is replaced by the stochastic step (S-Step). The randomness introduced in the S-Step causes the sequence of parameter estimates $\boldsymbol{\theta}^{(l)}$ to not converge monotonically towards the closest saddle point, plateau, or local maximum.

Celeux et al. (1996) argue that the sequence $\boldsymbol{\theta}^{(l)}$ is an irreducible, time-homogeneous Markov chain when $P_{Z|X,\boldsymbol{\theta}^{(l)}}(z \mid x)$ is positive for almost every $\boldsymbol{\theta}$ and z. If this sequence can be shown to be ergodic, it will converge to the unique stationary probability distribution ϕ of the Markov chain. A natural candidate for an estimator of $\boldsymbol{\theta}$ is the average of the last L draws from the chain; see e.g., Nielsen (2000).

Coming back to the GRSST algorithm, we observe that, disregarding the logarithm in Eq. (2.3), step 3 in Algorithm 2 is comparable to the S-Step in the SEM algorithm. Using the model's state from the previous iteration, $\hat{P}_Z^{(l-1)}(\boldsymbol{s})$, the algorithm completes the dataset by drawing samples from $P_{Z|X}^{(l)}(\boldsymbol{s}|x_i) \propto P_{X|Z}\,(x_i|\boldsymbol{s})\,\hat{P}_Z^{(l-1)}(\boldsymbol{s})$. The model defined in Eq. (2.1)–Eq. (2.2) does not make explicit parametric assumptions about the distribution of Z. Instead of producing a sequence of parameter estimates $\boldsymbol{\theta}^{(l)}$, the algorithm generates a sequence of KDE estimates, $\hat{P}_Z^{(l)}(\boldsymbol{s})$. Therefore, the draws in the GRSST's S-Step differ from the classic SEM because $\hat{P}_{Z|X}(\boldsymbol{s}|x_i)$ is constructed using last iteration's $\hat{P}_Z(\boldsymbol{s})$, unlike last iteration's $P_{Z|X,\boldsymbol{\theta}}(\boldsymbol{s})$ in the classic SEM. This general absence of parametric assumptions also affects the GRSST's M-Step. Instead of maximizing the (average, approximate) joint likelihood as in the EM, MCEM, or SEM algorithms, the GRSST algorithm minimizes the asymptotic mean integrated squared error of the kernel density estimator. The replacement of likelihood maximization with a surrogate function is a concept already applied in several generalizations of the EM algorithm (see e.g., Elashoff and Ryan (2004)). McLachlan and Krishnan (2007, ch. 7.7) provide an overview of the MM algorithm, which stands for Majorization-Minimization or Minorization-Maximization algorithm, a family of algorithms more general than the EM. The MM algorithm also allows for minimization in its second step (M-Step), and depending on the underlying model, a host of different surrogate functions can be constructed using relationships such as the Cauchy-Schwarz or Jensen's inequalities. The core difference between the GRSST and the MM algorithm family is that the surrogate function in GRSST is the RMISE function, which is minimized with respect to the chosen bandwidth in the KDE as opposed to model parameters. A rigorous treatment of convergence is still missing in the literature but is beyond the scope of this paper. This is why we also rely on simulation studies to judge the effectiveness of the AGRSST, just as in the original GRSST paper (Groß & Rendtel, 2016).

2.2.2 The Augmented Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator

The precision of the GRSST estimator is naturally constrained by the size of geographical areas which determine the sets $\{A_1, \ldots, A_d, \ldots, A_D\}$ (Groß et al., 2017). The algorithm does not require any additional information about the distribution inside the sets A_d. The AGRSST estimator is a heuristic which enables the use of an auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ aimed at improving the GRSST estimate $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$. Hereby, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is thought to be similar to $P_Z(\boldsymbol{s})$ due to theoretical considerations. In other words, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ serves as a proxy for the unknown true density $P_Z(\boldsymbol{s})$, and the analyst should select auxiliary variables based on substantive knowledge about

the spatial phenomenon of interest. A classical example, which we will investigate later on, is the distribution of human population $P_Z(\boldsymbol{s})$, where the distribution of nighttime light emissions is a natural candidate as an auxiliary density, because increased emissions of light may correlate with an increased human population. Both densities, $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ and $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, can be thought of as imperfect estimations of the true distribution $P_Z(\boldsymbol{s})$. The density $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ is imperfect because of the coarseness of the input data X, and $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ because of the generally unknown level of similarity between $P_Z(\boldsymbol{s})$ and $\hat{P}_{\text{AUX}}(\boldsymbol{s})$. Algorithm 3 depicts the steps to obtain the AGRSST estimator $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$.

Algorithm 3 Augmented Groß–Rendtel–Schmid–Schmon–Tzavidis Estimator

1: Run the standard GRSST algorithm to obtain $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$.
2: Evaluate $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ and $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ on a fine grid $\mathbf{G}$.
3: Calculate representative mean densities of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ and $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ on the level of A_d to obtain $\mathbf{m}_{\text{AUX}}$ and $\mathbf{m}_{\text{GRSST}}$.
4: Calculate the correlation $\hat{\gamma} = \text{cor}(\mathbf{m}_{\text{AUX}}, \mathbf{m}_{\text{GRSST}})$.
5: If $\hat{\gamma} < 0$, invert $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ on $\mathbf{G}$ by setting it to $\frac{\max\left[\hat{P}_{\text{AUX}}(\boldsymbol{s})\right] - \hat{P}_{\text{AUX}}(\boldsymbol{s})}{\Sigma_i\left[\hat{P}_{\text{AUX}}(\boldsymbol{s})\right]}$.
6: Determine $\hat{P}_Z^{\text{temp}}(\boldsymbol{s}) = \hat{\gamma}\hat{P}_{\text{AUX}}(\boldsymbol{s}) + [1 - \hat{\gamma}]\, \hat{P}_{\text{GRSST}}(\boldsymbol{s})$.
7: Define the geocoordinate counts per area as $c_d = \sum_{i \in n} \mathbb{1}\,(x_i = A_d)$.
8: Benchmark by sampling c_d geocoordinates in each A_d according to $\hat{P}_Z^{\text{temp}}(\boldsymbol{s})$. Run a KDE with these geocoordinates to obtain a final estimate $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$.

The core idea of the AGRSST estimator is to create a weighted combination of the standard GRSST density and an auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, which on average produces a better estimate. This is because if $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is unreliable, the estimate largely depends on the more robust GRSST. However, if $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is similar to the true distribution $P_Z(\boldsymbol{s})$, the combined estimate benefits by giving $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ a higher weight. In this context, $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ serves as a baseline estimate grounded in known (population) data, while $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ has the potential to represent certain aspects of $P_Z(\boldsymbol{s})$ much better than $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$.

However, it is generally not clear how close $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ is to $P_Z(\boldsymbol{s})$, and as a result, it is also not clear how much weight should be put on it. For this problem, we propose a basic heuristic of setting the weight on $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ to be equivalent to $\hat{\gamma} = \text{cor}(\mathbf{m}_{\text{AUX}}, \mathbf{m}_{\text{GRSST}})$, which is the correlation of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ and $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ on the level of the A_d geographical areas. The vectors $\mathbf{m}_{\text{AUX}}$ and $\mathbf{m}_{\text{GRSST}}$ are of length D and contain the average densities of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ and the GRSST input data (x_i). We compute $\mathbf{m}_{\text{AUX}}$ independently of the unit underlying the auxiliary data by taking the average of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ evaluated at the grid points inside each A_d. Since the unit of the GRSST input data is always geocoordinate counts, we can compute $\mathbf{m}_{\text{GRSST}}$ by dividing the geocoordinate counts per area by the actual geographical size of their corresponding areas A_d, so $c_d/\text{size}\,(A_d)$. Using linear correlation has the advantage of offering a measure that is inherently constrained

within the range of $[-1, 1]$, which makes it suitable as a weight of the convex combination $\hat{P}_Z^{\text{temp}}(\boldsymbol{s}) = \hat{\gamma}\hat{P}_{\text{AUX}}(\boldsymbol{s}) + [1 - \hat{\gamma}]\,\hat{P}_{\text{GRSST}}(\boldsymbol{s})$. Depending on the nature of the auxiliary variable behind $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, the intermediate density $\hat{P}_Z^{\text{temp}}(\boldsymbol{s})$ might lose some of the information contained in the GRSST input data. To correct for such distortions, step 8 benchmarks the process so that the final estimate $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$ is derived from a set of geocoordinates that align with the overall proportions specified by the GRSST input data. In some settings, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ might exhibit a negative correlation with $P_Z(\boldsymbol{s})$. For example, if $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ represents the density of trees, we might expect a negative relationship with $P_Z(\boldsymbol{s})$ if it represents human population density. For such cases, we propose a pragmatic inversion of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ as seen in step 6 of Algorithm 3. If $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is deemed to be quite informative of $P_Z(\boldsymbol{s})$, an intuitive approach of obtaining an estimate that combines the information of the auxiliary density with the one from the GRSST input data is to sample c_d many geocoordinates in each A_d weighted by $\hat{P}_{\text{AUX}}(\boldsymbol{s})$. This method is, apart from the missing addition of the smoothing parameter c in Algorithm 2's step 4, almost equivalent to setting $\hat{\gamma} = 1$ in the AGRSST. We decided to abstain from adding the constant c in order to not disturb the information of the auxiliary density. It is a natural alternative for the AGRSST; we refer to this estimator as AGRSST1 and denote its density by $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$, and also investigate it.

2.2.3 Obtaining Auxiliary Densities

In practice, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ will often need to be derived from raster data measuring some physical quantity. Algorithm 4 outlines our recommended method for converting this raw raster data into the $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ density. While it is possible to skip step (3) in Algorithm 4 and directly use $w^{std}(\boldsymbol{g}_j)$ in the AGRSST, our simulations indicate that creating $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ from a smoothed density, rather than the raw standardized data, generally leads to a slightly lower RMISE. If the auxiliary data is also given by a (more granular) choropleth count map, then we recommend using the standard GRSST to obtain $\hat{P}_{\text{AUX}}(\boldsymbol{s})$.

Algorithm 4 Obtaining $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ from raster data

1: Associate each grid point $\boldsymbol{g}_j$ in $\mathbf{G}$ with the raw data value $r(\boldsymbol{g}_j)$ of the corresponding raster cell it falls into.

2: Draw a sample from $\mathbf{G}$ using $w^{std}(\boldsymbol{g}_j) = r(\boldsymbol{g}_j) \cdot \left(\sum_{\mathbf{G}} r(\boldsymbol{g}_j)\right)^{-1}$ as a sample weight. The sample size needs to be sufficiently large and can be a scaled version of the population sample size $f \cdot n$.

3: Estimate a standard KDE on the sampled grid points resulting in $\hat{P}_{\text{AUX}}(\boldsymbol{s})$.

2.3 Simulation

2.3.1 Simulation Setup

In this section, we present the setup of the Monte Carlo simulation study used to assess the performance of the estimators introduced in Section 2.2. We use the map of the German state of Bavaria, which is obtained from the German Federal Agency for Cartography and Geodesy (2024). The simulation is based on $r \in \{1, 2, \ldots, R = 400\}$ iterations. In each iteration, an artificial true Bavarian population density $P_Z(\boldsymbol{s})$ is simulated. $P_Z(\boldsymbol{s})$ is obtained by randomly drawing 80 times from the geocoordinates $\boldsymbol{g}_j \in \{(s_j^o, s_j^a) \in \mathbb{R}^2\}$ on the fine grid $\mathbf{G}$. These geocoordinates are stored in the vector $\boldsymbol{\mu}$, which is set to the location parameter of a Gaussian mixture, such that

$$P_Z^{(r)}(\boldsymbol{s}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{s} \mid \mu_k, \boldsymbol{\Sigma}_k), \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leq \pi_k \leq 1, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & 0 \\ 0 & \sigma_{k2}^2 \end{bmatrix},$$

where the index $k \in \{1, 2, \ldots, K = 80\}$ enumerates each Gaussian component. The diagonal elements of the covariance matrices $\boldsymbol{\Sigma}_k$ are also randomly drawn; π_k is set to $1/80$. For each simulation run, we draw an artificial Bavarian population of $n = 250\,000$ from $P_Z^{(r)}(\boldsymbol{s})$ using $\mathbf{G}$. The geographical aggregation level for the GRSST is the German NUTS-3 district (*Kreise*) level. The evaluation grid $\mathbf{G}$ is set to a 500×500 grid, i.e., $J = 250\,000$, chosen to be as fine as computational resources permit. The input vector $\mathbf{x}$ for the GRSST is obtained by mapping the drawn geocoordinates $\mathbf{z}$ to the corresponding district A_d in which they are located. Figure 2.1 shows the simulated true density $P_Z^{(r)}(\boldsymbol{s})$ and the aggregated sampled geocoordinates on district level for an exemplary simulation run r.

We obtain the auxiliary density $\hat{P}_{\text{AUX}}^{(r)}(\boldsymbol{s})$ by taking the true density as a baseline $P_Z^{(r)}(\boldsymbol{s})$ and adding normally distributed distortions with varying variances. Generally, the GRSST (AGRSST) is a method designed to produce relatively reliable results for input data of low quality, due to aggregation. Depending on the target variable at hand, also $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ might be derived from data that is only available in a spatially aggregated form. In the simulation, we therefore base $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ on the German municipality (*Gemeinden*) level. The German municipalities are nested within the NUTS-3 level. Algorithm 5 depicts the process applied to obtain $\hat{P}_{\text{AUX}}^{(r)}(\boldsymbol{s})$ from $P_Z^{(r)}(\boldsymbol{s})$. Note that Algorithm 5 simply aggregates the true density, averages and standardizes it on the municipality level, draws geocoordinates according to these values, re-aggregates them on the municipality level, and applies a GRSST estimator on these aggregates to obtain a smooth density.

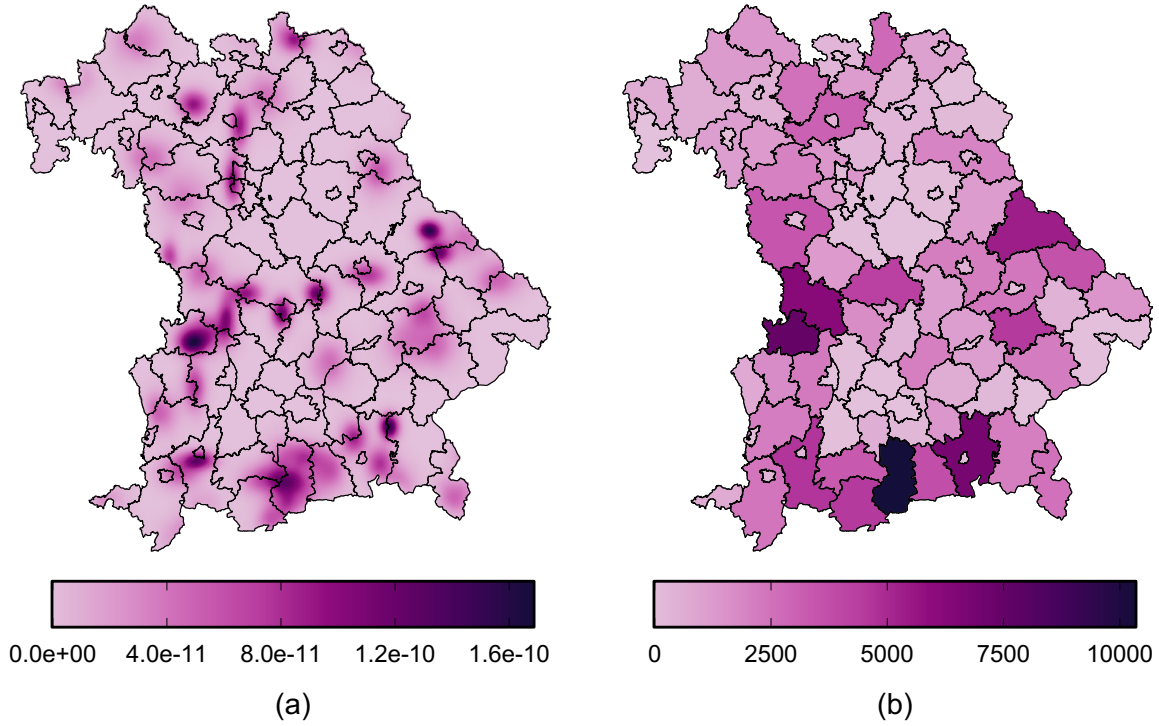


Figure 2.1: (a) Exemplary artificial Bavarian population density $P_Z^{(r)}(\boldsymbol{s})$ of an iteration run r and (b) the corresponding GRSST input data $\mathbf{x}^{(r)}$ resulting from $250\,000$ draws from $P_Z^{(r)}(\boldsymbol{s})$.

Algorithm 5 Obtaining $\hat{P}_{\text{AUX}}^{(r)}(\boldsymbol{s})$ from $P_Z^{(r)}(\boldsymbol{s})$.

1: Denote the sets of geocoordinates that fall inside the p-th municipality with $\{Q_1, \ldots, Q_p, \ldots, Q_P\}$.

2: Evaluate $P_Z^{(r)}(\boldsymbol{s})$ on $\mathbf{G}$ and calculate the mean density μ_{Q_p} of the grid points of $\mathbf{G}$ inside every set Q_p, $\mu_{Q_p} = \sum_{\boldsymbol{g}_j \in Q_p} P_Z^{(r)}(\boldsymbol{g}_j) \left(\sum_{\boldsymbol{g}_j \in Q_p} \mathbb{1}_{\boldsymbol{g}_j \in Q_p}(\boldsymbol{g}_j) \right)^{-1}$.

3: Draw a sample (we chose the population size $n = 250\,000$) from $\mathbf{G}$ using $w_{\mu_{Q_p}}^{std}(\boldsymbol{g}_j \in Q_p) = \mu_{Q_p} \left(\sum_{\boldsymbol{g}_j \in \mathbf{G}} w_{\mu_{Q_p}}(\boldsymbol{g}_j \in Q_p) \right)^{-1}$ as a sample weight, where $w_{\mu_{Q_p}}(\boldsymbol{g}_j \in Q_p) = \mu_{Q_p}$.

4: Aggregate the samples drawn in step 3 to the municipality level $\{Q_1, \ldots, Q_p, \ldots, Q_P\}$.

5: Run the standard GRSST algorithm with the aggregates of the previous step, which returns $\hat{P}_{\text{AUX}}^{(r)}(\boldsymbol{s})$.

To simulate auxiliary densities of different quality, we distort the μ_{Q_p} with normally distributed errors with different variances.

$$\mu_{Q_p,\mathtt{d}} = \mu_{Q_p} + \epsilon_{\mathtt{d}}, \quad \epsilon_{\mathtt{d}} \sim \mathcal{N}\left(0, \left(\mathtt{d} \cdot \mu_{Q_p} \cdot \left(\sum_{p=1}^{P} \mu_{Q_p}\right)^{-1}\right)^2\right),$$

where the standard deviation of ϵ is set to the standardized average density over all the municipalities scaled by $\mathtt{d} \in \{0, 0.5, 1, 2.5, 5, 10, 15, 20\}$. Replacing μ_{Q_p} with $\mu_{Q_p,\mathtt{d}}$ in step 2 of Algorithm 5 leads to auxiliary densities of varying quality, denoted by $\hat{P}^{(r)}_{\text{AUX},\mathtt{d}}(\boldsymbol{s})$.

Following Groß et al. (2017, eq. (5)), the primary measure we use to judge the quality of an estimate $\hat{P}_Z(\boldsymbol{s})$ of $P_Z(\boldsymbol{s})$ is the RMISE, which we approximate and bound with

$$\begin{aligned} RMISE\left(\hat{P}_Z(\boldsymbol{s})\right) &= \sqrt{\mathbb{E}\left[\int \left(P_Z(\boldsymbol{s}) - \hat{P}_Z(\boldsymbol{s})\right)^2 d\boldsymbol{s}\right]} \\ &\approx \sqrt{\frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{J}\left(P_Z^{(r)}(\boldsymbol{g}_j) - \hat{P}_Z^{(r)}(\boldsymbol{g}_j)\right)^2 \delta^2_{\boldsymbol{g}_j}} \\ &\geq \frac{1}{R}\sum_{r=1}^{R}\sqrt{\sum_{j=1}^{J}\left(P_Z^{(r)}(\boldsymbol{g}_j) - \hat{P}_Z^{(r)}(\boldsymbol{g}_j)\right)^2 \delta^2_{\boldsymbol{g}_j}} \\ &\propto \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(P_Z^{(r)}(\boldsymbol{g}_j) - \hat{P}_Z^{(r)}(\boldsymbol{g}_j)\right)^2}, \end{aligned} \tag{2.4}$$

The term δ refers to the side lengths of the square pixels, where the $\boldsymbol{g}_j$ are the centroids and $J = |\mathbf{G}|$ denotes the number of points in the evaluation grid $\mathbf{G}$, as introduced in Algorithm 2. The grid is evenly spaced, with $\delta^2_{\boldsymbol{g}_j}$ being a scalar for all $\boldsymbol{g}_j$. The approximation step in Eq. (2.4) replaces the expectation with the average over the R Monte Carlo replications, indexed by the superscript (r), and the integral with its Riemann sum on the finite grid (for the same construction, see, e.g., Geenens, 2014). The subsequent inequality is Jensen's: moving the square root inside the replication average can only decrease the value, so the average of the per-replication root integrated squared errors is bounded above by the approximated RMISE. The final proportionality drops the constant factor $\delta\sqrt{J}$, turning each summand into the root mean squared error between the true and the estimated density over the grid points. Since all estimators are evaluated on the same grid and the same replications, neither the Jensen gap nor the omitted constant affects any comparison. In the rest of this paper, we refer to the per-replication summand in the last term of Eq. (2.4) as the RMISE; the simulation study reports its distribution across the R replications, whereas in the real-data evaluation of Section 2.4 a single realization is available.

2.3.2 Simulation Results

In this section, we discuss the results obtained from the Monte Carlo simulation study laid out in the previous section. For the burn-in and additional iterations, we chose $B = 30$ and $L = 20$. Figure 2.2 depicts the behavior of $P_Z^{(r)}(\boldsymbol{s})$ at three different locations: one that showed high (medium, low) density in the first ($l = 1$) iteration of the algorithm. It can be seen that after the

pilot iteration ($l = 0$), the densities seem to quickly reach a behavior that suggests that $P_Z^{(r)}(\boldsymbol{s})$ is roaming around the stationary distribution of the Markov chain (if it exists). Additionally, choosing $B = 30$ and $L = 20$ surpasses the recommendations in the supplementary material of Groß et al. (2017). We therefore conclude that the chosen number of 50 iterations is sufficient in our context.

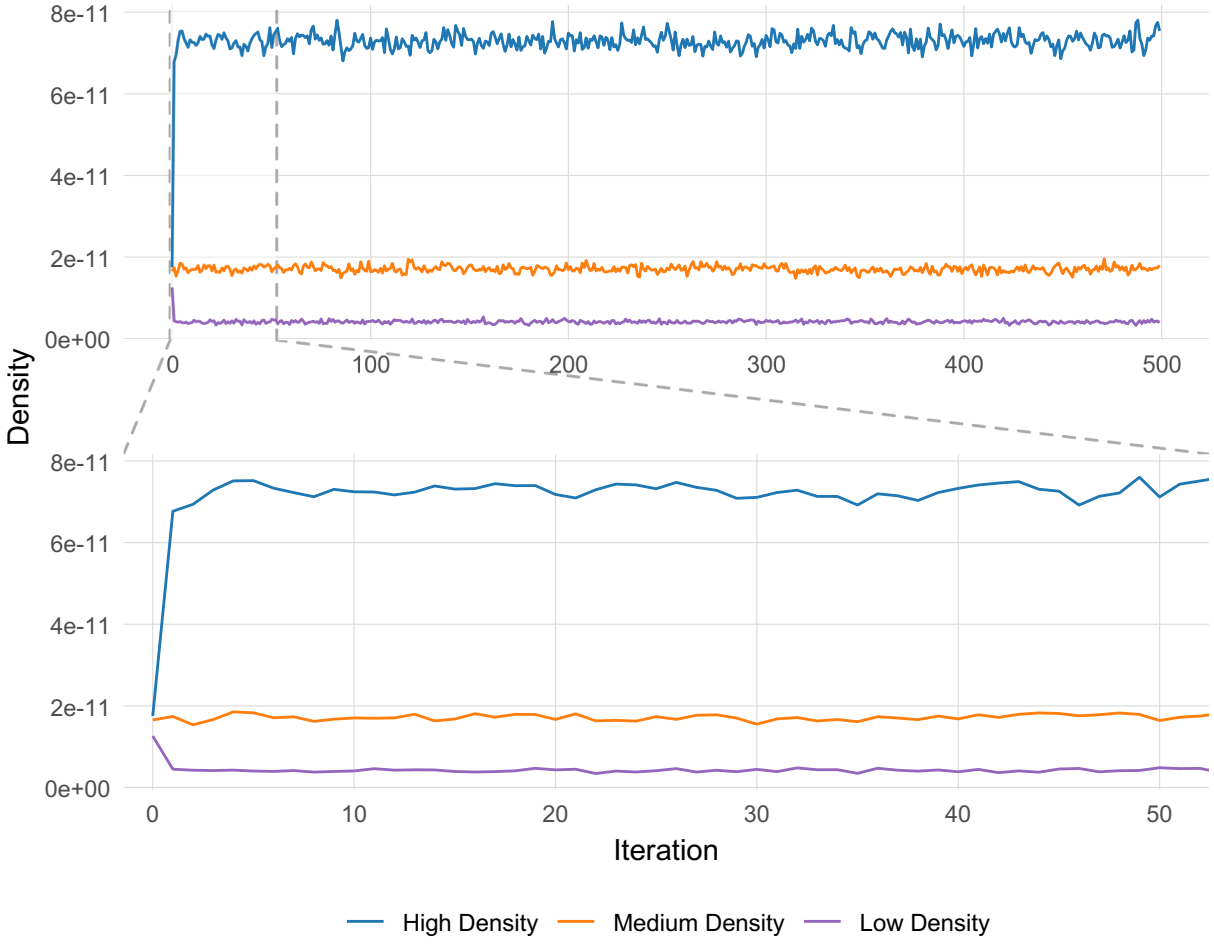


Figure 2.2: The development of $\hat{P}_Z^{(l)}(\boldsymbol{s})$ over $M = 500$ iterations of the GRSST evaluated at 3 different $\boldsymbol{g}_j$. Blue (purple, orange) represents a geocoordinate that displayed relatively high (medium, low) density in the first ($l = 1$) iteration. A zoomed-in view of the first 50 iterations is shown in the lower panel.

The boxplots in Figure 2.3 display the distribution of the RMISE across 400 simulation runs for each distortion level and estimator used: $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$, and $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$. As expected, the RMISE of the GRSST estimator remains almost constant across different distortion levels, as it is completely independent of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$; minor differences can be attributed to sampling uncertainty. All estimates that utilize the auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ perform significantly better than the GRSST for low levels of distortion. This advantage diminishes and even reverses in the case of $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ with increasing levels of distortion. The auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ behaves as anticipated: since it is obtained via a GRSST on the municipality level, its RMISE is lower than that of the GRSST on the district level. At elevated levels of distortion,

$\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is derived from samples of a density that exhibits increasing divergence from the true density $P_Z(\boldsymbol{s})$. Consequently, its RMISE is generally much higher than that of the GRSST estimator. The poor performance of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ at high distortion levels highlights how, in practice, simply choosing a seemingly similar density as an estimate for another should be avoided.

The methods that utilize the GRSST input data $\left(\hat{P}_{\text{AGRSST}}(\boldsymbol{s})\right.$ and $\left.\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})\right)$ mitigate the effects of a very poor auxiliary density. The comparison between $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$ and $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ reveals lower RMISE values for the AGRSST at both the lowest and high distortion levels, whereas $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ outperforms the AGRSST for moderate distortions. For highly informative auxiliary densities where the only source of divergence from the true density is the averaging at the municipality level (distortion = 0), the benchmarking step in $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$ and $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ does not improve the estimate but increases it due to the noise from the additional KDE process. In practice, this effect would only be relevant for auxiliary densities that are very similar to the true density and is therefore of little practical consequence. The AGRSST's advantage over the AGRSST1 at zero distortion likely stems from its ability to assign a small weight to the GRSST estimate, which slightly smooths the municipality-level based $\hat{P}_{\text{AUX}}(\boldsymbol{s})$. In the case of moderate distortions (100 - 500), the additional information about the true density incorporated into $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ outweighs this advantage, and $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ outperforms $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$. For very poor auxiliary densities, AGRSST performs better than $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ due to its ability to assign weight to the GRSST estimate, which is superior and more conservative because of its independence from the (heavily distorted) auxiliary density.

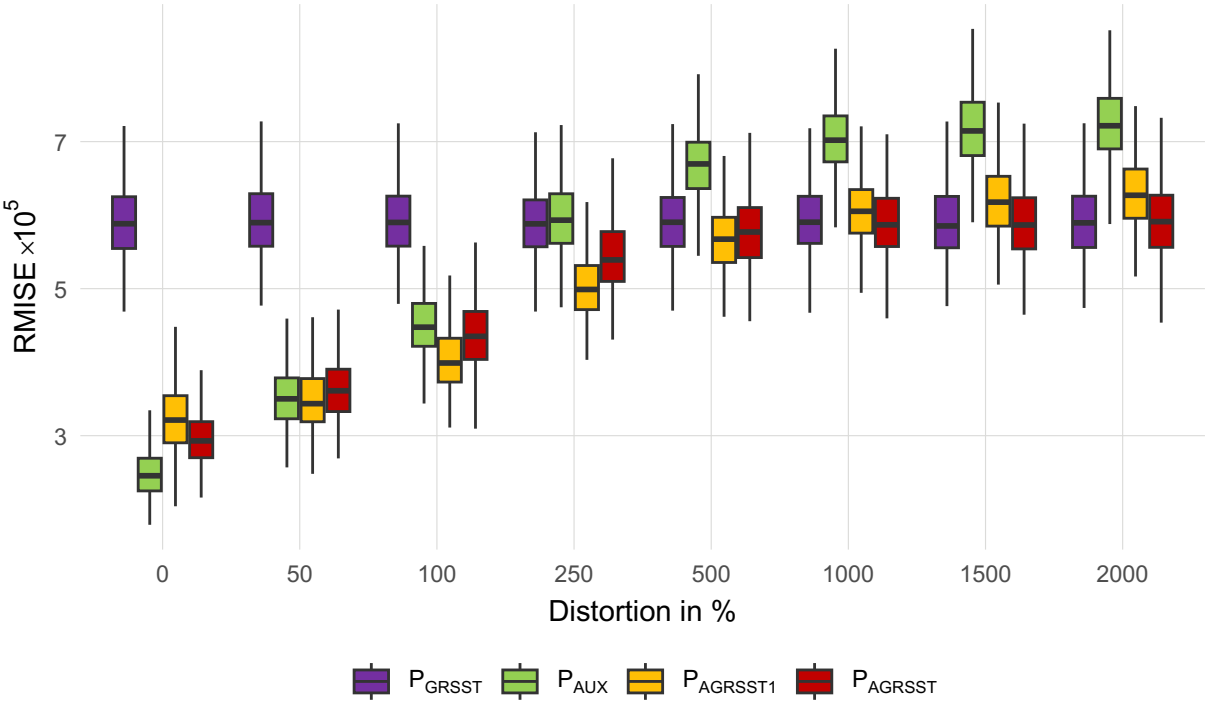


Figure 2.3: Boxplots of the RMISE of $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$ and $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ over increasing levels of distortion of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$.

The above-described behavior is confirmed by Figure 2.4, which shows the dependence of $\hat{\gamma}$ on the distortions. We deem the displayed relationship between $\hat{\gamma}$ and the distortion levels as generally desirable, because it drives the advantages of the AGRSST illustrated in Figure 2.3. Though, when examining the interplay between the true density $P_Z(\boldsymbol{s})$, $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ and $\hat{\gamma}$ the limitations of the AGRSST become apparent. The performance of the AGRSST estimator relies on the fact that $\hat{\gamma}$ properly reflects the similarity between the true and the auxiliary density. Confounding factors that can negatively impact the reflection of this relationship via $\hat{\gamma}$ can be:

- Nonlinearity: $P_Z(\boldsymbol{s})$ and $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ might be related in a nonlinear fashion. Since $\hat{\gamma}$ is essentially a standard empirical correlation, it might not, or only partly, capture such relationships.

- Ecological fallacy (Holt et al., 1996): Since all the information about $P_Z(\boldsymbol{s})$ is limited to the GRSST input data at its respective geographical aggregation level, we calculate $\hat{\gamma}$ at that level. The AGRSST assumes, though, that $\hat{\gamma}$ is a valid similarity measure at the level of the densities.

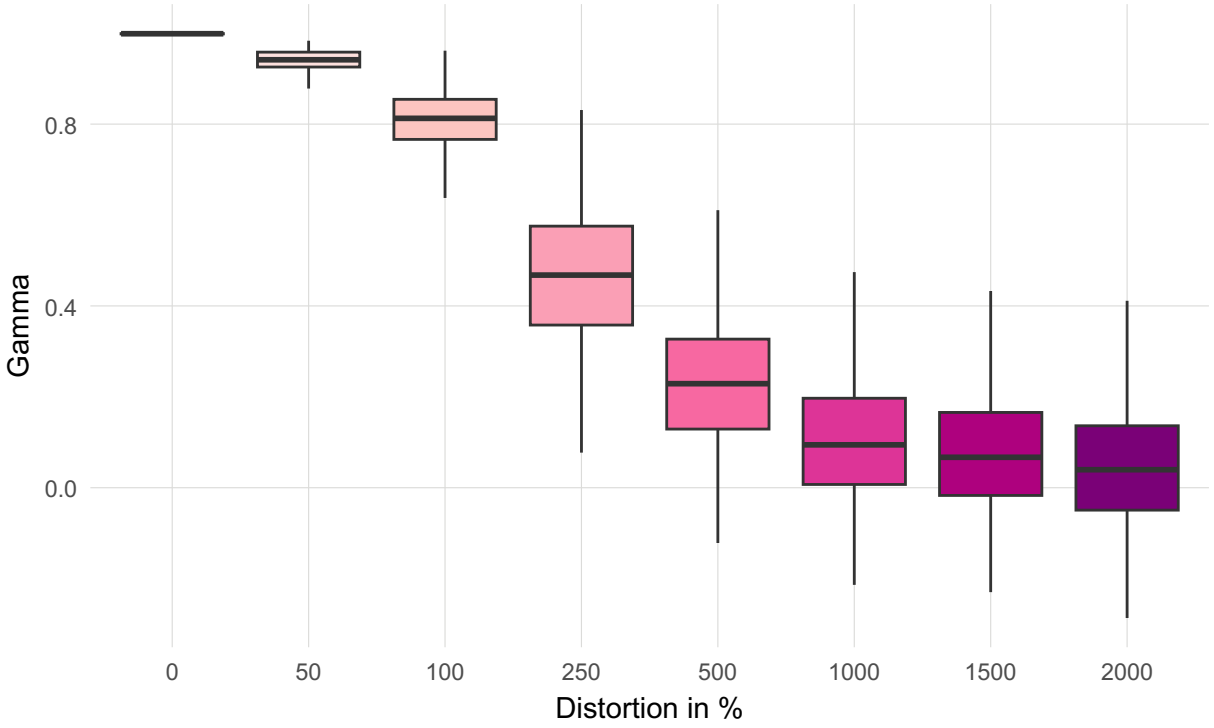


Figure 2.4: Boxplots of the calculated value of $\hat{\gamma}$ over increasing levels of distortion of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$.

The main tool to counteract problems caused by a $\hat{\gamma}$ that does not reflect the quality of $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ is the benchmarking (step 8 in Algorithm 3). It limits the amount of possible divergence of $\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$ from $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$. The worst-case scenario for the AGRSST is a highly divergent auxiliary density from the true density, which is still highly correlated with the true density at the GRSST input level. This scenario is covered by our simulation study; at the highest distortion

levels, $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ puts maximum weight on very poor auxiliary densities ($\hat{\gamma} = 1$). We observe that $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ performs worse than the GRSST, but the benchmarking step still ensures that divergence is limited. The AGRSST is basically equivalent to the GRSST at these high distortion levels. The main insights from our simulation study are that:

- $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$, which benchmarks a plausible auxiliary density with the GRSST input c_d, is an attractive alternative to the GRSST with large potential gains in RMISE but a potentially worse RMISE for poor auxiliary densities.
- In contrast to $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$, the AGRSST provides another layer of security to practitioners by limiting the weight placed on the auxiliary density when low correlation at the GRSST input level is detected. The trade-off for this increased security is a lower RMISE performance in the case of very informative auxiliary densities.

2.4 Evaluation Study on Population Data

2.4.1 Target Variable

The estimation of human population distributions is a classic area of interest in research in order to allow for targeted social- and environmental planning and policy development. Multiple contributions to the topic are, for example, added by the WorldPop research group; see, e.g., Stevens et al. (2015) or Sorichetta et al. (2015). In our first application, we use recently published data from the German Census 2022 that is conducted by the Federal Statistical Office of Germany (Destatis). Population counts are available in 100-meter raster cells (Destatis, 2024a) and also at the district level (Destatis, 2024b). We consider the raster data in relation to the size of Bavaria as granular enough to estimate a density $P_Z(\boldsymbol{s})$, which we assume to be the true population distribution. This density is estimated using a KDE based on geocoordinates located at the centroids of each raster. So, for a raster XYZ in which X-many people live, we place X-many geocoordinates at the centroid of XYZ. This is done for all rasters in Bavaria. The true distribution $P_Z(\boldsymbol{s})$ follows from a KDE on the resulting dataset (Figure 2.5 panel (a)). We obtain $\hat{P}_{\text{GRSST}}(\boldsymbol{s})$ by running the GRSST estimator on the district-level population data using a 700×700 evaluation grid (Figure 2.5 panel (b)). Since highly accurate data on the population distribution is already available, the principal objective of our first application is not to obtain an even more accurate estimate of the population density, but to use and compare our methods in a real-world scenario in which we also have access to a fairly accurate estimate of the ground truth $P_Z(\boldsymbol{s})$.

2.4.2 Auxiliary Variable

In practice, the auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ can be derived from various sources with differing underlying units and aggregation levels. For example, in the case of nighttime light satellite

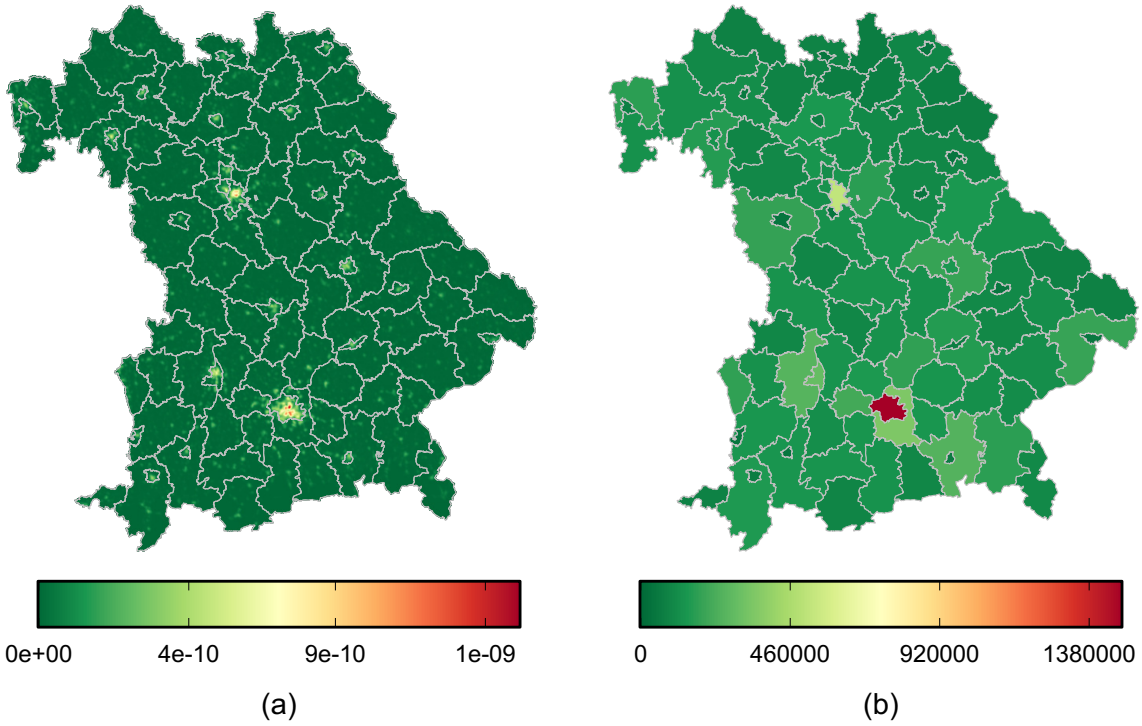


Figure 2.5: (a) The assumed ground truth, a KDE of the true Bavarian population based on the 2022 German Census 100-meter raster data and (b) the Bavarian population on the NUTS-3 district (*Kreise*) level, which serve as the GRSST input data.

raster data (Elvidge et al., 2021) that we use, the resolution of 15 arc-seconds ($\approx$ 300 meters at the latitude of Bavaria's geographical center) is fairly granular with respect to the sizes of the NUTS-3 districts. In this example, the radiance of the light is measured in watt per steradian per square meter ($W \cdot sr^{-1} \cdot m^{-2}$). For our application, we use the 2022 masked median radiance from monthly averages. The process to obtain these values includes multiple steps to adjust for distortions such as biomass burning or aurora and is documented in Elvidge et al. (2021). The general relationship between nighttime lights and population density is well established and, e.g., studied by Sutton (1997), Elvidge et al. (1999), Sutton et al. (2001), Wu et al. (2023).

In the scope of Bavaria, we deem the resolution of nighttime light radiance data granular enough to directly convert it into a density using Algorithm 4. Depending on the target variable at hand, the analyst might not have access to auxiliary data as granular as the nighttime light satellite data. In such cases, we recommend using a GRSST for the conversion, similar to Algorithm 5. As the sample size in Algorithm 4, we chose the unscaled Bavarian population ($n = 13\,038\,724$). Note that the analyst can scale this number, if computational resources allow, to further reduce the sampling uncertainty, although the gains are likely to be relatively small given the already large sample size. The smoothing effect by the KDE can be judged visually by comparing the (transformed) raw data to $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ as in Figure 2.6. If core features of the raw data are smoothed out, we recommend using the raw data or increasing n. The benchmarked auxiliary density $\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$ is analogously obtained from $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ as in the simulation study.

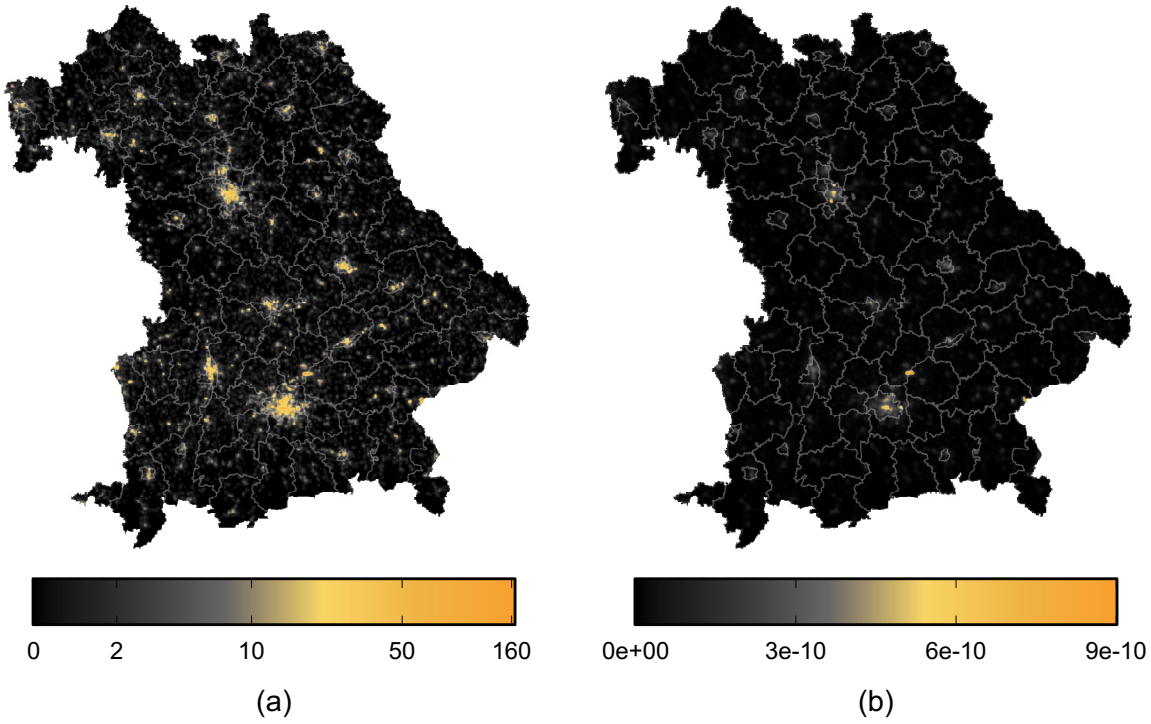


Figure 2.6: (a) The raw nighttime light satellite raster data, displayed on a pseudo log-transformed color scale for better contrast, and (b) the auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ derived by Algorithm 4. The auxiliary density is also pseudo log-transformed, and values above $9 \cdot 10^{-9}$ (located at Munich airport) are cut for improved visibility.

2.4.3 Evaluation Results

The value for γ is calculated as laid out in Section 2.2.2 and is $\hat{\gamma} \approx 0.945$, which means that the light density values are highly correlated with the Bavarian population densities at the district level. The availability of the true density allows us to determine RMISE values, which are depicted in Table 2.1.

Method	**RMISE** $\cdot 10^5$
$\hat{P}_{\text{GRSST}}(\boldsymbol{s})$	8.3252
$\hat{P}_{\text{AGRSST1}}(\boldsymbol{s})$	8.1803
$\hat{P}_{\text{AGRSST}}(\boldsymbol{s})$	**8.1396**
$\hat{P}_{\text{AUX}}(\boldsymbol{s})$	8.8786

Table 2.1: RMISE values of the different population density estimates, calculated based on census population data.

Compared to the simulation study, the improvements of the AGRSST over the GRSST in this real-world example appear rather small, with an RMISE reduction of 2.23%. A potential reason for such small gains could be large deviations between light intensity and actual population in areas such as airports or at the Theresienwiese in Munich (Figure 2.8). Interestingly, the AGRSST1 performs slightly worse than the AGRSST; we attribute this to the additional

smoothing that the mixture provides, especially in said areas with large deviations between light and population density, a dynamic that we also observed in our simulation study at the zero distortion level. The non-benchmarked $\hat{P}_{\text{AUX}}(\boldsymbol{s})$, in turn, has a 6.65% higher RMISE than the GRSST. This structure can be compared to the distortion levels of 250 and 500 in our simulation study, where the AGRSST and AGRSST1 outperform the GRSST, even with an auxiliary density that itself has a higher RMISE than the GRSST. Figure 2.7 depicts the resulting densities; upon visual inspection, the large-scale population density seems to be much better captured by AGRSST and AGRSST1, contrary to the small improvement in RMISE.

Figure 2.8, an enlarged view of Figure 2.7 focusing on Munich, illustrates the dynamics of all three methods. While the GRSST estimate appears largely uniform within Munich's boundaries, AGRSST1 and AGRSST manage to capture some of the true population dynamics, although both incorrectly depict the Theresienwiese area in central Munich – the site of the famous Oktoberfest – as having the highest density, rather than a low population. Given the relatively spiky nature of the auxiliary density, the more conservative nature of AGRSST compared to AGRSST1 is evident in its less spiky density. This example highlights the significant dependence of the AGRSST on the auxiliary density and underscores the importance of selecting these densities based on expert subject knowledge. Despite the generally limited data availability typically associated with GRSST usage, our overall evaluation study demonstrates the potential of AGRSST to enhance GRSST density estimates, both in terms of RMISE and visual representation.

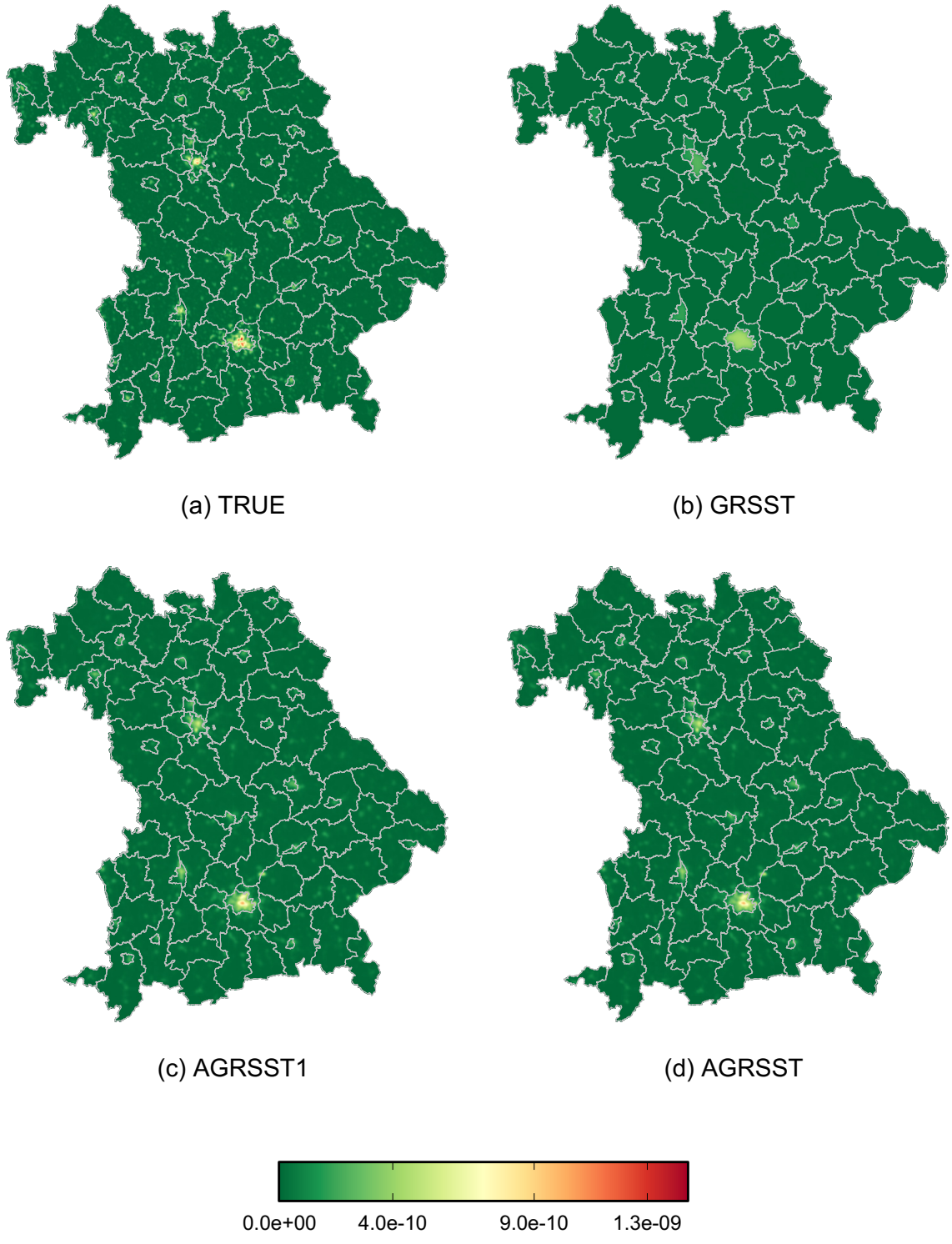


Figure 2.7: (a) The assumed ground truth, a KDE of the true Bavarian population based on the 2022 German Census 100-meter raster data, (b) the GRSST- , (c) AGRSST- and (d) the AGRSST1 estimate.

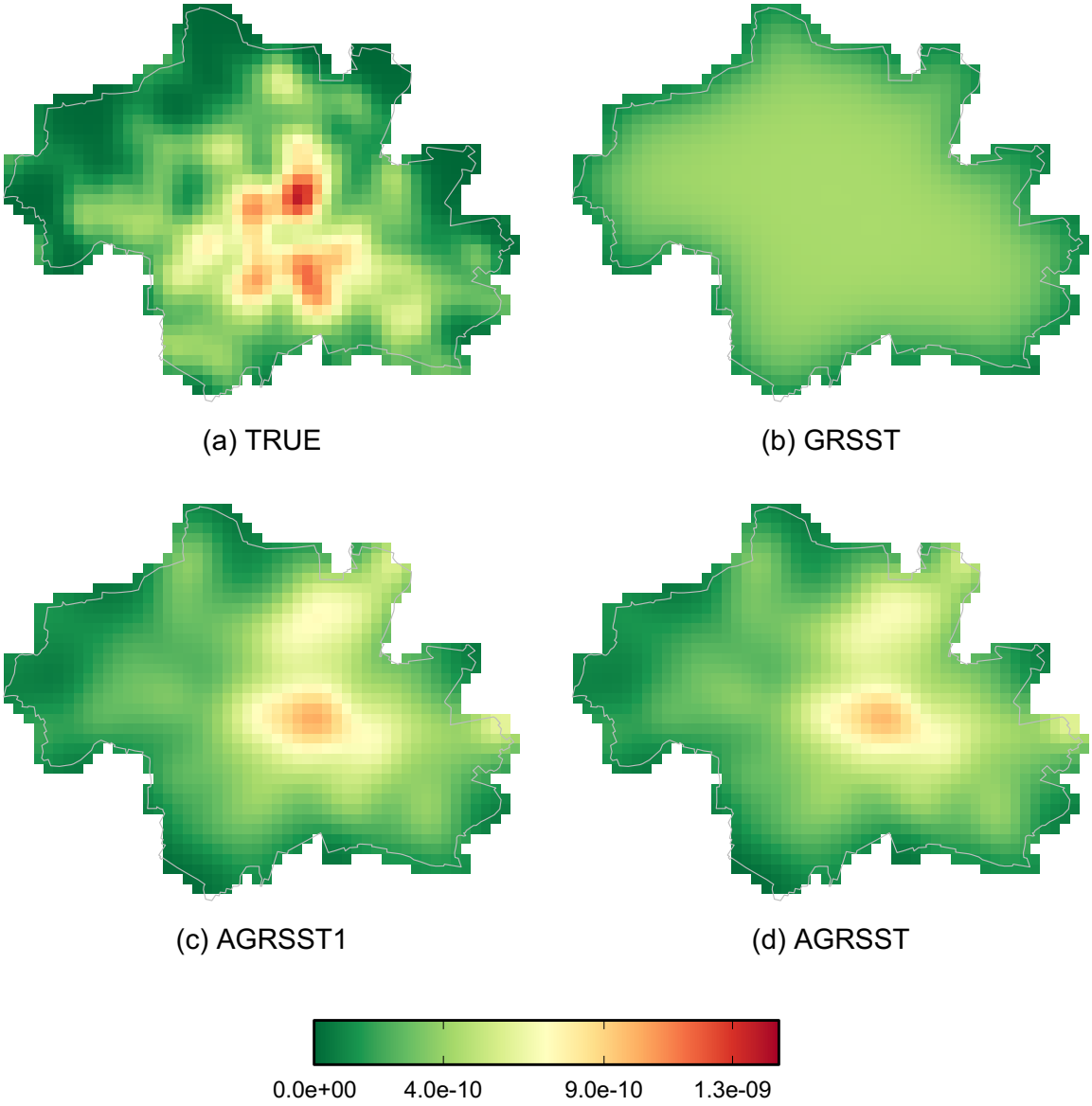


Figure 2.8: (a) The assumed ground truth, a KDE of Munich's true population based on the 2022 German Census 100-meter raster data, (b) the GRSST- , (c) AGRSST- and (d) the AGRSST1 estimate.

2.5 Application: Brown Hare Population Control

2.5.1 Target Variable

The *Lepus europaeus*, often called the brown hare or European hare, is a species of *Lepus* that is native to many European countries but has also been introduced into areas all over the world. Bock (2020) provides a comprehensive overview and literature collection on the general characteristics, distribution, ontogeny, reproduction, and other aspects of the mammal. Since the 1960s, there has been a steep decline in brown hare populations across Europe. The primary causes of this decline are believed to be the loss of landscape diversity and the emergence of diseases (see e.g., Smith et al. (2005)). Since 2015, the population of brown hares in Lower Saxony has shown signs of recovery, which is likely attributable to the dry conditions experienced in recent years, which have resulted in a decline in infection rates. The implementation of local initiatives by hunters, including the creation of wildflower strips and increased predator hunting, is another potential contributing factor to the recovery (Gräber et al., 2024). Sliwinski et al. (2019) studied the influence of different habitat variables on brown hare population density in Lower Saxony, using generalized additive mixed models (GAMMs) fit at the municipality level. We focus on identifying areas of high and low hunting intensity. To achieve this, we estimate a density using a 1000×1000 evaluation grid based on the number of brown hares killed per district in Lower Saxony and Bremen during the hunting season 2023/2024 (01.04.2023 - 31.03.2024) (Dammann-Tamke, 2025). The number of animals killed by hunting and/or traffic-related accidents is also referred to as the hunting bag or hunting bag statistic. While hunting bag statistics have been used as a proxy to infer trends in wildlife population densities, this practice is not always advisable for reasons such as temporary hunting bans (Gräber et al., 2024). As a complement to wildlife population densities, hunting bag statistics are still an important analytical tool used to determine overall variations in wildlife numbers (see e.g., Havet (2002)) and are needed to implement harvest management schemes (Aubry et al., 2020).

2.5.2 Auxiliary Variable

We utilize the NDVI as the underlying data for the purpose of obtaining an auxiliary density for the hunting bag data. The fundamental concept (see Huete et al. (2002)) underpinning vegetation indices, such as the NDVI, is based on the study of the spectral reflectance of leaves. The reflected radiation in the red band from leaves is relatively low due to the absorption of light by photosynthetically active pigments, such as chlorophyll. In contrast, a significant portion of near-infrared radiation (NIR) is reflected, primarily due to the internal structure of plant leaves. This contrast between red and near-infrared reflectance allows for the construction of the NDVI

as a sensitive index used to quantify vegetation, given by

$$NDVI = \frac{\rho_{\text{NIR}} - \rho_{\text{RED}}}{\rho_{\text{NIR}} + \rho_{\text{RED}}},$$

where ρ_{NIR} and ρ_{RED} are the surface bidirectional reflectance factors in the near-infrared and red bands. The raw data is collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) on board the Earth Observing System-Terra platform (Didan & Barreto-Muñoz, 2019). We use the `MOD13A1: 16-day 500m VI` data product (Didan, 2021), which provides 500-meter NDVI raster data in 16-day intervals. The specified interval (01.04.2023 - 31.03.2024) encompasses a total of 23 timestamps for raster layers. The mean of these raster values is calculated, resulting in a single NDVI raster layer for the designated hunting season (2023/2024). The NDVI is standardized between -1 and 1, with the majority of observed values falling between 0 (water or bare soil) and 0.8 (dense vegetation). Panel (a) in Figure 2.9 displays the NDVI raw data over Lower Saxony. There is a rich body of literature (e.g., Sliwinski et al. (2019), Schai-Braun et al. (2012), Smith et al. (2005)) on brown hare habitat preferences, which indicates that these depend on factors such as season, temperature, precipitation, and brown hare density itself. Generally, we can conclude that brown hares prefer heterogeneous agricultural landscapes with a mix of crops, grasslands, and semi-natural habitats, avoiding large monocultures, urban areas, and intensively farmed regions. The NDVI does not directly reflect all of these factors, such as agricultural heterogeneity, but the general assumption of a positive relationship between vegetation, expressed by NDVI, and the brown hare hunting bag seems plausible.

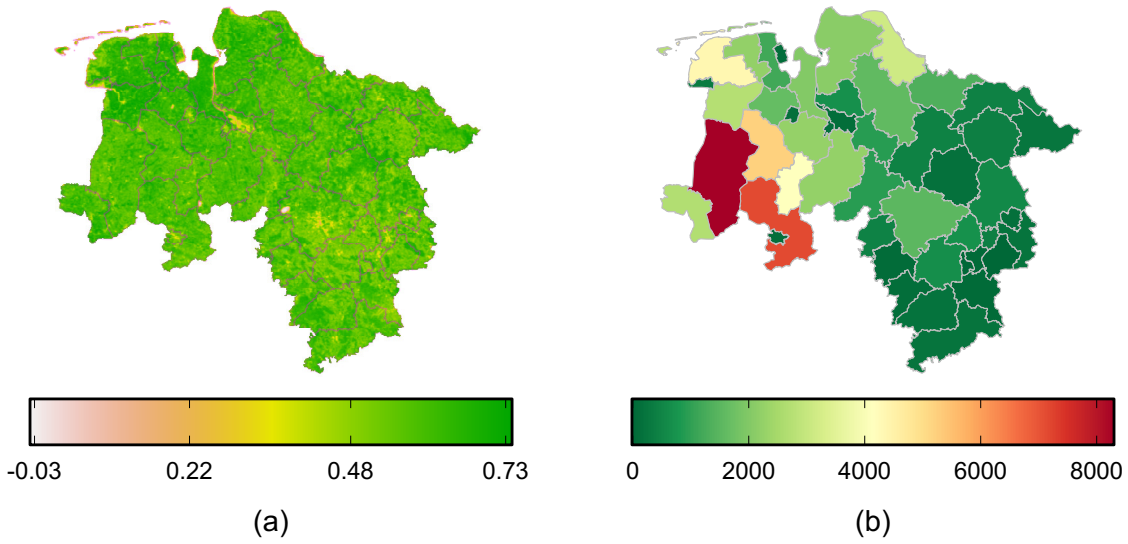


Figure 2.9: (a) The NDVI and (b) brown hare hunting bag data of Lower Saxony during the hunting season 23/24. The borders display the German NUTS-3 district (*Kreise*) level.

2.5.3 Application Results

Applying the steps outlined in Section 2.2.2 again yields a value of $\hat{\gamma} \approx 0.2067$, indicating a rather low, yet still significant correlation between NDVI and the brown hare hunting bag. The resulting density estimates are displayed in Figure 2.10. Unlike our evaluation study in the preceding section, it is not possible to determine RMISE values here because the brown hare hunting bag data is only available at the district level. Panel (a) of Figure 2.10 illustrates the auxiliary density $\hat{P}_{\text{AUX}}(\boldsymbol{s})$ derived from the raw NDVI data. It is important to note that panel (a) has its own distinct color scale and legend, whereas panels (b)-(d) share a common color scale to facilitate comparisons. The figure reveals that the auxiliary density is relatively flat compared to the GRSST. From the global perspective provided by Figure 2.10, we can observe the typical behavior of GRSST, exhibiting a rather smooth density within districts. In contrast, AGRSST1 and AGRSST display more structure within these districts. All three estimators indicate a west-to-east gradient with high bag densities in the west and declining densities to the east. In contrast to the absolute bag data (panel (b) in Figure 2.9), the KDE-based estimators identify a cluster of brown hare hunting in the three districts of Vechta, Cloppenburg, and Osnabrück, with Vechta being the district exhibiting by far the highest bag density. Figure 2.11 zooms in on the district of Vechta to better visualize the discrepancies between the three estimates; the color scales remain consistent.

The AGRSST1 significantly reflects some of the structure of the auxiliary density shown in panel (a); the dark green high vegetation hotspots north and south of Vechta can be clearly identified. Note that we observed a much larger impact of the auxiliary density in our evaluation study. The limited effect of the NDVI is due to two characteristics of the input data. First, the total count of $n = 64\,779$ causes the determinants of the bandwidth matrices $\mathbf{H}$ to be relatively large compared to the setting in our evaluation study, which in turn causes the resulting density estimates to be smoother. Second, the NDVI auxiliary density is generally much smoother than the GRSST density, which is quickly confirmed by comparing the maximum values: $2.5 \cdot 10^{-11}$ for the auxiliary density versus $8.0 \cdot 10^{-11}$ for the GRSST. Also, contrary to our expectation, the AGRSST estimate seems to have higher maximum density values than the AGRSST1 and the GRSST. Two factors contribute to this: first, the AGRSST's derivation without the additional constant in step 3 of Algorithm 2 results in a spikier estimate compared to the GRSST; second, the relatively flat auxiliary density leads to a flatter AGRSST1 estimate. Considering these aspects, we prefer the AGRSST1 approach for this particular application, as it is sensitive to the NDVI input data while maintaining a high degree of smoothness. This balance already helps limit potential errors compared to the GRSST, even without relying on the AGRSST rationale.

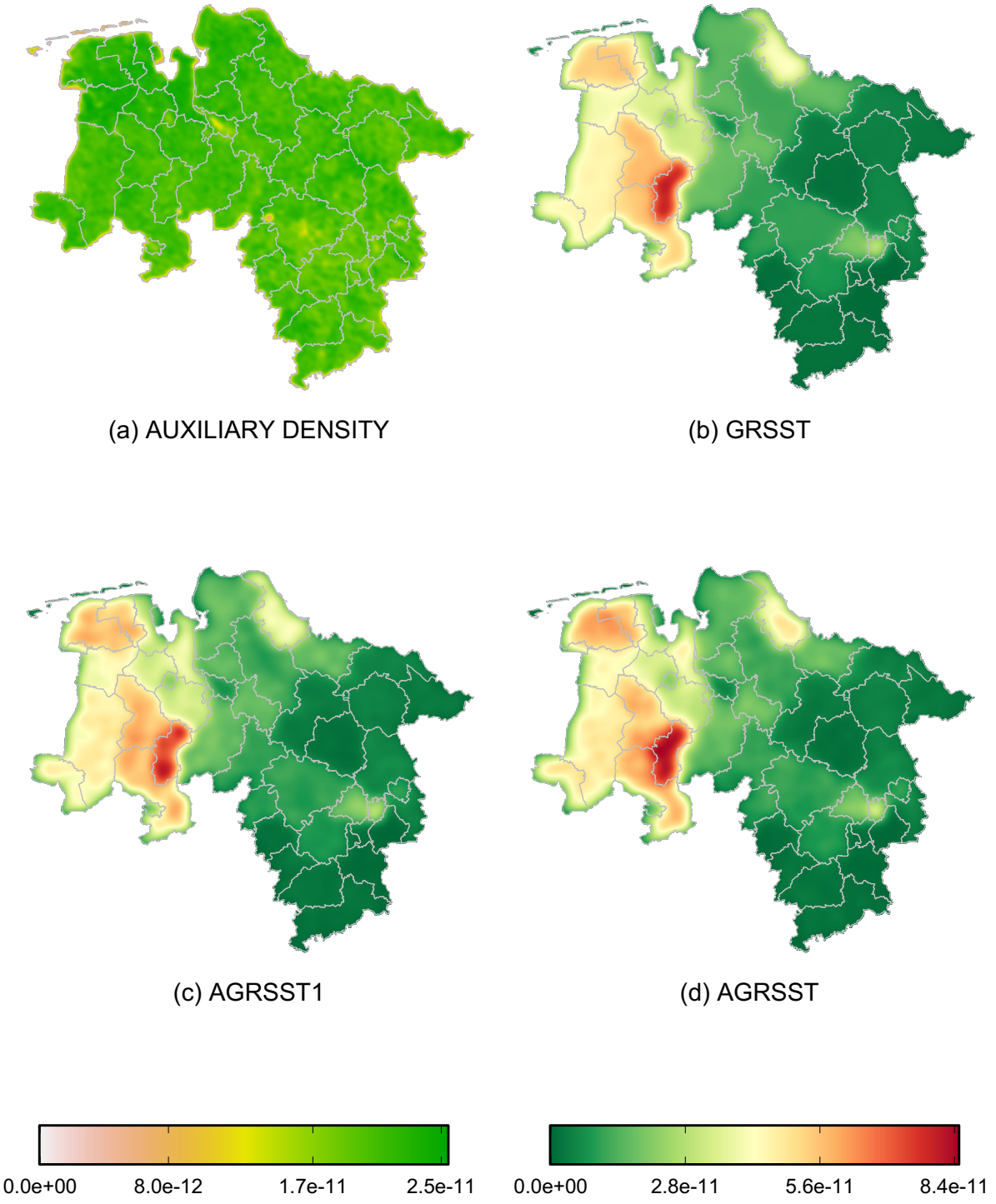


Figure 2.10: (a) The auxiliary density, obtained by applying Algorithm 4 on the raw NDVI data (the corresponding color scale is on the bottom left), (b) the GRSST, (c) AGRSST, and (d) the AGRSST1 estimate (their color scale is on the bottom right).

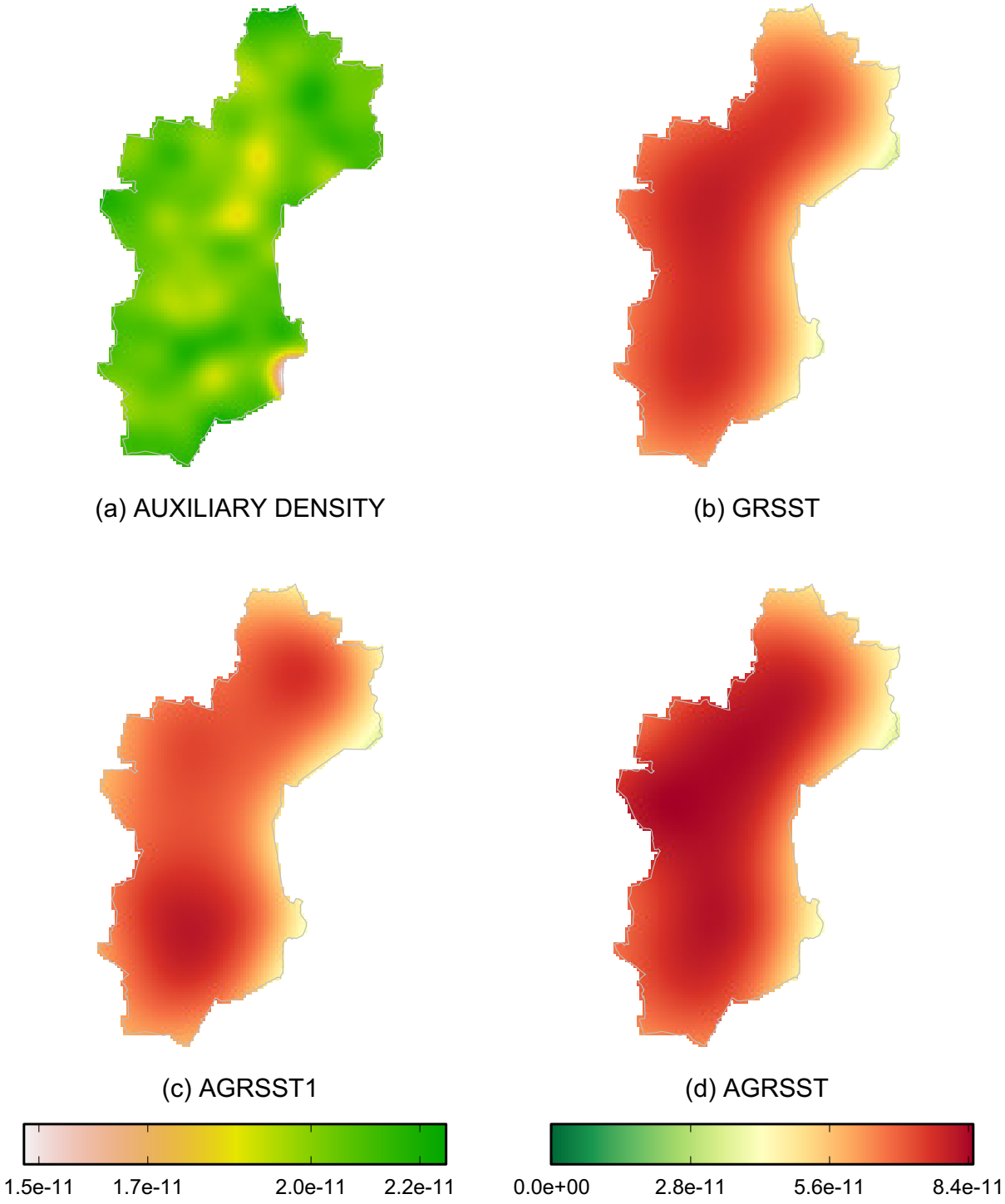


Figure 2.11: A version of Figure 2.10, zoomed-in on the district of Vechta. (a) The auxiliary density, obtained by applying Algorithm 4 on the raw NDVI data (the corresponding color scale is on the bottom left), (b) the GRSST, (c) AGRSST, and (d) the AGRSST1 estimate (their color scale is on the bottom right).

2.6 Conclusion

In this paper, we propose an extension to the GRSST estimator, which we name AGRSST. The GRSST is a method that broadens the applicability of KDE to challenging scenarios involving aggregated data, enabling more profound insights compared to basic choropleth maps displaying absolute counts. These advantages include facilitating the identification of spatial clusters and enabling the transformation of area aggregates between non-hierarchical geographical systems. However, the inherent accuracy of the GRSST remains limited by the relative geographical size of its input areas, as no direct information about the density distribution within these areas is provided.
The AGRSST estimator linearly combines the auxiliary and GRSST densities. The weight assigned to the auxiliary density is determined by its correlation with the GRSST-derived density at the level of the aggregated spatial units. This allows the AGRSST to adapt to the informativeness of the auxiliary data, placing more emphasis on it when it exhibits a strong positive correlation and relying more on the conservative GRSST estimate when the correlation is weak. Furthermore, a benchmarking step ensures that the final density estimate is based on the original aggregated counts. We also highlighted a special version of the AGRSST, the AGRSST1, which directly samples within the aggregated units according to the auxiliary density.
To evaluate the effectiveness of the AGRSST estimator, we conducted a simulation study using the administrative map of Bavaria, Germany. We simulated various true population densities and generated auxiliary densities of varying quality by introducing controlled levels of distortion. The RMISE was used as the primary metric to assess the performance of the standard GRSST and the proposed AGRSST variants across these different distortion levels. The results of our simulation study show that depending on the quality of the auxiliary density, large accuracy improvements are possible when using the AGRSST. It also became apparent that the correlation-based convex combination serves as a mechanism which limits the negative influences of poor auxiliary densities. When fully weighting the auxiliary density (AGRSST1), we observe the potential for greater improvements over the GRSST, but also the risk of worse estimates in terms of RMISE, contingent on the quality of the auxiliary density. These findings are corroborated in a real-world evaluation using German Census data and remotely sensed nighttime lights, where, owing to the high correlation between nighttime lights and population density, both augmented estimators outperform the GRSST, with the AGRSST attaining the lowest RMISE, slightly ahead of AGRSST1.

Lastly, applying AGRSST and AGRSST1 to Lower Saxony's brown hare hunting bag data reveals a west-to-east density gradient, with intense hunting clustered in the Vechta, Cloppenburg, and Osnabrück districts. Comparing the estimators, AGRSST is spikier than GRSST, while only AGRSST1 effectively integrates information from the NDVI data. Considering these factors, we conclude that AGRSST1 is the preferred method for this application.

The findings of this paper open several interesting avenues for future research in this domain. A potentially powerful enhancement could involve the integration of auxiliary densities within each iteration of the GRSST algorithm. Indeed, during the course of this research, we explored one such approach, wherein the standard KDE step in the GRSST was replaced by weighted KDE (see, e.g., Wang and Wang (2007)), with weights determined by the corresponding values of the auxiliary density at the locations of the input geocoordinates. Unfortunately, this specific implementation did not yield significant improvements over the AGRSST presented herein and was thus omitted for the sake of brevity. Another promising direction for future research concerns the incorporation of multiple auxiliary densities. While the optimal strategy for weighting these distinct densities to derive a final estimate remains an open question, a potential initial approach could involve formulating an overall auxiliary density as a linear convex combination of several individual auxiliary densities, with weights potentially determined through a multi-dimensional grid search aimed at maximizing $\hat{\gamma}$, the estimated correlation. However, we also note that a significantly improved (auxiliary) data situation enables more sophisticated methodologies such as random forest-based dasymetric mapping (Stevens et al., 2015). This generally raises questions regarding the relationship between AGRSST and such methods, including when one approach should be preferred over the other, and what potential combinations might be beneficial. An important limitation of the current proposal is the absence of formal uncertainty quantification for the density estimates. Neither the GRSST nor the AGRSST provides confidence bands or standard errors, and developing such measures is an important avenue for future research. Furthermore, future work with significant practical implications includes the extension of the `Kernelheaping` package to incorporate the AGRSST methodology, as well as the introduction of refinement schemes already established for the GRSST. These include the boundary correction technique and the capacity to exclude (uninhabited) areas known to have a density of zero from the estimation process (Erfurth et al., 2022).

Overall, our research demonstrates the considerable potential that auxiliary densities offer in the context of density estimation with aggregated data. We recommend that practitioners consider using AGRSST when the emphasis is on limiting potential adverse effects from incorporating an auxiliary density, and AGRSST1 when the focus is on maximizing potential accuracy. In any case, considering the generally poor initial data situation, which largely prohibits advanced statistical modeling and uncertainty quantification techniques such as cross-validation, a sound theoretical justification for the candidate auxiliary density is indispensable.

Declarations

Funding and/or Conflicts of interests/Competing interests

All authors declare no conflicts of interest.

Code Availability Statement

The code used in this study is available upon request from the corresponding author.

Data Statement

The hunting bag statistics for Bremen and Lower Saxony are available in the Handbook of the German Hunting Association (DJV), whose president is Dammann-Tamke (2025), and can be purchased online. All other datasets used are freely available under their corresponding citations.

Chapter 3

Ordinal Random Forests in Language Data Analysis

Abstract

Ordinal outcome variables are common in different branches of language research. For data analysis, linguists typically convert responses into numbers and then analyze them using averages or (mixed-effects) regression models. As this strategy can yield misleading insights, it is to be welcomed that the methodological literature has started to give due consideration to alternative, more adequate procedures. The current paper joins this discourse by demonstrating a modelling approach based on Random Forests, which have been gaining ground in language research. Due to their ability to uncover complex relationships in the data while actively guarding against overfitting, Random Forests are particularly attractive for exploratory work. The present paper relies on the frequency-adjusted borders ordinal forest (fabOF) framework, which has seen a recent mixed-effects upgrade that handles hierarchically structured datasets in a principled way. To unlock the full interpretative potential of this tool, we have created (previously unavailable) partial dependence routines in R. As we illustrate using a case study on lexical preference ratings in Maltese English, mixed-effects fabOFs can not only exceed ordinal mixed-effects regression models in terms of predictive utility, but also afford nuanced insights into patterns in the data that can be visualized effectively.

Keywords: ordinal data, random forests, partial dependence plots, latent variable model

3.1 Introduction

Ordinal response variables are frequently encountered in linguistic research. They are particularly common in sub-disciplines that rely on rating scales or questionnaires for data collection. In experimental syntax, for instance, participants are often asked to rate the acceptability of a sentence on an ordinal scale (see Cowart, 1997, Schütze and Sprouse, 2014), and in research on social and regional patterns of language variation, informants report on their (perceived) language use (see Krug and Sell, 2013, Dollinger, 2015). The statistical analysis of ordinal outcome variables has received considerable attention in the methodological literature (e.g. Knapp, 1990, Norman, 2010; for linguistics, see Endresen and Janda, 2017, Sönning, 2024).

A systematic review by Sönning et al. (2024) has shown that ordinal outcomes are chiefly analyzed using what Agresti (2010, p. 137) refers to as a mean response model: The ordinal categories are replaced with numeric scores, which are then modelled using procedures such as linear (mixed-effects) regression. While this approach can run into serious problems (see, e.g., Liddell and Kruschke, 2018, Rohrer and Arslan, 2021), it is easy to implement and yields simple and interpretable results. The same is not true for statistically more adequate ordinal regression models, which may explain the preponderance of mean response models in applied work.

A recent paper by Sönning et al. (2024) has illustrated how a particular form of ordinal regression, the cumulative-link model, can be interpreted by reference to a latent-variable framework, which allows for analogous interpretation strategies on a fine-grained continuous scale. The current paper aims to contribute new impulses to the methodological discourse by exploring an alternative modeling approach based on random forests (RFs), a class of predictive models that is becoming increasingly popular in linguistics (see Gries, 2021, ch. 7, Levshina, 2020). Specifically, we will concentrate on frequency-adjusted borders ordinal forests (fabOFs), which were proposed by Buczak (2025b). This framework has recently been extended to deal with hierarchically structured data (Buczak, 2025c), which makes it attractive for linguistic applications, where such data layouts are frequently encountered (see Winter and Grice, 2021, Sonderegger, 2023). A distinctive feature of the fabOF algorithm, which is coherent with the latent-scale regression approach discussed in Sönning et al. (2024), is that it likewise operates on the basis of an unobserved continuous variable. As a result, it comes with the same interpretative advantages, as it is capable of bringing into view fine-grained patterns in the data that can be visualized effectively. Apart from this, it handles data with a hierarchical (or grouping) structure in a principled way. Finally, and perhaps most importantly, RFs provide a very flexible modeling framework, which is particularly attractive for exploratory work. They are capable of capturing non-linear associations that may hold between continuous predictors and the outcome, and they can pick up on (complex) interaction patterns while actively guarding against overfitting.

The aim of the present paper, then, is to showcase the added value of ordinal RFs for language data analysis. For illustration, we draw on an extended version of the dataset used by Sönning et al. (2024). To unlock the interpretative potential of this modeling approach, we illustrate how variable importance scores and partial dependence plots may be used, in tandem, to shed light on systematic patterns of variation in the data. To this end, the first author of this paper has implemented (previously unavailable) partial dependence routines for this type of ordinal random-forest model. To establish continuity with the preceding discourse on ordinal language data analysis, and to allow for an informed discussion of the strengths of the ORF approach, we will refer to a mixed-effects ordinal regression model as a standard of comparison.

With these goals in mind, we have structured the paper as follows. Section 3.2 introduces the dataset for our case study. The two methods we compare are outlined in Section 3.3, starting with a brief recap of the mixed-effects cumulative regression model and an extended treatment of (mix)fabOF models. Section 3.4 then lays out our modeling approach. The results are presented and compared in Section 3.5, followed by a concluding discussion in Section 3.6.

3.2 Data

The data used in this study are drawn from the *Bamberg Survey of Language Variation and Change*, a questionnaire database documenting the use of lexical and grammatical structures in varieties of English (Krug & Sell, 2013). Similar to Sönning et al. (2024), we concentrate on lexical preference ratings, where participants are asked to indicate which of two (largely synonymous) expressions they (tend to) use. These lexical binaries consist of a (traditionally) American English (AmE) variant, and a (traditionally) British English (BrE) variant (e.g. *truck–lorry*, *realization–realisation*; Algeo, 2006). As illustrated in the questionnaire excerpt given in Figure B.1, participants responded on a 5-point scale flanked by the two expressions. Response options ranged from "I always use this expression" at the endpoints of the scale, to the less extreme rating "I use this expression more often" and, finally, "I have no preference" as a neutral midpoint (see Krug and Sell, 2013 for methodological details on the data collection procedure). As we will see further below, the symmetric verbalization of the response scale will have consequences for analysis.

We rely on ratings provided by 1,391 speakers from Malta, which were collected between 2008 and 2023. The initial dataset included 68 lexical pairs and 1,813 individuals. To prepare the data for analysis, we applied several preprocessing steps. First, participants with missing information on relevant socio-demographic variables had to be excluded: For 17% (n = 310), information on the highest qualification was missing, for 12% (n = 222), no data on the language used at home while growing up was available, and for 1% (n = 16), there was no information on gender. This led to the removal of 348 speakers. Then we excluded 7 speakers who had spent less than half their life in Malta and 5 speakers whose home language(s) included neither Maltese nor English. Following Sönning et al. (2024), item pairs with more than 15% missing responses were excluded, leaving 63 lexical pairs for analysis. Finally, only participants who responded to 40 or more item pairs (out of the remaining 63) were retained, which led to the exclusion of 62 individuals. This left us with 1,391 participants. Removal of 2,741 missing responses yielded 84,892 observations (i.e. ratings) in total.

Table 3.1 summarizes the distribution of the variables in our dataset. The first row reports on the spread of responses across the five ordinal categories and reveals a concentration of ratings in the extreme categories (in particular, "always BrE"; see also Figure 3.1 further below). In terms of data structure, the 84,892 observations in our case study are clustered by Participant (n

Variable	Description	Scale and distribution
Rating	Response recorded by participant	Ordinal; Always AmE (n = 14,312 observations, 17%), AmE more often (n = 3,543, 4%), No preference (n = 11,304, 13%), BrE more often (n = 8,200, 10%), Always BrE (n = 47,533, 56%)
Item pair	Pair of expressions	Nominal; 63 item pairs
Participant	Maltese informant	Nominal; 1,391 speakers
Participant-level variables		
Date of birth	Date of birth	Continuous; Range [1920, 2008], deciles: 1959, 1973, 1983, 1989, 1991, 1992, 1994, 1997, 2000
Gender	Self-reported gender	Binary; Female (n = 795, 57%), Male (n = 593, 43%)
Year	Year of data collection	Continuous; 2008 (n = 87, 6%), 2009 (n = 282, 20%), 2013 (n = 429, 31%), 2018 (n = 378, 27%), 2023 (n = 215, 15%)
Highest qualification	Highest level of qualification (completed or ongoing)	Nominal; No/other qualification (n = 217, 15%), Secondary education (n = 283, 20%), Work-based learning (n = 57, 4%), Professional/vocational (n = 97, 7%), Higher education (n = 738, 53%)
Ratio	Proportion of life spent in Malta	Continuous; Range [.53, 1], n = 1,213 (87%) with a Ratio of 1, n = 94 (7%) with Ratio $<$ .90
Home language	Language(s) used at home while growing up	Nominal; Maltese (n = 574, 41%), Maltese and English (n = 525, 38%), English and Maltese (n = 237, 17%), English (n = 55, 4%)

Table 3.1: Descriptive statistics for the variables in the dataset.

= 1,391). This hierarchical layout, which is quite typical of linguistic datasets (see Winter and Grice, 2021, Sonderegger, 2023, pp. 242-244, 251–252) has consequences for the statistical treatment of the data. The analyses we report in the present paper concentrate on two types of predictors. The first is the variable Item pair, which we treat as a fixed effect in the ensuing analyses. Apart from this, interest centers on several participant-level socio-demographic variables, which are listed in Table 3.1. The data used in the present study are available on Zenodo (Krug et al., 2025).

3.3 Methodology

While the main focus in the current paper is on the ordinal RF approach implemented in mixed-effects frequency-adjusted borders ordinal forest (mixfabOF), we also analyze the data using an ordinal regression model. This maintains continuity with Sönning et al. (2024) and, more importantly, provides a basis of comparison for judging the relative performance of mixfabOFs. We therefore start with a brief recapitulation of cumulative link mixed models (CLMMs) in Section 3.3.1. Section 3.3.2 then outlines in more detail the mixfabOF approach to constructing ordinal RFs, including a discussion of the variable importance measures (VIMs) and the newly developed partial dependence plots (PDPs) for model interpretation.

3.3.1 Cumulative Link Mixed Models

The family of ordered regression models includes three basic types (Bürkner and Vuorre, 2019, Fullerton and Xu, 2016). We concentrate on the most commonly used variant, the cumulative link model (CLM; McCullagh, 1980, McKelvey and Zavoina, 1975), which also allows for a latent-variable interpretation (see Wieditz et al., 2024 and Sönning et al., 2024 for more background). The hierarchical structure of the data in our case study (observations grouped by participant) suggests a mixed-effects modeling approach, with Participant as a random-effects term. We therefore turn to a CLMM, which extends the basic cumulative link model (CLM) to accommodate nested data structures by incorporating random effects (see Agresti and Natarajan, 2001 for a survey of mixed-effects models for ordinal data). The comprehensive vignette (Christensen, 2023) of the `ordinal` package in `R` provides details on the structure and implementation of CLMs and serves as a basis for our notation, which we extend below to the mixed-effects case.

CLMMs assume that the observed ordinal response Y_{ig} arises from an unobserved continuous latent variable S_{ig}, where $i = 1, \ldots, n_g$ indexes observations within group $g = 1, \ldots, G$ (participants in the present study). The latent variable is modeled as a linear mixed model, with the superscript * indicating that coefficients are defined on the unobserved latent scale:

$$S_{ig} = \alpha^* + \mathbf{x}_{ig}^\top \boldsymbol{\beta}^* + u_g^* + \varepsilon_{ig}^*, \tag{3.1}$$

where:

- α^* is the overall intercept on the latent scale,
- $\mathbf{x}_{ig}^\top \boldsymbol{\beta}^*$ represents the contribution of fixed effects, with $\mathbf{x}_{ig}$ as the covariate vector and $\boldsymbol{\beta}^*$ as the corresponding coefficients on the latent scale,
- $u_g^* \sim \mathcal{N}(0, \sigma_{u^*}^{*2})$ is the normally distributed random intercept for group g, accounting for unobserved group-level variability,
- $\varepsilon_{ig}^* \sim \mathcal{N}(0, \sigma_{\varepsilon^*}^{*2})$ is the individual-level error term.

The observed ordinal response Y_{ig} takes the value $j \in \{1, \ldots, J\}$ if the latent variable S_{ig} falls between two consecutive thresholds: $\theta_{j-1}^* < S_{ig} \leq \theta_j^*$, where $-\infty \equiv \theta_0^* < \theta_1^* < \cdots < \theta_{J-1}^* < \theta_J^* \equiv \infty$ and J is the number of ordinal categories.

Since the random intercept u_g^* is itself a random variable, cumulative probabilities are formulated conditionally on it, i.e., as group-specific probabilities given the realized value of u_g^*. The cumulative probability for observing $Y_{ig} \leq j$ is then given by $P(Y_{ig} \leq j \mid u_g^*) = P(S_{ig} \leq \theta_j^* \mid u_g^*)$. Substituting the expression for S_{ig} from Equation (3.1) and rearranging results in

$$P(Y_{ig} \leq j \mid u_g^*) = P\left(\frac{\varepsilon_{ig}^*}{\sigma_{\varepsilon^*}^*} \leq \frac{\theta_j^* - \alpha^* - \mathbf{x}_{ig}^\top \boldsymbol{\beta}^* - u_g^*}{\sigma_{\varepsilon^*}^*}\right),$$

where the probability on the right-hand side refers to the distribution of ε_{ig}^* alone, which is independent of u_g^*.

In the model above, $\frac{\varepsilon_{ig}^*}{\sigma_{\varepsilon^*}^*}$ follows a standard normal distribution. We can therefore express this probability using the standard normal cumulative distribution function Φ (the probit link; alternatively, the logit link would assume ε_{ig}^* to follow a logistic distribution). Defining the standardized parameters $\theta_j = (\theta_j^* - \alpha^*)/\sigma_{\varepsilon^*}^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}^*/\sigma_{\varepsilon^*}^*$, and $u_g = u_g^*/\sigma_{\varepsilon^*}^*$, the conditional cumulative probability simplifies to:

$$P(Y_{ig} \leq j \mid u_g) = \Phi(\theta_j - (\mathbf{x}_{ig}^\top \boldsymbol{\beta} + u_g)).$$

Model interpretation often concentrates on the scale of the latent variable, S_{ig}, by adopting standard identifying assumptions: $\varepsilon_{ig}^* \sim \mathcal{N}(0, 1)$ (i.e., $\sigma_{\varepsilon^*}^* = 1$) and $\alpha^* = 0$. Under these constraints, the latent thresholds $\boldsymbol{\theta}^*$ and coefficients $\boldsymbol{\beta}^*$ become equivalent to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, which can be estimated. This allows $\eta_{ig} = \mathbf{x}_{ig}^\top \boldsymbol{\beta} + u_g$ to be directly interpreted as the conditional mean of the latent variable S_{ig}.

The effect of changes in covariates can therefore be understood as shifts in the latent distribution: when η_{ig} increases, the distribution of $S_{ig} \sim N(\eta_{ig}, 1)$ shifts rightward while the

thresholds $\boldsymbol{\theta}$ remain fixed, moving probability mass across the threshold boundaries.

3.3.2 Mixed-Effects Ordinal Forest

This section develops the rationale underlying the mixed-effects ordinal forest approach proposed by Buczak et al., 2025, a tree ensemble method designed to accommodate ordinal responses in data with a hierarchical structure. It extends the simpler fabOF approach (Buczak, 2025b) by incorporating group-specific effects, such as the random intercepts u_g established in the CLMM context (Section 3.3.1). It should be noted, however, that the current state of the mixfabOF implementation only supports random effects for a single grouping variable. We will return to this caveat in Section 3.6. Like the CLMM, both methods adopt the latent variable framework wherein the ordinal response arises from an unobserved continuous variable S_{ig} discretized via category thresholds. We first consider the basic fabOF algorithm (Section 3.3.2). Section 3.3.2 then moves on to its mixed-effects extension (mixfabOF), the method that takes center stage in the present study. This is followed by a brief discussion of VIM for mixfabOF (Section 3.3.2) as well as the newly developed software implementation of PDPs for this modeling framework (Section 3.3.2).

Before we go further, it is worth noting that both fabOF and mixfabOF represent recent RF-based implementations of the latent-variable approach to modeling ordinal data. Ordinal Forest (Hornung, 2019) and Ordinal Score Optimization Algorithm (Buczak et al., 2025) can be viewed as predecessors to fabOF, which apply optimization algorithms to find category borders. A score-free alternative to fabOF is the Split-based Ordinal Forest (Tutz, 2021), whereas Ordinal Mixed-Effect Random Forest (Bergonzoli et al., 2024), which also includes random-effects terms, offers an alternative to mixfabOF. The focus in the present paper is on fabOF and mixfabOF due to their conceptual simplicity, computational advantages and competitive predictive performance (see Buczak, 2025b and Buczak, 2025c). For an extensive simulation study comparing many of the above-mentioned RF methods and CLMM, we refer to Buczak et al. (2025).

Frequency-adjusted borders ordinal forest: fabOF

Like the CLMM, fabOF considers the observed ordinal response Y_l as a coarse realization of an unobserved continuous latent variable S_l, where $l = 1, \ldots, N$ indexes individual observations. It adapts a standard RF for regression to the ordinal prediction task through the following steps:

1. **Score Assignment:** A numeric auxiliary score $S^{(j)}$ (e.g., $1, 2, \ldots, J$) is assigned to each of the J ordinal response categories.

2. **Regression RF:** A standard regression RF is trained on the assigned numeric auxiliary scores $S_l^{(j)}$ using the covariate vectors $\mathbf{x}_l$. This provides a non-linear estimate of the fixed effects component, denoted as $f(\mathbf{x})$.

3. **Border Determination (Frequency-Adjusted Heuristic):** Instead of relying on parameter optimization, fabOF determines the $J-1$ category borders (termed thresholds in the CLMM framework) $\boldsymbol{b} = (b_1, \ldots, b_{J-1})$ using a frequency-based heuristic. Specifically, if the original training data contains proportions $p_1, p_2, \ldots, p_J$ of observations in categories $1, 2, \ldots, J$ respectively, the borders are set as the (p_1)-th, $(p_1 + p_2)$-th, ..., $(p_1 + \cdots + p_{J-1})$-th empirical quantiles of the out-of-bag (OOB) numeric predictions. This ensures that the predicted category distribution matches the observed training distribution.

A numeric RF prediction $S_l = f(\mathbf{x}_l)$ is transformed back into an ordinal category by finding the interval (defined by the borders $\boldsymbol{b}$) that contains the prediction. Step 1 of the fabOF device shows that this method relies on a scoring system, i.e. a set of numeric values that are assigned to the response options. It should be noted that in cases where the ordered categories are not perceived as representing equally-spaced increments on the underlying continuum, custom scoring systems may be used (see Labovitz, 1967, p. 155, Worcester and Burns, 1975, p. 191, Sönning, 2024 for discussion).

Mixed-effects fabOF: mixfabOF

The mixfabOF method extends the fabOF algorithm to clustered data settings. This is to say that it is capable of accounting for group-level variability by incorporating the group-specific random effects u_g, directly extending the latent variable assumption to hierarchical data

$$S_{ig} = f(\mathbf{x}_{ig}) + u_g + \varepsilon_{ig},$$

where $f(\mathbf{x}_{ig})$ is the non-linear prediction from the RF ensemble (the fixed-effects part), $u_g \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for group (in our case, participant) g, and $\varepsilon_{ig} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual error. This structure directly parallels the latent variable formulation of the CLMM (Equation (3.1)), with the key distinction being that mixfabOF replaces the linear predictor $\mathbf{x}_{ig}^\top \boldsymbol{\beta}$ with the non-linear function $f(\mathbf{x}_{ig})$, which is estimated through the RF. The estimation is based on an iterative procedure (similar to an expectation-maximization (EM) algorithm), adapting the strategy of the Mixed-Effect Random Forest (MERF; Hajjem et al., 2014). As in fabOF, numeric auxiliary scores $S^{(j)}$ (e.g., $1, 2, \ldots, J$) are first assigned to each ordinal category. The procedure then initializes all random effects to zero ($\hat{u}_g = 0$ for all groups g), and alternates between two main steps:

1. **Fixed Effects (RF) Step:** A standard regression RF is trained to estimate $f(\mathbf{x}_{ig})$ using a pseudo-response variable $\tilde{S}_{ig} = S_{ig} - \hat{u}_g$, constructed by subtracting the current estimate of the random effect $\hat{u}_g$ from the numeric score S_{ig}. This separates the group-level effect from the fixed-covariate relationships, allowing the RF to focus on estimating the non-linear function $f(\mathbf{x}_{ig})$.

2. **Random Effects (Mixed Model) Step:** The residuals from the RF fit are used to estimate the current group-specific random effects $\hat{u}_g$ and the corresponding variance components (e.g., σ_u^2).

This iterative process refines the estimates of the non-linear fixed-effects model and the random-effects structure until convergence, which is determined by monitoring the change of a generalized log-likelihood criterion (cf. Hajjem et al., 2014) across iterations. In the final step, the frequency-based heuristic is applied to the predictions to define the final set of borders $\boldsymbol{b}$.

The prediction on the latent-variable scale for observation i in group g, then, is given by:

$$\hat{S}_{ig} = \hat{f}(\mathbf{x}_{ig}) + \hat{u}_g.$$

This numeric prediction $\hat{S}_{ig}$ is then mapped to a response category based on the determined borders $\boldsymbol{b}$, providing a prediction that benefits from the predictive power of RFs while statistically accounting for the non-independence of observations through the $\hat{u}_g$ term.

Variable Importance Measures

VIMs (see e.g. Molnar, 2025 for a practical overview) quantify the relative contribution of each predictor to model performance, addressing a key question in predictive modeling: which covariates are most influential in determining the outcome? For mixfabOF, we employ a permutation-based importance metric adapted to the model's hierarchical structure with random effects.

The standard permutation importance approach for RFs (Breiman, 2001) operates by randomly permuting each predictor's values and measuring the resulting drop in prediction accuracy. However, this approach requires modification for mixfabOF in two respects. First, because the response is ordinal rather than continuous, prediction performance should arguably be evaluated using a metric that pays attention to the ordered category structure and penalizes ordinally distant misclassifications more heavily than adjacent ones. We therefore use the linearly weighted Cohen's Kappa (κ_w; see Section 3.5) rather than mean squared error or simple accuracy. Second, because mixfabOF incorporates random effects $\hat{u}_g$ in addition to the RF-based fixed effects $\hat{f}(\mathbf{x}_{ig})$, predictions must include both components when evaluating performance: $\hat{S}_{ig} = \hat{f}(\mathbf{x}_{ig}) + \hat{u}_g$.

Following Buczak (2025b) and Buczak (2025c), the mixfabOF variable importance procedure uses out-of-bag (OOB) predictions to avoid overfitting bias, with each tree's predictions evaluated only on observations that were not used in training that tree. For statistical stability, the permutation is repeated multiple times and importance values are averaged across replications to reduce sampling variability. Algorithm 6 formalizes the complete procedure.

Algorithm 6 Variable Importance for mixfabOF

Input: Fitted mixfabOF model, number of importance replications R

Output: VIM $= (\text{VIM}_1, \ldots, \text{VIM}_p)$ for p predictors

for each observation i in group g **do**
  $\hat{S}_{ig}^{\text{OOB}} \leftarrow \text{OOB-predict}(\hat{f}, \mathbf{x}_{ig}) + \hat{u}_g$
  $\hat{Y}_{ig} \leftarrow \text{categorize}(\hat{S}_{ig}^{\text{OOB}}, \boldsymbol{b})$
$\kappa_{\text{orig}} \leftarrow \text{linearKappa}(\mathbf{Y}, \hat{\mathbf{Y}})$
for each predictor $j = 1, \ldots, p$ **do**
  Initialize $\kappa_{\text{perm}}[j] \leftarrow$ vector of length R
  for replication $r = 1, \ldots, R$ **do**
    if clusterwise permutation **then**
      for each group g **do**
        $\mathbf{x}_{(\cdot)j}^{\text{perm}}[g] \leftarrow \text{permute}(\mathbf{x}_{(\cdot)j}[g])$
    else
      $\mathbf{x}_{(\cdot)j}^{\text{perm}} \leftarrow \text{permute}(\mathbf{x}_{(\cdot)j})$
    $\mathbf{x}^{\text{perm}} \leftarrow (\mathbf{x}_{(\cdot)1}, \ldots, \mathbf{x}_{(\cdot)j}^{\text{perm}}, \ldots, \mathbf{x}_{(\cdot)p})$
    for each observation i in group g **do**
      $\hat{S}_{ig}^{\text{OOB, perm}} \leftarrow \text{OOB-predict}(\hat{f}, \mathbf{x}_{ig}^{\text{perm}}) + \hat{u}_g$
      $\hat{Y}_{ig}^{\text{perm}} \leftarrow \text{categorize}(\hat{S}_{ig}^{\text{OOB, perm}}, \boldsymbol{b})$
    $\kappa_{\text{perm}}[j][r] \leftarrow \text{linearKappa}(\mathbf{Y}, \hat{\mathbf{Y}}^{\text{perm}})$
  $\text{VIM}_j \leftarrow \kappa_{\text{orig}} - \text{mean}(\kappa_{\text{perm}}[j])$
return VIM

The importance value VIM_j quantifies the average degradation in ordinal prediction performance (measured by κ_w) when predictor x_j is randomly permuted. Variables with high importance values cause substantial performance loss when their association with the response is broken, indicating strong predictive relationships. Conversely, near-zero or negative importance values indicate weak or absent associations; negative values arise from random sampling variability when permutation occasionally improves performance by chance, and can be interpreted as effectively zero importance (see Janitza et al., 2018).

An alternative approach is cluster-wise permutation, wherein predictor values are permuted only within each group g rather than across the entire dataset (Buczak, 2025c). This approach preserves any group-level associations between predictors and unmeasured cluster characteristics, isolating within-cluster predictive contributions. However, cluster-wise permutation has an important limitation for hierarchical data: participant-level variables (e.g. Date of birth or Gender), which are constant within clusters (i.e. individuals), receive zero importance by construction, since permuting these values among observations within the same participant leaves predictions unchanged. In preliminary analyses, applying cluster-wise permutation did not

substantially alter the importance ranking of the item variable (which varies within participants and can be meaningfully permuted at the observation level). We therefore employ standard (across-cluster) permutation importance in the present analysis.

Partial Dependence Plots

While VIM identifies which predictors are important, it does not reveal the functional form of their relationship with the outcome. Partial dependence plots (PDPs) (Friedman, 2001) address this gap by visualizing the marginal effect of a predictor x_p on the predicted response, averaging over the distribution of all other covariates. For a fitted model $\hat{f}(\mathbf{x})$, the partial dependence function for predictor x_p is defined as

$$\text{PD}(x_p) = \mathbb{E}_{\mathbf{x}_{-p}}\left[\hat{f}(x_p \mid \mathbf{x}_{-p})\right] \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_p \mid \mathbf{x}_{i,-p}),$$

where $\mathbf{x}_{-p}$ denotes all predictors except x_p, the subscript in $\mathbb{E}_{\mathbf{x}_{-p}}$ indicates that the expectation is taken with respect to the marginal distribution of $\mathbf{x}_{-p}$, and the expectation is approximated by averaging predictions across all n observations in the training data, each time substituting the target value x_p while retaining the individual's actual values for all other predictors.

For mixfabOF with random effects, predictions take the form $\hat{S}_{ig} = \hat{f}(\mathbf{x}_{ig}) + \hat{u}_g$, combining the RF-estimated fixed effects $\hat{f}(\mathbf{x}_{ig})$ with the estimated random intercept $\hat{u}_g$ for group g. When constructing PDPs, random effects can either be excluded (yielding population-level marginal effects) or included (yielding group-specific predictions).

In addition to the aggregated partial dependence scores, individual conditional expectation (ICE) scores (Goldstein et al., 2015) express $\hat{S}_{ig}$ as a function of x_p for each observation i separately; "conditional" therefore refers to holding fixed the specific covariate values of that case (which, in our case study, would correspond to a particular observation, i.e. rating). ICE scores reveal heterogeneity in predictor effects across observations, which can be suggestive of potential interactions that may be obscured by averaging in standard PDPs (see Grafmiller, 2023 for a discussion in the context of language data analysis).

PDPs can be further refined to isolate potential interactions among predictor variables by conditioning on an additional explanatory variable x_q, where p and q index different predictors (Strobl et al., 2026, Goldstein et al., 2015). Different partial dependence functions for the predictor x_p are then approximated, one for each level of x_q. This reveals whether (and how) the marginal effect of predictor x_p varies across subgroups defined by x_q. We will therefore distinguish between:

- **Standard PDP**: Evaluates $\text{PD}(x_p) = n^{-1}\sum_{i=1}^{n} \hat{f}(x_p \mid \mathbf{x}_{i,-p})$, showing the average marginal effect of x_p on the latent score across all observations.

- **Conditional PDP (c-PDP)**: For a conditioning variable x_q (e.g., Gender or Item), computes separate partial dependence scores for each level ℓ of x_q: c-PD$(x_p \mid x_q = \ell) = n_\ell^{-1} \sum_{i:x_{iq}=\ell} \hat{f}(x_p \mid \mathbf{x}_{i,-p})$, where n_ℓ is the number of observations with $x_{iq} = \ell$. By stratifying ICE curves according to x_q and computing group-specific PDPs, c-PDPs reveal whether the effect of x_p varies across levels of x_q, thereby exposing potential (two-way) interaction effects. For example, a c-PDP of Date of birth conditional on Gender would show whether age effects differ between male and female respondents.

When visualizing PDPs for mixfabOF, we overlay the estimated category borders $\boldsymbol{b}$ as reference lines, allowing direct interpretation of how changes in x_p influence the predicted ordinal category by shifting the latent score $\hat{S}_{ig}$ across category boundaries.

Software Implementation

To facilitate reproducibility and enable broader adoption of these methods in linguistic research, we provide an extended implementation of the mixfabOF framework. We forked the original `fabOF R` package (Buczak, 2025a) and contributed several methodological enhancements specifically for the mixed-effects variant: (1) partial dependence plotting functionality for mixfabOF, enabling the visualization techniques described in Section 3.3.2; (2) a dedicated plotting function for mixfabOF variable importance (Section 3.3.2); and (3) parallel computation support for variable importance calculations, to reduce computation time for large datasets. The extended package is freely available via GitHub at https://github.com/muehlbauer-m/fabOF.

3.4 Modeling

In this section, we discuss our core modeling decisions and motivate the selection and scaling of predictor variables. Following some general comments, we deal with aspects specific to the CLMM analysis (Section 3.4.1) and the mixfabOF approach (Section 3.4.2).

To start with, the main drivers of our choice for CLMM and mixfabOF are (1) the ordinal scale of our dependent variable Y_{ig} (see Figure 3.1) and (2) the hierarchical data structure arising from multiple observations per participant. Both of these features, which regularly appear in applied research, may be accounted for by various modeling approaches (see Agresti and Natarajan, 2001). They also typically arise in linguistic work relying on acceptability judgments or questionnaires.

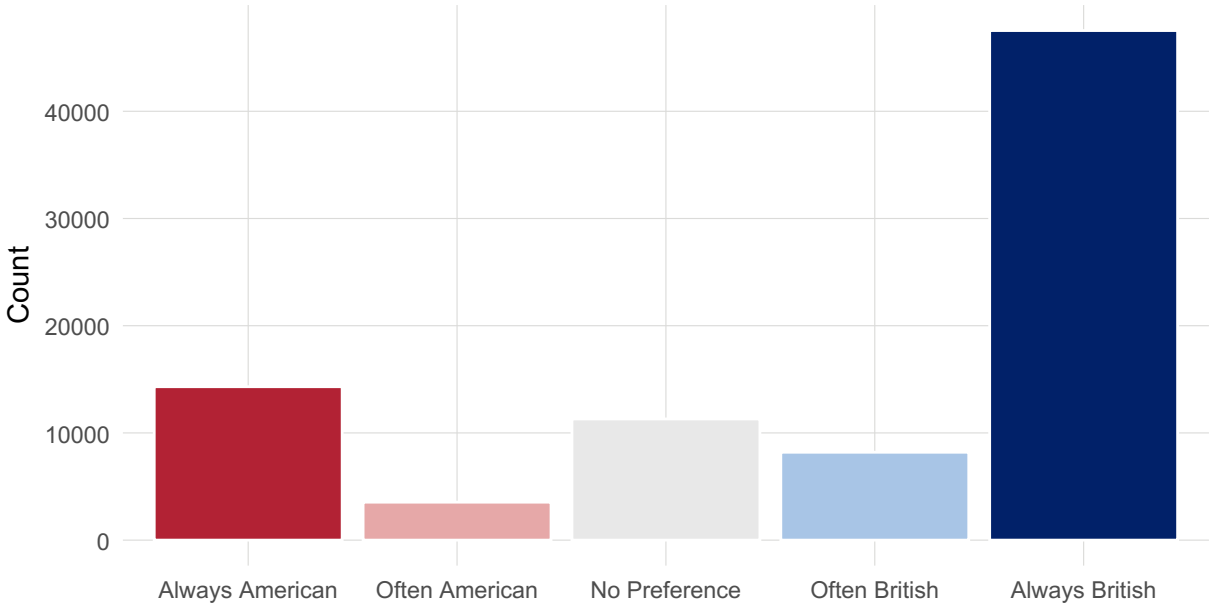


Figure 3.1: Distribution of the Y_{ig} variable Rating.

Our case study takes an exploratory approach to the analysis of lexical variation in Malta. Apart from the variables Item pair and Date of birth, which have previously been documented to show an association with speakers' lexical preferences, we examine a number of additional socio-demographic characteristics, which are listed in Table 3.1 above. These cover a range of sociolinguistic categories that have been observed to be linked to language variation and change more generally.

3.4.1 Cumulative Link Mixed Model

For regression modeling, a choice must be made on the scaling of variables. We opted for the following:

- Categorical predictors (Gender, Item pair, Year, Home language, Highest qualification) are coded using sum-to-zero contrasts, ensuring that coefficient estimates represent deviations from the conditional grand mean of the latent variable across all factor levels rather than comparisons to a reference category.
- Continuous predictors (Date of birth, Ratio) were standardized prior to modeling to support model convergence (see e.g. Gill et al., 2019).

In CLMMs, estimation of the $J-1$ threshold parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{J-1})^\top$ (corresponding to $\theta_1^*, \ldots, \theta_{J-1}^*$ from the latent-variable formulation in Section 3.3.1) may be constrained in different ways (referred to as structured thresholds in Christensen, 2023). Given the symmetric wording of our ordinal response scale, we decided that thresholds should likewise be symmetric (see Sönning et al., 2024 for discussion). Instead of estimating $J-1$ thresholds independently, symmetric thresholds are derived from a smaller set of K underlying parameters $\boldsymbol{\alpha} \in \mathbb{R}^K$ with

$K < J - 1$ via a transformation matrix $\mathbf{J} \in \mathbb{R}^{(J-1)\times K}$:

$$\boldsymbol{\theta} = \mathbf{J}\boldsymbol{\alpha}$$

For a response variable with $J = 5$ categories, we employ the 'symmetric' threshold structure as implemented in the function `clmm()` of the `R` package `ordinal` (Christensen, 2023), which constrains the thresholds to be symmetric around a central value. Incorporating both the random effects and the structured thresholds, then, the full extended ordinal regression model is:

$$P(Y_{ig} \leq j \mid u_g) = \Phi((\mathbf{J}\boldsymbol{\alpha})_j - (\mathbf{x}_{ig}^{\top}\boldsymbol{\beta} + u_g)) \tag{3.2}$$

The linear predictor for the baseline CLMM (Eq. (3.2)) takes the specific form

$$\begin{aligned} P(Y_{ig} \leq j \mid u_g) = \Phi\big((\mathbf{J}\boldsymbol{\alpha})_j - \big(&\mathbf{w}_i^{\text{item}\top}\boldsymbol{\beta}_{\text{item}} + \beta_{\text{dob}} \cdot \text{dob}_{ig} + (\mathbf{w}_i^{\text{item}\top}\boldsymbol{\beta}_{\text{item:dob}}) \cdot \text{dob}_{ig} \\ &+ \mathbf{w}_i^{\text{gender}\top}\boldsymbol{\beta}_{\text{gender}} + \mathbf{w}_{ig}^{\text{year}\top}\boldsymbol{\beta}_{\text{year}} + \beta_{\text{ratio}} \cdot \text{ratio}_{ig} \\ &+ \mathbf{w}_{ig}^{\text{lang}\top}\boldsymbol{\beta}_{\text{lang}} + \mathbf{w}_{ig}^{\text{qual}\top}\boldsymbol{\beta}_{\text{qual}} + u_g)\big), \end{aligned}$$

with fixed effects incorporated via sum-to-zero contrast coding for categorical predictors and standardization for continuous predictors. Specifically, $\mathbf{w}_i^{\text{item}}$ denotes the contrast vector for the factor Item at observation i, with corresponding coefficient vector $\boldsymbol{\beta}_{\text{item}}$ containing 62 estimated parameters. Under sum-to-zero contrast coding, the coefficient for the base level (item pair 63) is implicitly defined as $-\sum_{a=1}^{62}\beta_{\text{item}_a}$, ensuring that coefficients sum to zero across all 63 items. The Item × Date of birth interaction is modeled as $(\mathbf{w}_i^{\text{item}\top}\boldsymbol{\beta}_{\text{item:dob}}) \cdot \text{dob}_{ig}$, allowing items to show different apparent-time trajectories (see Sönning et al., 2024). Similarly, $\mathbf{w}_i^{\text{gender}}$ represents the Gender contrast vector with coefficient vector $\boldsymbol{\beta}_{\text{gender}}$, while $\mathbf{w}_{ig}^{\text{year}}$, $\mathbf{w}_{ig}^{\text{lang}}$, and $\mathbf{w}_{ig}^{\text{qual}}$ encode Year, Home language, and Highest-qualification effects, respectively, with their corresponding coefficient vectors $\boldsymbol{\beta}_{\text{year}}$, $\boldsymbol{\beta}_{\text{lang}}$, and $\boldsymbol{\beta}_{\text{qual}}$.[1] The continuous predictors, standardized Date of birth (dob_{ig}) and standardized proportion of life spent in Malta (ratio_{ig}), enter the model directly with scalar coefficients β_{dob} and β_{ratio}.

To account for the hierarchical structure of the data (multiple observations per individual g), by-participant random intercepts $u_g \sim N(0, \sigma_u^2)$ are included. On the one hand, these capture individual-specific deviations from the overall mean and thereby document the residual between-speaker variability in the overall inclination toward BrE or AmE, which is of linguistic interest. Further, and more importantly, random effects account for the fact that multiple observations from the same participant are correlated rather than independent, ensuring that standard errors and significance tests appropriately reflect the actual number of independent

[1] For example, the sum-to-zero contrast vector $\mathbf{w}_{ig}^{\text{year}\top}$ takes the form $(1, 0, 0)$ for a survey year 2008 observation, and $(-1, -1, -1)$ for the constrained base level year 2018.

participants rather than the total number of observations. The inclusion of by-speaker random effects has therefore become common practice in linguistic research (see Baayen et al., 2008, Barr et al., 2013, Winter and Grice, 2021).[2]

3.4.2 mixfabOF

Where possible, we adopt the aforementioned design choices for our mixfabOF modeling approach. We deviate by not explicitly modeling the Item × Date of birth interaction due to the inherent ability of RFs to automatically capture complex relationships between Y_{ig} and $\mathbf{x}_{ig}$. Categorical predictors enter as native factors (via ranger's `respect.unordered.factors = "order"` option; Wright and Ziegler, 2017), so each categorical variable receives a single (mix-)fabOF adjusted VIM (see Buczak et al., 2025 and Buczak, 2025b). This global variable-importance assessment is followed by an inspection of (conditional) partial dependence plots in Section 3.5.3, to examine the relationship between the predictors and the response variable. Here, our linguistic interest primarily attaches to the 63 item pairs, which have been observed to exhibit distinct patterns of variation in previous work (Krug et al., 2016, Krug et al., 2020).

3.5 Results

We come now to the results of our analysis. Section 3.5.1 starts by comparing the predictive performance of the CLMM and mixfabOF models. Section 3.5.2 then considers the use of VIMs for comparing the predictive capacity of the continuous predictor variables and subgroups represented by categorical predictors. We then turn to the use of partial dependence and conditional partial dependence plots for model interpretation. Section 3.5.3 showcases the utility of these newly implemented routines for shedding light on patterns of association and how they vary across the predictor space.

3.5.1 Predictive Performance

We start the analysis of our results by evaluating them on a 20% hold-out test sample. Let $\mathcal{T}$ denote the set of test observations, where each case is indexed by (i, g) representing observation i by participant g. The two measures used are accuracy and the linearly weighted Cohen's Kappa (κ_w, 1968). The first metric, accuracy, is the proportion of correctly classified observations in the test set, defined as:

$$\text{Accuracy} = \frac{\sum_{(i,g)\in\mathcal{T}} \mathbb{1}_{\left\{\hat{Y}_{ig}=Y_{ig}\right\}}}{|\mathcal{T}|}$$

where $|\mathcal{T}|$ is the total number of observations in the test set, Y_{ig} is the observed response for observation i by participant g, $\hat{Y}_{ig}$ is the predicted response, and $\mathbb{1}(\cdot)$ is the indicator function.

[2]This theoretical justification is further supported by marginal improvements in out-of-sample predictive performance when comparing mixed-effects versions to non-mixed-effects counterparts. For brevity, we focus exclusively on the mixed-effects specification in this paper.

Second, κ_w is a robust metric for ordinal data that accounts for chance agreement and penalizes misclassifications proportionally to the distance between the observed and predicted categories. It is defined as:

$$\kappa_w = \frac{P_o^w - P_e^w}{1 - P_e^w}$$

where P_o^w is the observed weighted agreement, and P_e^w is the expected weighted agreement under chance. The metric ranges from -1 to 1, where higher values indicate better predictive performance: $\kappa_w = 1$ signifies perfect agreement, $\kappa_w = 0$ indicates performance no better than random chance, and negative values suggest systematic disagreement worse than chance. Let p_{jk} denote the proportion of observations where the observed category is j and the predicted category is k, and let $p_{j\cdot}$ and $p_{\cdot k}$ be the respective marginal probabilities. The observed and expected weighted agreements are:

$$P_o^w = \sum_{j=1}^{J}\sum_{k=1}^{J} w_{jk}p_{jk} \quad ; \quad P_e^w = \sum_{j=1}^{J}\sum_{k=1}^{J} w_{jk}p_{j\cdot}p_{\cdot k}$$

For linear weighting, the weights w_{jk} are calculated as:

$$w_{jk} = 1 - \frac{|j-k|}{J-1}.$$

Figure 3.2 illustrates the confusion matrix structure showing how the proportions p_{jk}, marginals $p_{j\cdot}$ and $p_{\cdot k}$, and weights w_{jk} combine to compute weighted Cohen's Kappa.

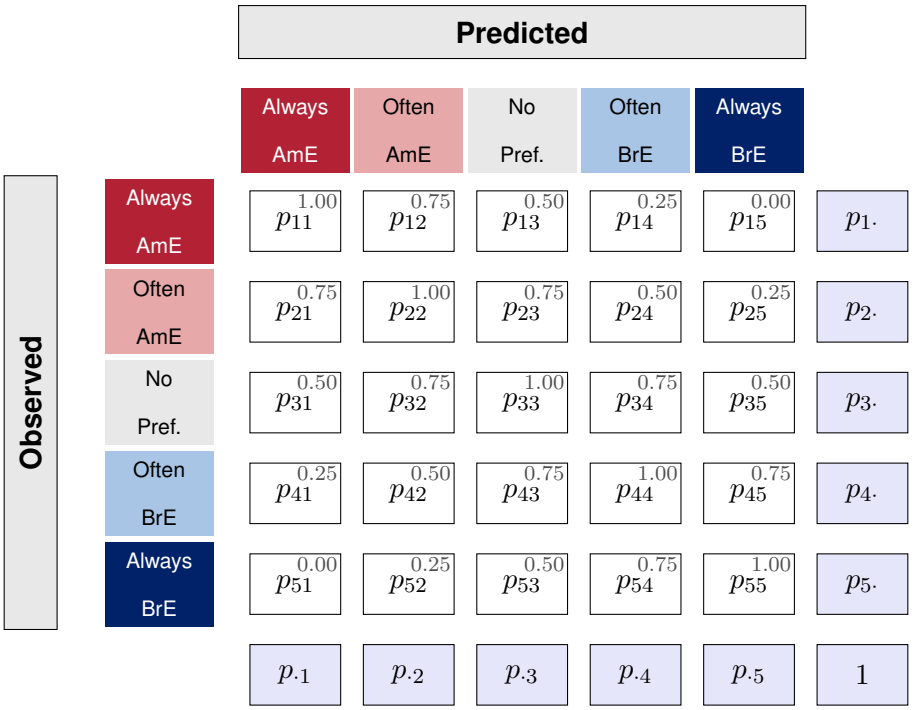


Figure 3.2: Confusion matrix for weighted Cohen's Kappa. Each cell p_{jk} shows the proportion where the observed category is j and the predicted is k. Weights w_{jk} (gray, at top right corners) decrease linearly with ordinal distance. Marginals $p_{j\cdot}$ and $p_{\cdot k}$ (lavender) are used for the expected agreement calculation (P_o^w).

Model	Accuracy	Cohen's Kappa
`mixfabOF`	0.539	**0.418**
`clmm`	**0.603**	0.345

Table 3.2: Model performance comparison on test data.

Table 3.2 depicts the performance of the two fitted models on the 20% ($|\mathcal{T}| = 16,979$) hold-out test set. The results reveal a trade-off between the two evaluation metrics. The CLMM achieves a higher accuracy value (0.603 vs. 0.539), while the mixfabOF model demonstrates superior performance when assessed using Cohen's Kappa (0.418 vs. 0.345). According to conventional benchmarks suggested by Landis and Koch (1977), these would be considered "moderate" and "fair", respectively. This points to a limitation of using accuracy as the sole evaluation criterion for ordinal data, particularly when the distribution of categories is highly imbalanced. As shown in Figure 3.1, the observed ratings are heavily concentrated in the extreme categories. Specifically, approximately 56% of observations fall in category 5 ("always British") and 17% in category 1 ("always American"), while the middle categories contain substantially fewer observations (category 2: 4%; category 3: 13%; category 4: 10%).

An examination of the predictions from the CLMM model (see Figure 3.3) reveals what may be considered a degenerate prediction pattern: the model predicts exclusively categories 1 and 5, making zero predictions for the intermediate categories 2, 3, and 4. This has to do with the

way in which ordered regression models generate predictions. As discussed by Sönning et al. (2024, p. 92), the model forecasts the response category that receives the largest probability mass. If intermediate categories are infrequent in the data, the spacing of the thresholds will be quite narrow, meaning that the amount of probability mass that is found between thresholds is constrained. Even if the predicted value on the latent scale falls between two thresholds, this does not necessarily mean that the response category represented by this interval will have the highest probability. In other words, the predicted value of the S_{ig} need not coincide with the predicted response category.

The mixfabOF approach, by contrast, produces predictions across all five ordinal categories. When misclassifications occur, they tend to involve adjacent or nearby categories, resulting in ordinally coherent errors that are appropriately rewarded by the Kappa metric. This difference in prediction patterns suggests that the CLMM's parametric structure—which assumes a linear relationship between predictors and the latent response scale—may struggle to fully capture the complexity of relationships in these data, whereas the flexible, non-parametric nature of RFs appears better suited to accommodate potential non-linearities and interactions.[3] Of course, the relative inflexibility of CLMMs can to some extent be cushioned by specifying splines or smooth terms, which allows for non-linear effects of quantitative predictors. However, the detection of complex interaction patterns remains a challenge in the regression framework, as these need to be specified (i.e. actively built into the model) by the analyst.

The RF-based methods, by contrast, make no parametric assumptions about functional form, allowing them to adaptively learn interactions and non-linearities that produce ordinally coherent predictions across all response categories. This pattern is consistent with findings from Buczak et al. (2025), who compared CLMM performance to various RF-based predecessors of mixfabOF in a comprehensive simulation study. Their results demonstrate a similar drop in linearly weighted Cohen's Kappa for CLMM relative to RF-based methods when the data-generating process is highly non-linear and the sample size is large, suggesting that parametric constraints become increasingly problematic as the complexity of the underlying relationships and the amount of available information increase.

3.5.2 Variable Importance

Variable importance for the mixfabOF model is assessed using the permutation-based VIM introduced in Section 3.3.2. (Figure 3.4) ranks the predictors based on this measure. The variable Item pair variable dominates by a wide margin (VIM $= 0.380$), followed at a substantial

[3]The CLMM interaction model achieved imperfect convergence (convergence code 4: "Stopped by zero step from line search") with a maximum gradient of 8.698×10^{-3}, which exceeds the default tolerance (1×10^{-6}) but remains relatively small in absolute terms. Robustness checks across different model specifications (varying random effects, covariate sets, and randomly disturbed starting values) yielded highly similar parameter estimates, confirming that the convergence status reflects numerical optimization challenges rather than fundamental parameter uncertainty.

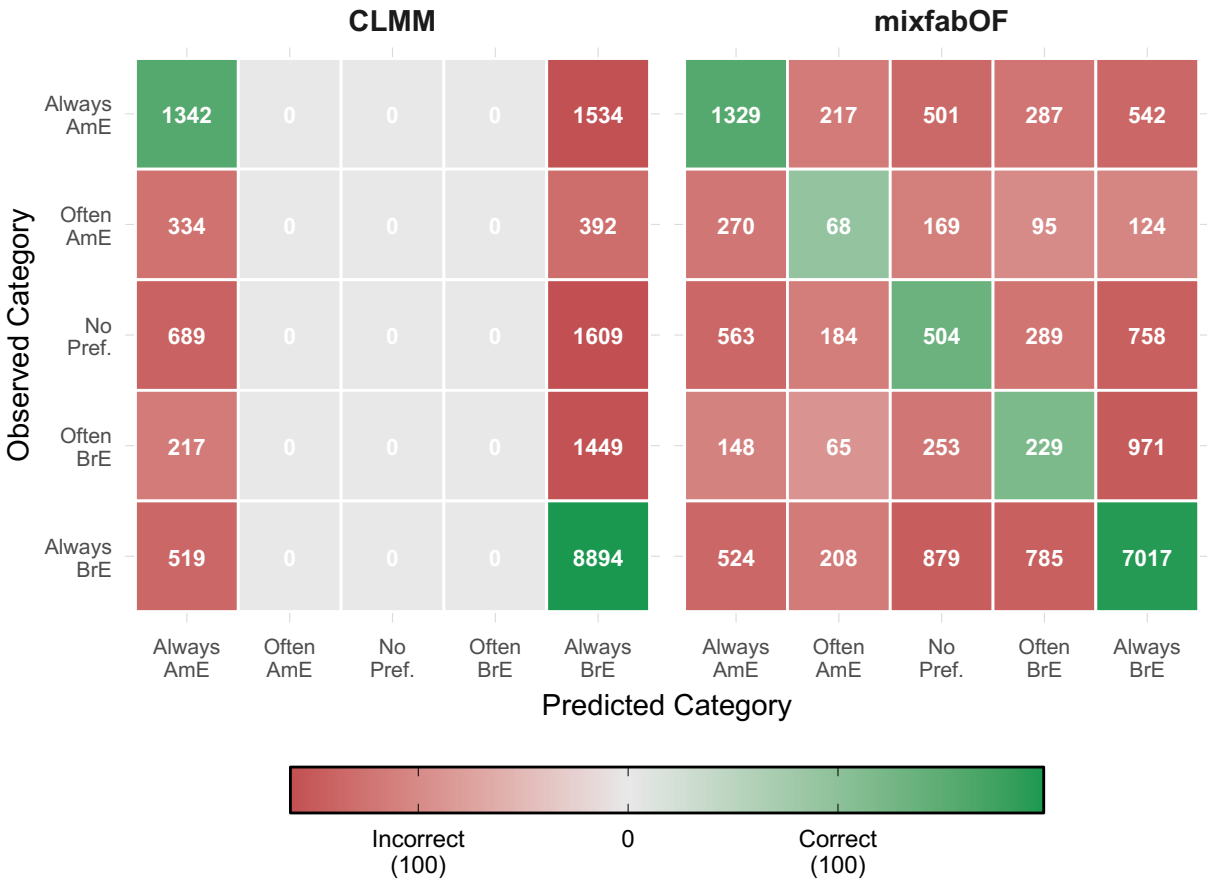


Figure 3.3: Confusion matrices comparing prediction patterns of CLMM and mixfabOF models on the test set. The CLMM predicts only categories 1 and 5, while mixfabOF produces predictions across all five ordinal categories. Cell values show the count of observations, and color intensity represents the proportion of total predictions.

distance by Date of birth (VIM $= 0.034$); Highest qualification, Year, Home language, Gender, and Ratio all have VIMs below 0.01 and make no meaningful contribution to predictive performance. Since VIM scores are marginal rather than conditional (Strobl et al., 2008), the high aggregate importance of Lexical pair can reflect both a strong population-level item effect and strong interactions between item and the other predictors. To distinguish these scenarios and to recover per-pair structure—which pairs drive the item effect and in which direction—we examine the item PDP (Figure 3.7) in the next section.

To quantify the collective importance of the variable Item pair from a second angle, we conducted an ablation study by fitting a mixfabOF model excluding variable (i.e. retaining only demographic and temporal predictors). Test set performance dropped substantially: weighted Cohen's Kappa decreased from 0.418 (standard mixfabOF) to 0.128 ($\Delta = 0.290$, a 69.5% reduction), while accuracy fell from 0.539 to 0.436 ($\Delta = 0.103$, a 19.2% reduction). The ablation and the native-factor item VIM (0.380, Figure 3.4) are complementary aggregate measures: the VIM captures how much the fitted model relies on the item variable (permuting it in the trained model collapses κ_w from 0.418 to ≈ 0.04), while the ablation captures how well demographics alone can substitute for item when the model is retrained without it ($\kappa_w = 0.128$). Both reinforce the same conclusion: item is the dominant predictor and has no effective substitute among the remaining covariates.

Among the demographic and temporal predictors, Date of birth emerges as the sole variable with non-negligible importance (VIM $= 0.034$, ranked second overall after Item pair). This finding underscores the role of apparent-time trends in the linguistic preferences, in line with earlier findings for both Maltese English (Krug and Sönning, 2018) as well as other varieties (Krug et al., 2020).

The remaining demographic and contextual variables exhibit minimal or effectively zero predictive importance. This includes Ratio, Highest qualification, Year, Gender, and Home language. The negligible importance of these variables indicates that they make no contribution to predictive performance, neither through main effects nor through interaction effects with other predictors. For example, Gender's near-zero importance implies that men and women do not differ in their overall vocabulary preferences, and further that there are no meaningful Item pair × Gender interactions: men and women do not systematically differ in their preferences for specific BrE vs. AmE variants.

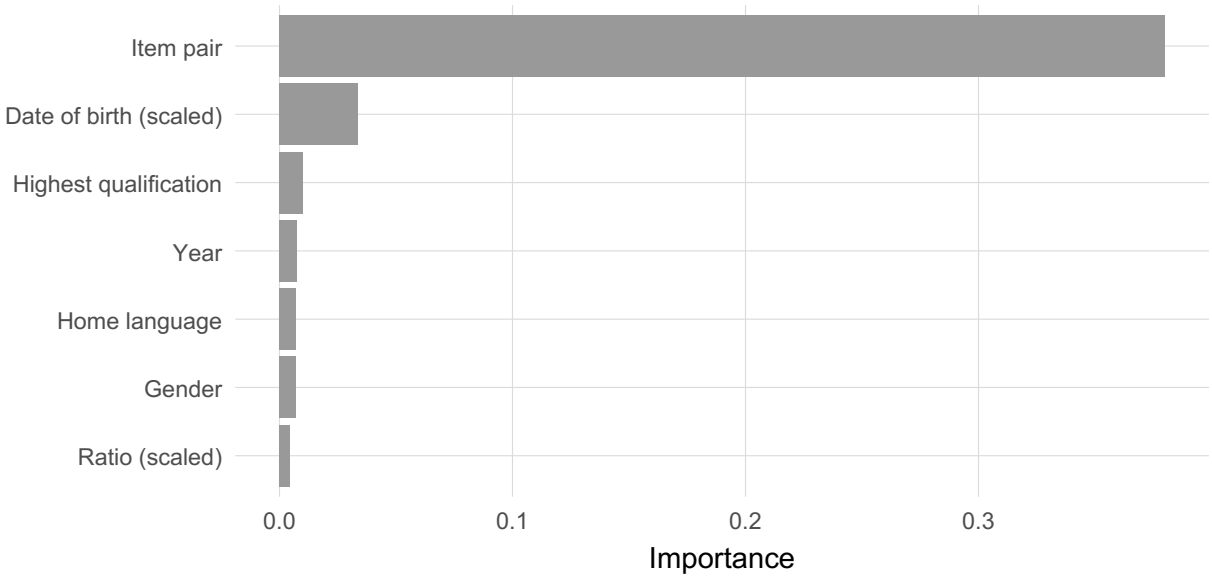


Figure 3.4: Variable importance plot for the mixfabOF model. Each predictor contributes a single bar, whether it enters the model as a native factor (Item pair, Gender, Year, Home language, Highest qualification) or as a metric variable (Date of birth, Ratio). Importance is the permutation-based decrease in weighted Cohen's Kappa when the predictor is randomly shuffled.

3.5.3 Partial Dependence

Following the VIM analysis, we now turn to PDPs to examine systematic patterns of variation in the data. As a point of departure, we will ask the mixfabOF model to predict a latent score S_{ig} for each observation in the data. In general, this is a useful first step after model fitting (see, e.g. Long and Freese, 2006, p. 207). It will also allow us to recapitulate how mixfabOFs work and how partial dependence scores are calculated.

Figure 3.5 shows the distribution of predicted scores for all 67,913 observations in the training data. The model outputs numeric predictions on a continuous latent scale (horizontal axis, range 1–5), which are then converted to ordinal categories using frequency-adjusted borders (dashed vertical lines). As discussed in Section 3.3.2, these borders are set as empirical quantiles of the OOB predictions to match the observed category frequencies: for example, the boundary between "Always AmE" and "AmE more often" sits at approximately 3.0 (not at 1.5, the midpoint between the values used for Score Assignment). The key takeaway is that the latent scale itself has no inherent meaning—what matters is where predictions fall relative to these quantile-based borders, which determine the final ordinal classification.

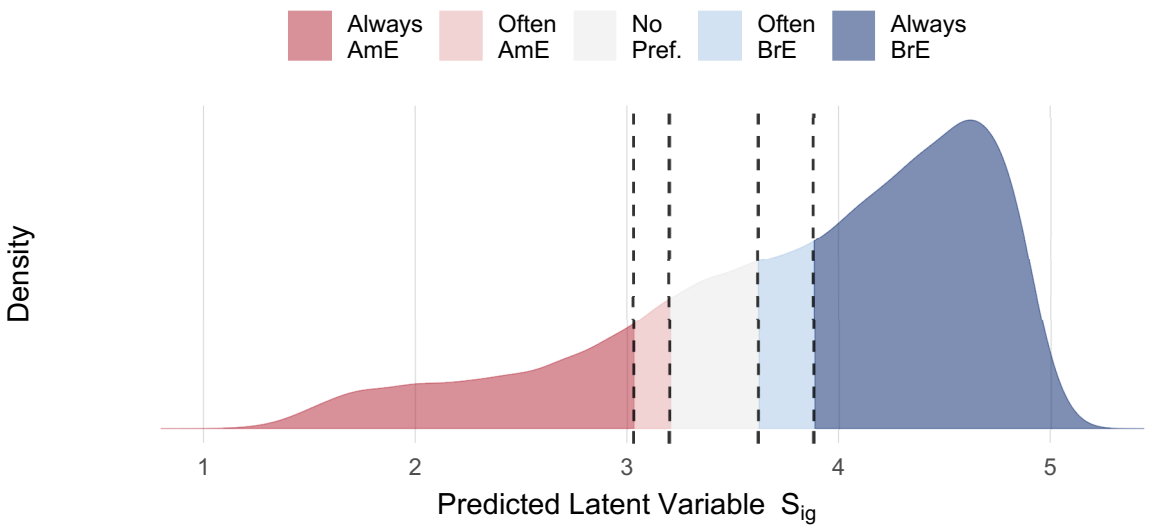


Figure 3.5: Distribution of predictions for the estimation sample, i.e. all observations in the training data.

The four borders in Figure 3.5 partition the latent scale into five regions (shown with different fill colors), one for each ordinal category. The predicted category is determined by which region it falls into: for instance, a predicted score of 3.5 lands in the grey region, corresponding to "No preference". The uneven spacing of the borders—particularly their clustering toward the center—reflects the empirical distribution in our training data, where most ratings concentrate in the outer categories "Always BrE" and "Always AmE" (see Figure 3.1). Accordingly, most predictions fall in the "Always BrE" region.

Our PDPs visualize how predicted latent scores shift when a predictor changes while holding all other variables at their observed values. For a two-level categorical predictor like Gender, we create counterfactual predictions: setting Gender to "Male" for *all* training observations yields one distribution, setting it to "Female" yields another. Figure 3.6 shows both distributions with summary overlays: filled circles represent the mean, and error bars show 66% and 95% quantile intervals. These intervals summarize the distribution of predictions (rather than average predictions or expected values) and therefore do not have the same interpretation as, say, confidence intervals. Consistent with Gender's low VIM score (Figure 3.4), the distributions largely overlap, indicating minimal systematic differences between male and female respondents. We now turn to the high-importance predictors Item pair and Date of birth. For completeness, we provide PDPs for the other variables in the Appendix.

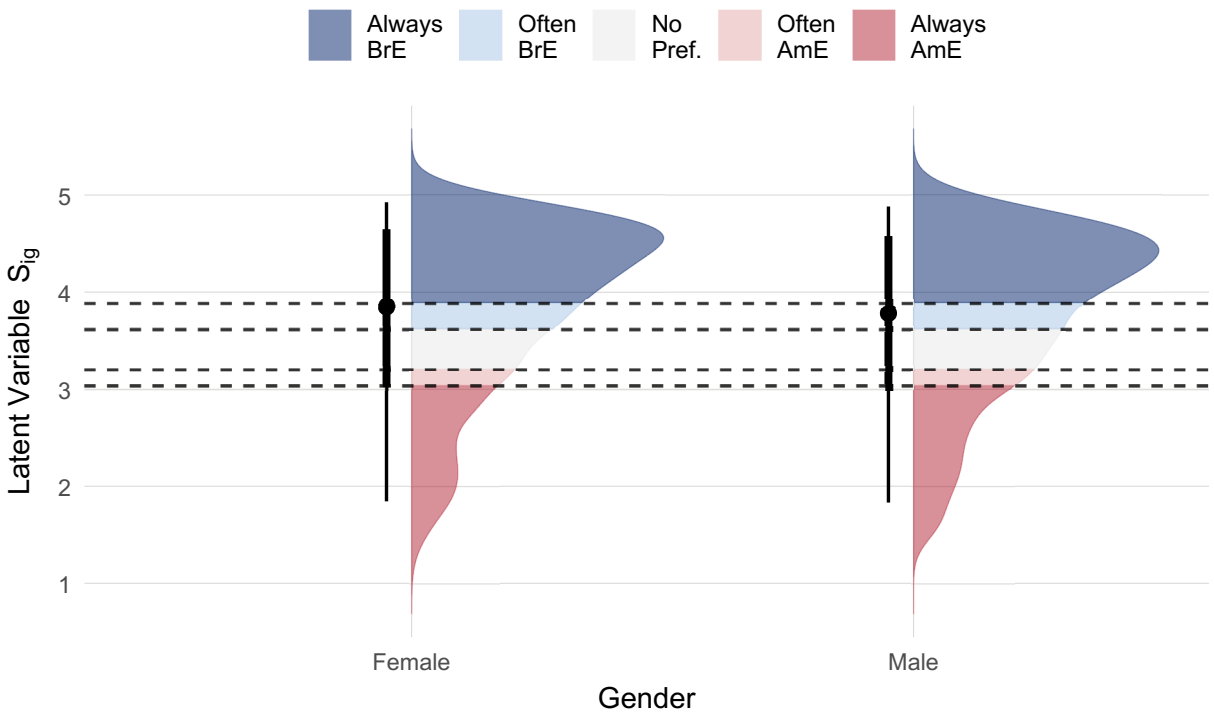


Figure 3.6: Partial dependence plot for the predictor Gender.

Figure 3.7 shows a PDP for all 63 item pairs, ordered by overall BrE preference. Each panel displays a density ridge showing the distribution of predicted latent scores when all training observations are set to that item, overlaid with summary statistics (mean and quantile intervals). At the top, items strongly favor British variants (e.g., *roundabout* over *traffic circle*, *pushchair* over *stroller*), with densities concentrated toward the "Always BrE" end of the latent scale. At the bottom, respondents prefer American variants (e.g., *truck* over *lorry*, *sports* over *sport*), with densities shifted toward "Always AmE". While substantial variation exists across items, most lean toward BrE. The density visualizations resemble the conditional probability distributions from CLMMs (see Figure 12b in Sönning et al., 2024, which plots conditional expectations), though mixfabOF's frequency-adjusted border approach does not provide a formal probabilistic framework for interpreting density areas as category probabilities. Nonetheless, the distributions reveal item-specific preference patterns and enable fine-grained comparisons across lexical pairs.

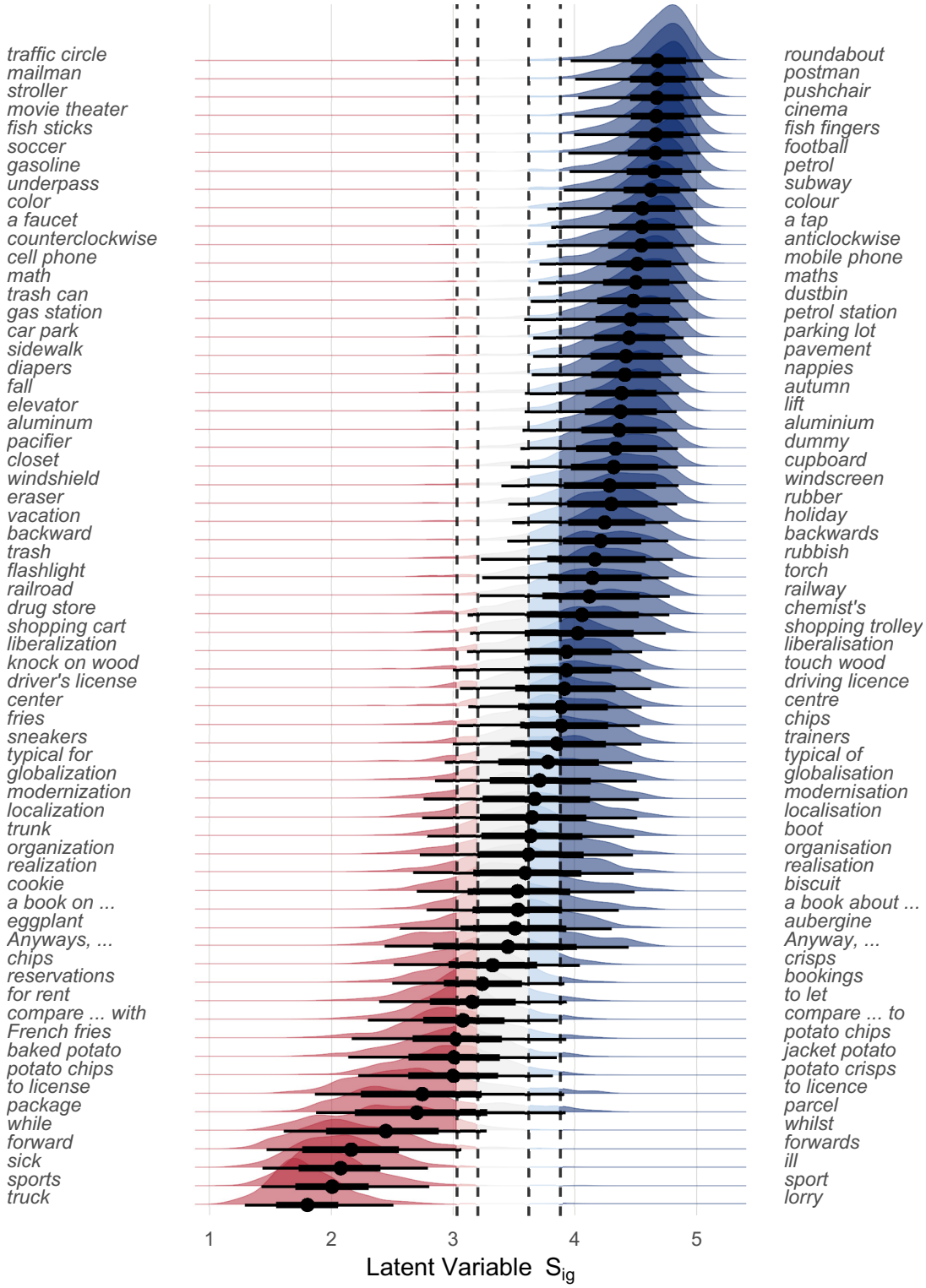


Figure 3.7: PDP for the predictor Item pair, showing density distributions and summary statistics (means and quantile intervals) for each of the 63 item pairs, ordered by overall BrE preference.

Next, we consider apparent-time trends for the individual item pairs. Since we are now interested in how model-based predictions vary with changes in *two* predictor variables (Date of

birth and Item pair), we will draw a c-PDP. For an initial overview of the breadth of variation in apparent-time trends, Figure 3.8 draws 63 profiles, one for each item pair. These show how the average model-based prediction changes with Date of birth. A number of insights emerge from Figure 3.8. First, the layering of the profiles, which stretches from the dark red (AmE) to the dark blue (BrE) region of the plot, echoes the insights afforded by Figure 3.7: Item pairs span the entire range of the scale, with a concentration at the British end. Further, there is little suggestion of linear, straight-line trends in apparent time. Rather, the profiles show noticeable non-local fluctuations. In particular, it appears that the contours are pulled apart most noticeably in the middle of the scale. Younger cohorts appear to show patterns of confluence, with items that earlier generations located firmly at either end of the scale being rated less categorically. In contrast, our CLMM constrains apparent-time trends to straight-line patterns, similar to Figure 13b in Sönning et al., 2024.

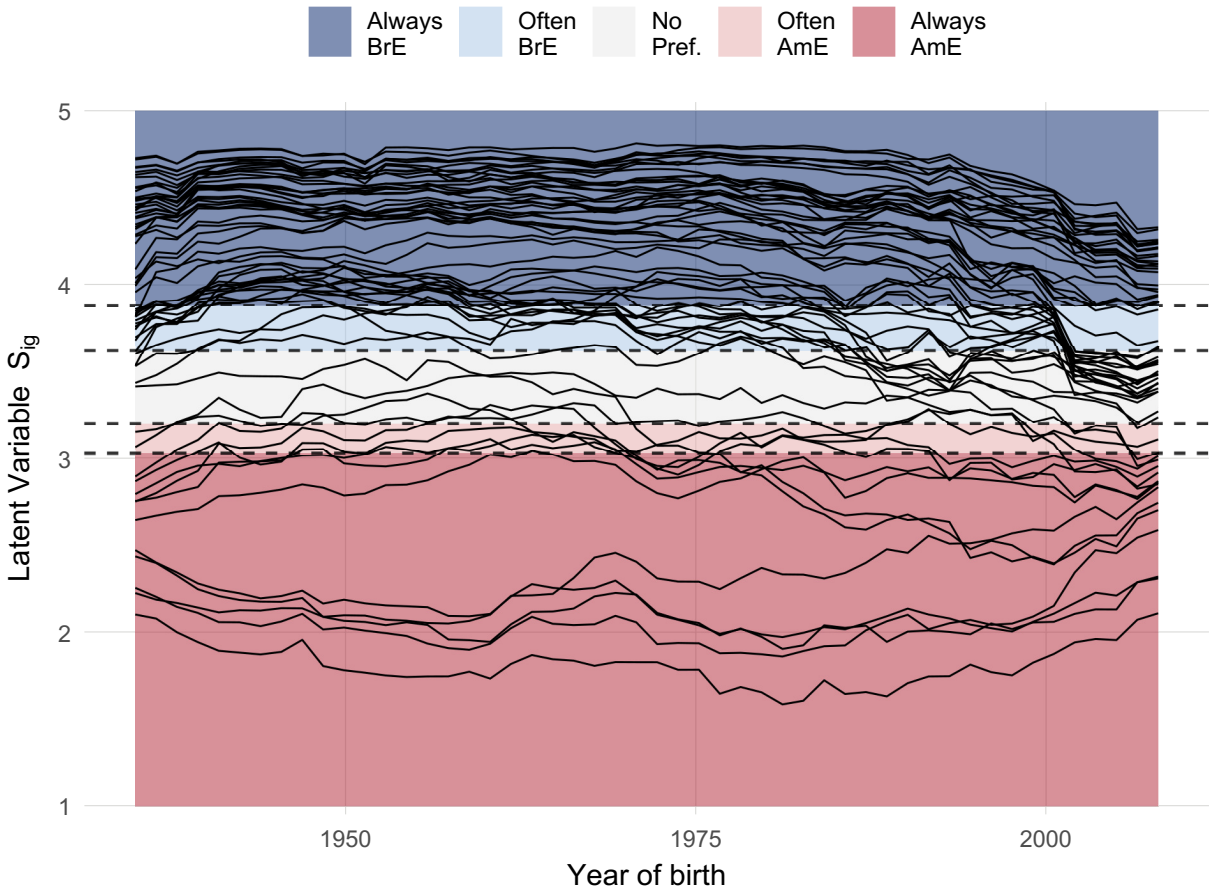


Figure 3.8: A c-PDP for the predictor Item pair, showing the average prediction for each of the 63 item pairs in the data.

One caveat is that PDPs can be misleading when data density is low; predictions should be interpreted with less confidence where observations are scarce (see Molnar, 2025). Importantly, however, the non-parametric flexibility of mixfabOF's RF component automatically discovers curved and non-monotonic trajectories without requiring the analyst to pre-specify polynomial

terms or transformation functions—a key advantage over parametric approaches that demonstrates why mixfabOF substantially outperforms CLMM on ordinal prediction performance (Table 3.2).

To demonstrate another capability of partial dependence plots, let us finally consider the apparent-time pattern for a single item pair. To allow for cross-comparisons with Sönning et al., 2024, we opt for *package–parcel*. Reference back to Figure 3.7, where it ranks 6th from the bottom, shows that it is among the "most AmE" item pairs in the pool.

Before turning to the mixfabOF c-PDP for this item pair, we first inspect the corresponding CLMM fit to make the comparison concrete. Figure 3.9 visualises item-specific apparent-time trajectories by displaying the CLMM linear predictors as a function of year of birth. These trajectories are constructed, in a manner inspired by estimated marginal means (EMMs) (Lenth, 2025), from item-specific intercepts and corresponding slopes with respect to Date of birth, both computed with continuous covariates at their means and categorical covariates averaged over. Thresholds are retained rather than averaged over, so that the plotted trajectories show the pure linear predictor $\eta_{ig} = \mathbf{x}_{ig}^{\top}\boldsymbol{\beta} + u_g$ directly, and shifts in η_{ig} across Date of birth can be read off against the fixed thresholds. Even with the Item × Date of birth interaction included in the model specification, our CLMM constrains each item's age effect to a strictly linear trajectory on the latent scale. The *package–parcel* line (purple) shows a steady decline in Britishness as respondents get younger, moving from a strong preference for the British variant "Parcel" in older cohorts ($\approx +1$ on the latent scale) to a clear preference for the American variant "Package" in the youngest cohorts (≈ -1.5).

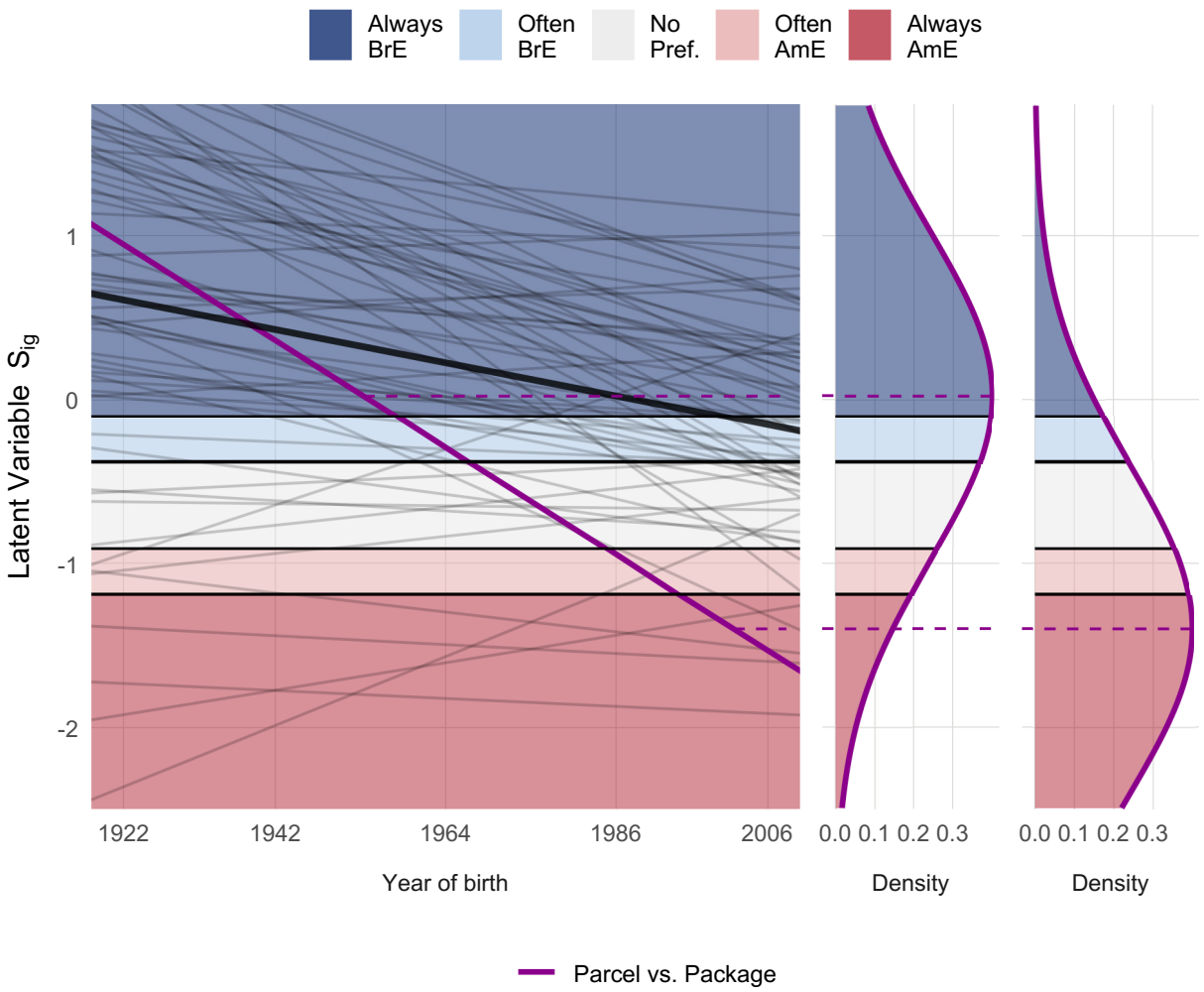


Figure 3.9: CLMM item-specific age trends and latent variable interpretation. The left panel displays linear predictors as a function of year of birth for all items (grey lines), with *package–parcel* highlighted in purple and the overall mean trend shown as a thick black line. The right panels show the conditional normal densities $S_{ig} \sim N(\eta_{ig}, 1)$ for the *package–parcel* item at two age points (an older and a younger cohort), with dashed purple lines tracing the predicted latent means across from the left panel.

For a more complete account of the rating patterns associated with this item pair, we now turn to the mixfabOF c-PDP, which shows variation in ratings across age cohorts without imposing a linear functional form. In Figure 3.10, the item pair appears in purple, and it is contrasted with the overall trend for Date of birth, which appears as a black line. The thick purple profile represents the conditional partial dependence score for *package–parcel*. The thin purple lines surrounding it are individual conditional expectation (ICE) curves, which reveal substantial heterogeneity in observation-level trajectories that the average PDP smooths over.

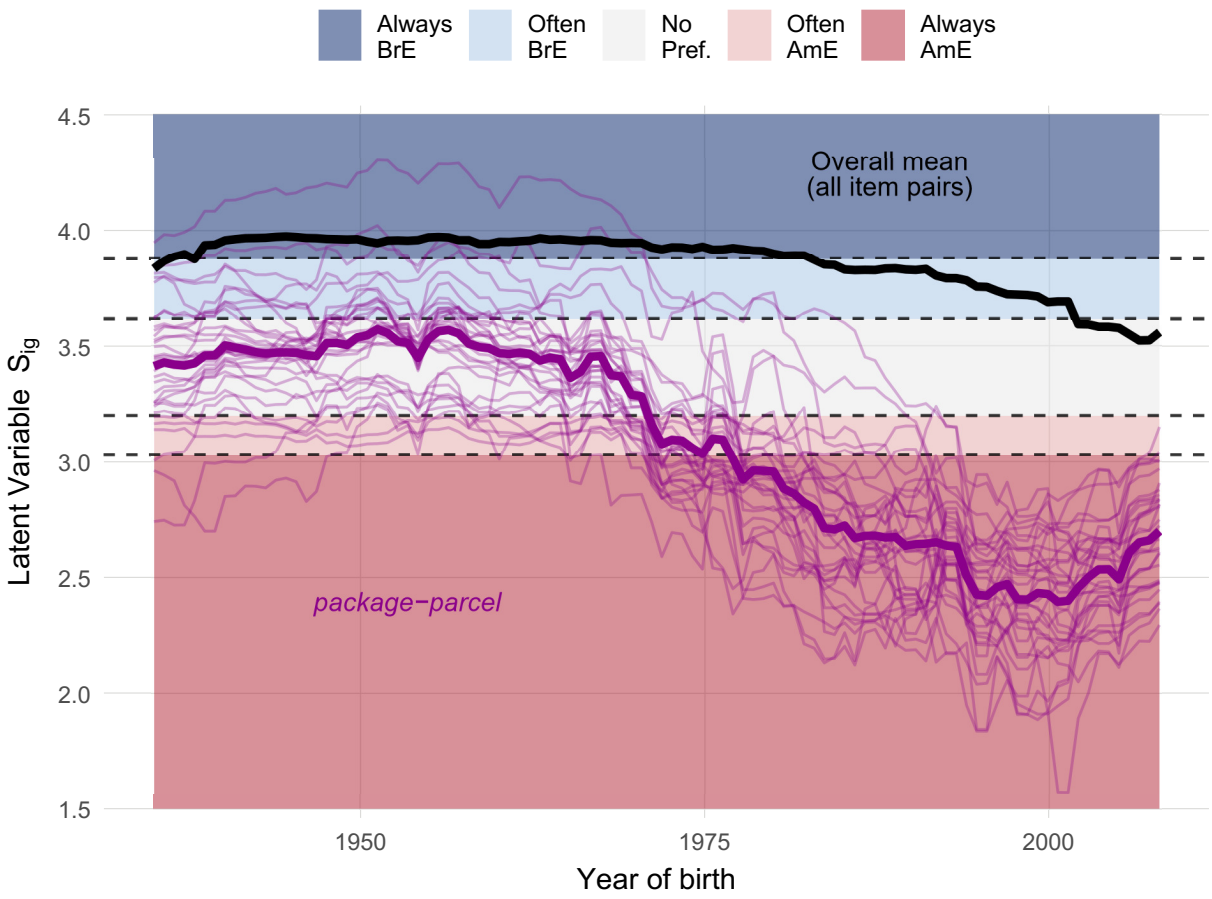


Figure 3.10: Conditional partial dependence plot for the predictor Date of birth, conditioning on a single item pair (*package–parcel*). The thin purple lines represent ICE curves, the black line the partial dependence profile for Date of birth across all item pairs.

Compared to the CLMM's straight-line trajectory in Figure 3.9, the c-PDP trace in Figure 3.10 reveals a distinct non-linear pattern. Rather than exhibiting a steady linear trend toward the AmE end of the scale, *package–parcel* hovers around the neutral midpoint of the scale until approximately 1965, and then drops quite sharply toward the AmE pole, with the youngest cohorts exhibiting a subtle recovery, with the inclination toward AmE becoming slightly less pronounced.

The PDP analysis helps explain why mixfabOF outperforms the parametric CLMM ($\kappa_w = 0.418$ vs. 0.345). The CLMM's linear constraints force it to approximate curved age trajectories with straight lines, resulting in systematic prediction errors—particularly for younger cohorts where we observe less extreme response behavior overall (see Figure 3.7). In contrast, mixfabOF's random forest component automatically discovers these non-linear patterns through recursive partitioning, effectively creating piecewise-constant approximations to arbitrary functional forms without requiring the analyst to actively allow for such patterns. This non-parametric flexibility comes at the cost of reduced interpretability relative to parametric coefficients, but the PDP framework—combined with Variable Importance metrics—provides

a powerful interpretive tool for understanding complex patterns in the fitted model.

3.6 Discussion

The aim of the current study was to contribute to the methodological discourse on the analysis of ordinal outcomes in language research. Motivated by recent endorsements of latent-variable modelling approaches to this type of response variable, we have demonstrated the utility of the frequency-adjusted borders OF approach developed by Buczak (2025b, 2025c) for the analysis of a highly structured dataset. The recent extension of this tool to clustered data structures via mixed-effects components makes it particularly appealing for language research, which often deals with clustered data structures. This is particularly true for sub-branches of the language sciences that make routine use of ordinal rating scales (e.g. rating-task experiments or questionnaire studies). Our case study has shown that this framework achieves better out-of-sample predictive accuracy than a mixed-effects cumulative regression model. A distinct advantage of RFs is their computational efficiency with large and high-dimensional datasets, their ability to reveal non-linear relationships and complex interaction patterns in the data, and their in-built safeguards against overfitting the data. These unique advantages make RFs a valuable addition to the linguistic toolkit.

The frequency-adjusted borders Ordinal Forest approach foregrounded in the present paper ties in with cumulative-link regression models by stipulating an underlying latent variable. It therefore offers similar interpretative advantages as CLMMs, which have been discussed in detail by Sönning et al. (2024). Importantly, this allows for meaningful cross-comparisons between the two modelling approaches and may direct our attention to certain oversimplifying features of linear regression approaches. In that spirit, mixfabOFs may also serve as useful adjuncts in regression modelling.

In order for these interpretative advantages to unfold, however, purpose-built software facilities are required. A particularly useful tool are partial dependence plots. This method of interpretation has so far not been available for (mix)fabOF models. We therefore implemented partial dependence routines for this class of model, which are made available to the community via GitHub. Our case study allowed us to illustrate the use and added value of these newly developed functions. Thus, while VIMs suggest the presence of systematic patterns in the data, PDPs allow us to follow up on these initial leads and uncover their nature and shape. We demonstrated how conditional PDPs can be used to look for (and appreciate) interaction patterns in the data, and how ICE curves may be used to shed light on heterogeneity that may be masked by the averaging process that is inherent to forming partial dependence scores.

From a methodological perspective, several limitations warrant acknowledgment. First and foremost, the mixfabOF framework currently only supports a single grouping variable. Since

our case study showed an explicit interest in the behavior of item pairs, it was appropriate to treat the variable Item pair as a fixed effect. In many other settings that make use of ordinal rating-scale data, items are instead considered (and therefore statistically treated as) a random effect, a typical example being an acceptability-judgment study. For the moment, this limits the wider utility of mixfabOFs for linguistic research. It would therefore be desirable for future work within the fabOF framework to extend this scheme to crossed random effects.

Second, it should be borne in mind that random forests are suited for exploratory rather than confirmatory work. Thus, their main strength lies in their ability to handle a (potentially very) wide range of predictor variables and to detect non-linear relationships as well as complex interaction patterns. CLMMs, on the other hand, offer a statistically more adequate approach for confirmatory research – they are able to account for multiple sources of variation in the data (e.g. crossed random factors) and offer substantively informative effect sizes measures as well as indications of statistical uncertainty (e.g. confidence intervals). The utility of mixfabOFs therefore also depends on the stated goals of the analysis, and they are best considered as complementary to CLMMs.

A further limitation of mixfabOFs is their inability to incorporate survey weights, which may be necessary if the sample deviates systematically from the population of interest in ways that affect substantive conclusions. Extending mixfabOF to accommodate survey weights would enable principled inference to target populations and is an important methodological priority. This extension would likely require modifying both the RF fitting procedure (to use weighted sampling) and the border estimation heuristic (to respect weighted quantiles).

Finally, the VIMs we employ are marginal rather than conditional, meaning they reflect the overall predictive contribution of each variable including any interaction effects involving that variable. The marginal importance thus conflates main effects with higher-order interactions, and does not isolate the unique contribution of a variable's main effect alone. Conditional variable importance methods (e.g., Debeer & Strobl, 2020; Strobl et al., 2008) could provide finer-grained insights by isolating main effects, but adapting these methods to the mixfabOF framework would require methodological research beyond the scope of this study. We note further, following Debeer and Strobl (2020), that marginal and conditional ("partial") variable importance are best read as two extremes on one continuum rather than as competing definitions of a single true importance, with the choice of perspective depending on the research question.

We hope to have illustrated the added value of (mix)fabOFs for ordinal language data analysis. In terms of model interpretation, its latent-variable grounding puts it on a par with cumulative-link regression models, as it likewise provides informative and nuanced insights into patterns in the data. At the same time, the underlying recursive-partitioning scheme yields several advantages that allow for flexible and effective data analysis. For these reasons, we would argue that the (mix)fabOF framework is a serious contender for a permanent place in the toolkit of

quantitative linguist(ic)s.

Declarations

Funding and/or Conflicts of interest/Competing interests

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG; grant number: 548274092). The authors declare no conflicts of interest.

Code Availability Statement

The code used in this study is available upon request from the corresponding author.

Data Statement

The dataset used in the current study is available on Zenodo (Krug et al., 2025).

B Appendix

	I always use this expression	I use this expression more often	I have no preference	I use this expression more often	I always use this expression		I never use either expression	Explanation / Comment
a drop in the ocean	○	○	○	○	○	a drop in the bucket	○	
a faucet	○	○	○	○	○	a tap	○	
aluminum	○	○	○	○	○	aluminium	○	
anticlockwise	○	○	○	○	○	counterclockwise	○	
eggplant	○	○	○	○	○	aubergine	○	*(fruit/vegetable)*
fall	○	○	○	○	○	autumn	○	*(season of the year)*
backward	○	○	○	○	○	backwards	○	
bicentenary	○	○	○	○	○	bicentennial	○	
cookie	○	○	○	○	○	biscuit	○	*(something sweet to eat)*
bookings	○	○	○	○	○	reservations	○	

Figure B.1: Excerpt from questionnaire. The full version is available on Zenodo (Krug et al., 2025).

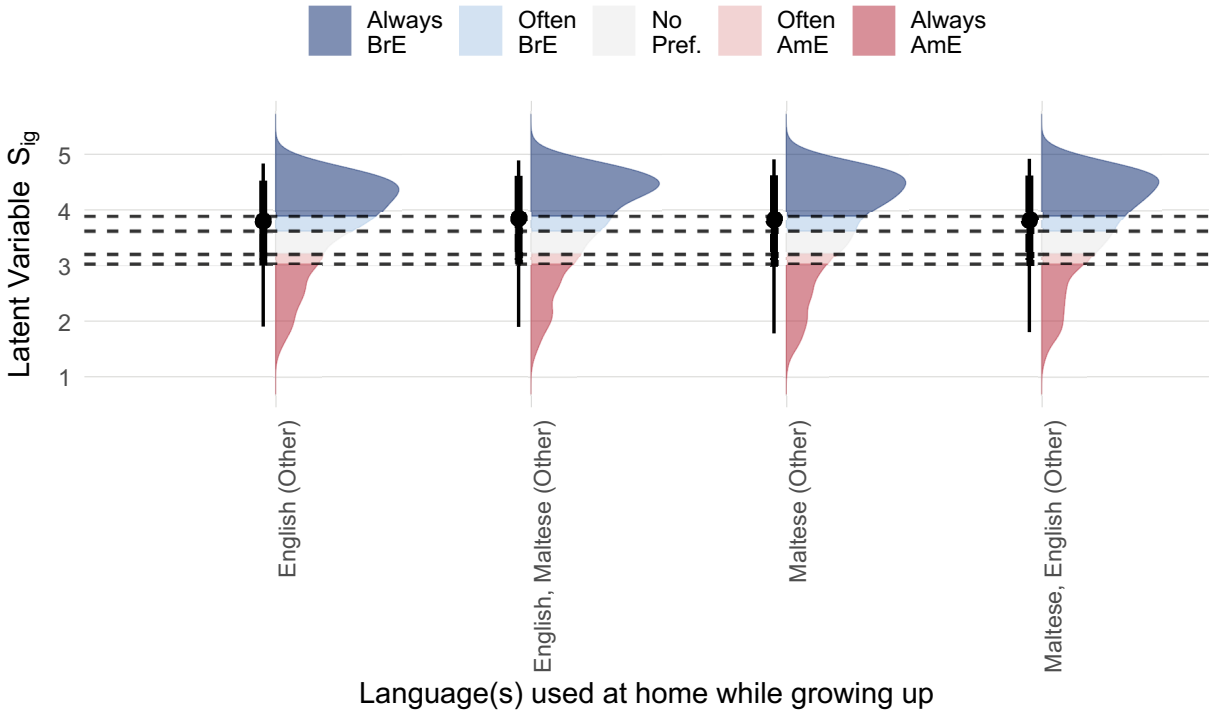


Figure B.2: Partial dependence plot for the predictor Language(s) used at home while growing up.

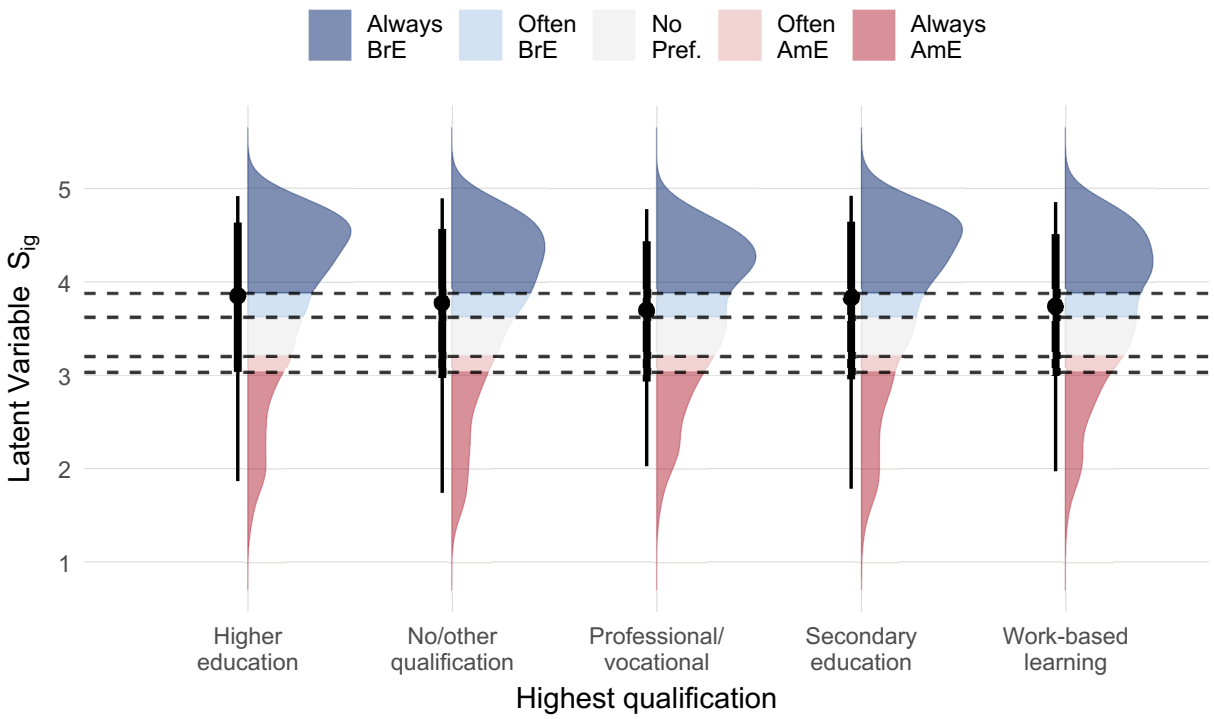


Figure B.3: Partial dependence plot for the predictor Highest qualification.

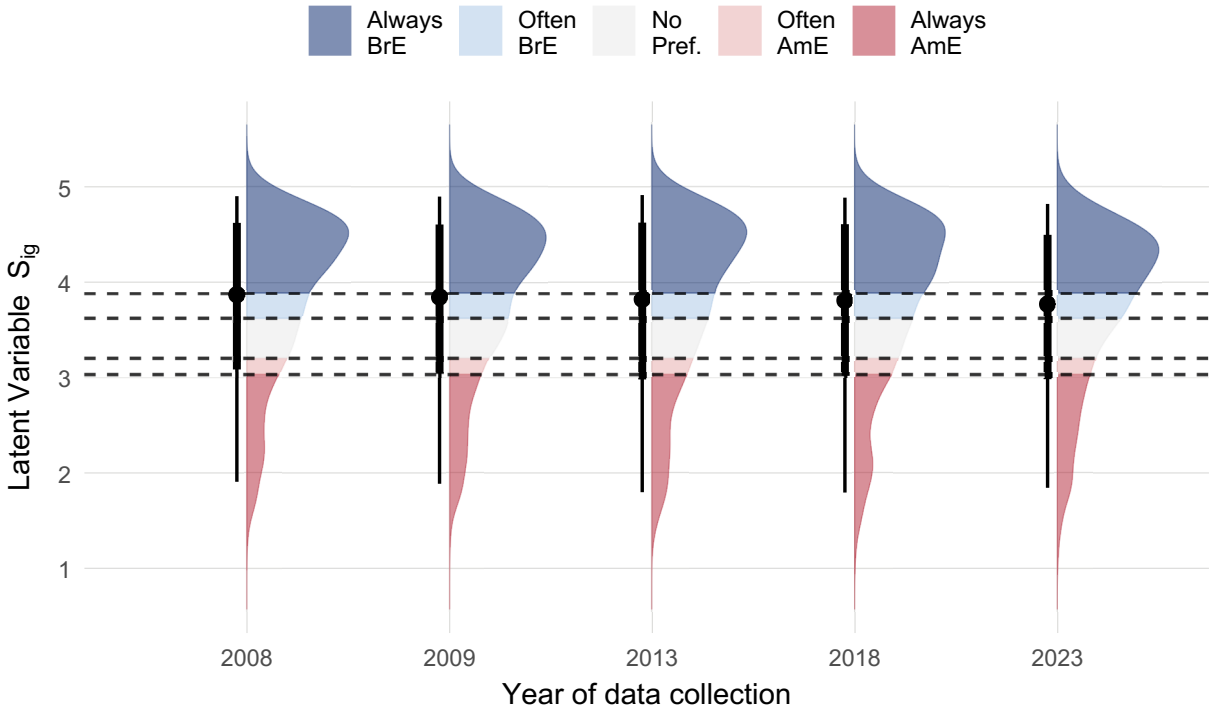


Figure B.4: Partial dependence plot for the predictor Year of data collection.

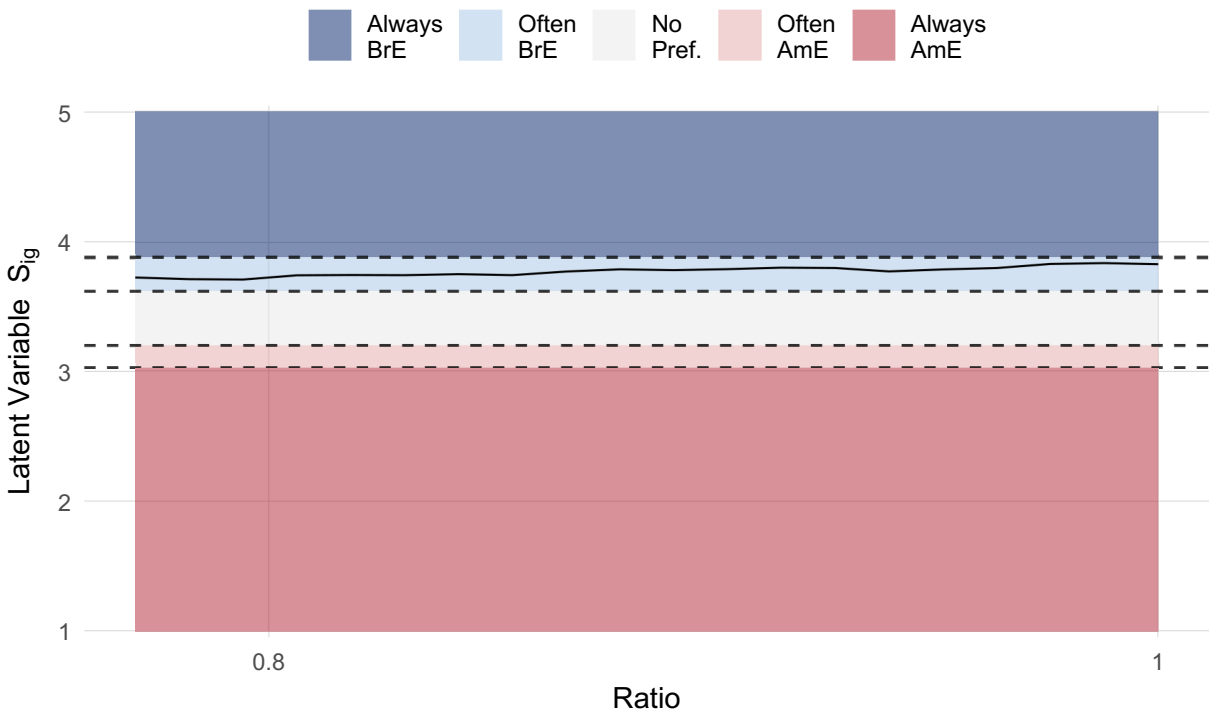


Figure B.5: Partial dependence plot for the predictor Ratio (proportion of life spent in Malta).

Bibliography

Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley. https://doi.org/10.1002/9780470594001

Agresti, A., & Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review / Revue Internationale de Statistique*, *69*(3), 345–371. https://doi.org/10.2307/1403450

Alfons, A., & Templ, M. (2012). Estimation of social exclusion indicators from complex surveys: The R package laeken. https://doi.org/10.2139/ssrn.2244876

Algeo, J. (2006). *British or American English?: A handbook of word and grammar patterns*. Cambridge University Press. https://doi.org/10.1017/cbo9780511607240

Al-Khasawneh, M. B., & Cirillo, C. (2024). Using small area estimation to produce reliable transportation statistics: The case of household trips estimation at the census tract level. *Data Science for Transportation*, *6*(3), 21. https://doi.org/10.1007/s42421-024-00105-1

Anjoy, P. (2023). Hierarchical Bayes measurement error small area model for estimation of disaggregated level workers mobility pattern in India. *Journal of Quantitative Economics*, *21*(2), 339–361. https://doi.org/10.1007/s40953-023-00338-x

Aubry, P., Guillemain, M., Jensen, G. H., Sorrenti, M., & Scallan, D. (2020). Moving from intentions to actions for collecting hunting bag statistics at the European scale: Some methodological insights. *European Journal of Wildlife Research*, *66*(4). https://doi.org/10.1007/s10344-020-01400-2

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36. https://doi.org/10.1080/01621459.1988.10478561

Bäumer, M., Hautzinger, H., & Pfeiffer, M. (2018). Small-Area-Methoden zur Schätzung von Verkehrskennzahlen in kleinräumiger Gliederung. Retrieved June 19, 2023, from https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Bericht_Regionalisierung_MiD-Ergebnisse_Small_Area-Verfahren_1218.pdf

Bäumer, M., & Pfeiffer, M. (2025). *Mobilitätskennzahlen für Deutschland in kleinräumiger Gliederung auf der Basis von modellgestützten Small Area-Schätzungen* (Technical Report). infas, DLR, IVT Research, infas 360. Bonn. https://www.mobilitaet-in-deutschland.de

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press. https://doi.org/10.1515/9781400874668

Bergonzoli, G., Rossi, L., & Masci, C. (2024). Ordinal mixed-effects random forest. https://doi.org/10.48550/arXiv.2406.03130

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, *41*(3), 1111–1141. https://doi.org/10.1214/13-AOS1096

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale de Statistique*, *51*(3), 279–292. https://doi.org/10.2307/1402588

Bock, A. (2020). Lepus europaeus (Lagomorpha: Leporidae). *Mammalian Species*, *52*(997), 125–142. https://doi.org/10.1093/mspecies/seaa010

Brandão, F. J. S., Correia, R. M., & Paio, A. (2018). Measuring urban renewal: A dual kernel density estimation to assess the intensity of building renovation—case study in Lisbon. *Urban Science*, *2*(3), 91. https://doi.org/10.3390/urbansci2030091

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buczak, P. (2025a). FabOF: Frequency adjusted borders ordinal forest for ordinal prediction. https://github.com/phibuc/fabOF

Buczak, P. (2025b). Frequency-adjusted borders ordinal forest: A novel tree ensemble method for ordinal prediction. *British Journal of Mathematical and Statistical Psychology*, *78*(2), 594–616. https://doi.org/10.1111/bmsp.12375

Buczak, P. (2025c). Mixed-effects frequency-adjusted borders ordinal forest: A tree ensemble method for ordinal prediction with hierarchical data. *Multivariate Behavioral Research*, *60*(6), 1237–1252. https://doi.org/10.1080/00273171.2025.2547416

Buczak, P., Horn, D., & Pauly, M. (2025). Old but gold or new and shiny? comparing tree ensembles for ordinal prediction with a classic parametric approach. *Journal of Classification*, *42*(2), 364–390. https://doi.org/10.1007/s00357-024-09497-9

Buhl, Y., Kovacevic, Z., Le Jeune, D., Nguyen, L., & Ritter, J. (2025). *restatis: R wrapper to access a wide range of Germany's federal statistical system databases based on the GENESIS web service RESTful API of the German federal statistical office (Statistisches Bundesamt/Destatis)*. manual. https://doi.org/10.32614/CRAN.package.restatis

Bundesinstitut für Bau-, Stadt- und Raumforschung. (2025). INKAR – Indikatoren und Karten zur Raum- und Stadtentwicklung. Retrieved December 1, 2025, from https://www.inkar.de/

Burgard, J. P., Krause, J., & Kreber, D. (2019). *Regularized area-level modelling for robust small area estimation in the presence of unknown covariate measurement errors* (Research Papers in Economics No. 4/19). Universität Trier, Fachbereich IV - Volkswirtschaftslehre. Trier. https://hdl.handle.net/10419/207036

Burgard, J. P., Krause, J., Kreber, D., & Morales, D. (2021). The generalized equivalence of regularization and min–max robustification in linear mixed models. *Statistical Papers*, *62*(6), 2857–2883. https://doi.org/10.1007/s00362-020-01214-z

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Celeux, G. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly*, *2*, 73–82.

Celeux, G., Chauveau, D., & Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of statistical computation and simulation*, *55*(4), 287–314. https://doi.org/10.1080/00949659608811772

Christensen, R. H. B. (2023). *Cumulative link models for ordinal regression with the R package ordinal* (tech. rep.). CRAN. https://CRAN.R-Project.org/package=ordinal

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage. https://us.sagepub.com/en-us/nam/experimental-syntax/book6067

Dammann-Tamke, H. (2025). *DJV Handbuch 2025*. Deutscher Jagdverband. https://www.grube.de/p/djv-handbuch-2025/P76-381-2025/

Debeer, D., & Strobl, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics*, *21*(1), 307. https://doi.org/10.1186/s12859-020-03622-2

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Destatis. (2024a). Bevölkerungszahlen in Gitterzellen. https://www.zensus2022.de/static/Zensus_Veroeffentlichung/Zensus2022_Bevoelkerungszahl.zip

Destatis. (2024b). Personen: Bevölkerungszahl, Code: 1000A-0000. https://ergebnisse.zensus2022.de/datenbank/online/url/dc1a89b7

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.2307/2290268

Didan, K., & Barreto-Muñoz, A. (2019). *MODIS collection 6.1 (C61) vegetation index product user guide* (tech. rep.). The University of Arizona, Vegetation Index and Phenology Lab: Tucson, AZ, USA. Retrieved March 26, 2025, from https://lpdaac.usgs.gov/documents/621/MOD13_User_Guide_V61.pdf

Didan, K. (2021). MODIS/Terra vegetation indices 16-day l3 global 500m SIN grid v061. https://doi.org/10.5067/MODIS/MOD13A1.061

Dollinger, S. (2015). *The written questionnaire in social dialectology: History, theory, practice*. John Benjamins Publishing Company. https://doi.org/10.1075/impact.40

Dumelle, M., Higham, M., Ver Hoef, J. M., Olsen, A. R., & Madsen, L. (2022). A comparison of design-based and model-based approaches for finite population spatial sampling and inference. *Methods in Ecology and Evolution*, *13*(9), 2018–2029. https://doi.org/10.1111/2041-210x.13919

Eggs, J., Follmer, R., Gruschwitz, D., Nobis, C., Bäumer, M., & Pfeiffer, M. (2018). *Mobilität in Deutschland - MiD: Methodenbericht* (tech. rep.). Infas, DLR, IVT and infas 360. https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Methodenbericht.pdf

Elashoff, M., & Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, *13*(1), 48–65. https://doi.org/10.1198/1061860043092

Elvidge, C. D., Baugh, K. E., Dietz, J. B., Bland, T., Sutton, P. C., & Kroehl, H. W. (1999). Radiance calibration of DMSP-OLS low-light imaging data of human settlements. *Remote Sensing of Environment*, *68*(1), 77–88. https://doi.org/10.1016/s0034-4257(98)00098-4

Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F.-C., & Taneja, J. (2021). Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, *13*(922). https://doi.org/10.3390/rs13050922

Endresen, A., & Janda, L. A. (2017). Five statistical models for Likert-type experimental data on acceptability judgments. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 217–250. https://doi.org/10.1558/jrds.30822

Erfurth, K., Groß, M., Rendtel, U., & Schmid, T. (2022). Kernel density smoothing of composite spatial data on administrative area level: A case study of voting data in Berlin. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *16*(1), 25–49. https://doi.org/10.1007/s11943-021-00298-9

European Commission. (2020). *Energy, transport and environment statistics—2020 edition* (tech. rep.). Publications Office of the European Union Luxembourg. https://ec.europa.eu/eurostat/web/products-statistical-books/-/ks-dk-20-001

Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Fay, R. E., & 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*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Federal Agency for Cartography and Geodesy. (2024). Verwaltungsgebiete 1:250 000. https://sgx.geodatenzentrum.de/web_public/gdz/datenquellen/Datenquellen_vg_nuts.pdf

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, *222*, 309–368. https://doi.org/10.1098/rsta.1922.0009

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications. https://collegepublishing.sagepub.com/products/applied-regression-analysis-and-generalized-linear-models-3-237254

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fullerton, A. S., & Xu, J. (2016). *Ordered regression models: Parallel, partial, and non-parallel alternatives*. Chapman and Hall/CRC. https://doi.org/10.1201/b20060

Geenens, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, *109*(505), 346–358. https://doi.org/10.1080/01621459.2013.842173

Ghosh, M., & Rao, J. N. (1994). Small area estimation: An appraisal. *Statistical Science*, *9*(1), 55–76.

Gill, P. E., Murray, W., & Wright, M. H. (2019). *Practical optimization*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975604

Gill, R. D., van der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In D. Y. Lin & T. R. Fleming (Eds.), *Proceedings of the first seattle symposium in biostatistics* (pp. 255–294). Springer US.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gomes, D. G. E. (2022). Should i use fixed effects or random effects when i have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, *10*, e12794. https://doi.org/10.7717/peerj.12794

Gräber, R., Strauß, E., Rölfing, F., & Johanshon, S. (2024). *Wild und Jagd: Landesjagdbericht* (tech. rep.). Lower Saxony Ministry of Agriculture. https://www.ml.niedersachsen.de/download/211767/Landesjagdbericht_2023_2024.pdf

Grafmiller, J. (2023). Visualizing grammatical similarities in comparative variationist analysis. In L. Sönning & O. Schützler (Eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. VARIENG. https://urn.fi/URN:NBN:fi:varieng:series-22-4

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gries, S. T. (2021). *Statistics for linguistics with R*. De Gruyter Mouton. https://doi.org/10.1515/9783110718256

Gril, L., Steinkemper, L., Groß, M., & Rendtel, U. (2025). Kernel heaping - kernel density estimation from regional aggregates via measurement error model. *The R Journal*, *16*(3), 115–133. https://doi.org/10.32614/RJ-2024-026

Groß, M., & Fuchs, L. (2022). Kernelheaping: Kernel density estimation for heaped and rounded data. https://doi.org/10.32614/cran.package.kernelheaping

Groß, M., Kreutzmann, A.-K., Rendtel, U., Schmid, T., & Tzavidis, N. (2020). Switching between different non-hierachical administrative areas via simulated geo-coordinates: A case study for student residents in Berlin. *Journal of Official Statistics*, *36*(2), 297–314. https://doi.org/10.2478/jos-2020-0016

Groß, M., & Rendtel, U. (2016). Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology*, *4*(3), 339–361. https://doi.org/10.1093/jssam/smw011

Groß, M., Rendtel, U., Schmid, T., Schmon, S., & Tzavidis, N. (2017). Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(1), 161–183. https://doi.org/10.1111/rssa.12179

Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one". In V. Godambe & D. Sprott (Eds.), *Foundations of statistical inference*. Holt, Rinehart and Winston.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Havet, P. (2002). Besoins de recherche et orientations d'action resultant de l'analyse de l'enquête sur les tableaux de chasse en France (1998–1999). *Zeitschrift für Jagdwissenschaft*, *48*(S1), 222–235. https://doi.org/10.1007/bf02192412

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, *81*(4), 701. https://doi.org/10.2307/2337073

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, *19*(4), 2244–2253. https://doi.org/10.1214/aos/1176348396

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. *Geographical Analysis*, *28*(3), 244–261. https://doi.org/10.1111/j.1538-4632.1996.tb00933.x

Hornung, R. (2019). Ordinal forests. *Journal of Classification*, *37*(1), 4–17. https://doi.org/10.1007/s00357-018-9302-x

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.2307/2280784

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., & Ferreira, L. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Infas. (2023). Mobilität in deutschland, webpage. Retrieved May 25, 2023, from https://www.mobilitaet-in-deutschland.de/archive/index.html

Infas, DLR, & IVT. (2019). Mid 2017. https://daten.clearingstelle-verkehr.de/279/

Infas, DLR, IVT, & Infas 360. (2025). *Mobilität in Deutschland – MiD Ergebnisbericht* (Technical Report No. FE-Nr. VB600001). Bundesministerium für Verkehr. Bonn, Berlin. https://www.mobilitaet-in-deutschland.de/pdf/MiD2023_Ergebnisbericht.pdf

Janitza, S., Celik, E., & Boulesteix, A.-L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*(4), 885–915. https://doi.org/10.1007/s11634-016-0276-4

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, *15*(1), 1–96. https://doi.org/10.1007/BF02595419

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, *39*(2). https://doi.org/10.1097/00006199-199003000-00019

Kreutzmann, A.-K., Marek, P., Runge, M., Nicola, S., & Schmid, T. (2022). The Fay–Herriot model for multiply imputed data with an application to regional wealth estimation in Germany. *Journal of Applied Statistics*, *49*(13), 3278–3299. https://doi.org/10.1080/02664763.2021.1941805

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, *91*(7), 1–33. https://doi.org/10.18637/jss.v091.i07

Krug, M., Schützler, O., & Werner, V. (2016). Patterns of linguistic globalization: Integrating typological profiles and questionnaire data. In O. Timofeeva, A.-C. Gardner, A. Honkapohja, & S. Chevalier (Eds.), *New approaches to English linguistics* (pp. 35–66). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.177.03kru

Krug, M., Schützler, O., & Werner, V. (2020). How British is Gibraltar English? In P. Núñez-Pertejo, M. J. López-Couso, B. Méndez-Naya, & J. Pérez-Guerra (Eds.), *Crossing linguistic boundaries: Systemic, synchronic and diachronic variation in English* (pp. 153–186). Bloomsbury. https://doi.org/10.20378/irb-48103

Krug, M., & Sell, K. (2013). Designing and conducting interviews and questionnaires. In *Research methods in language variation and change* (pp. 69–98). Cambridge University Press. https://doi.org/10.1017/cbo9780511792519.007

Krug, M., & Sönning, L. (2018). Language change in Maltese English: The influence of age and parental languages. In P. Paggio & A. Gatt (Eds.), *The languages of Malta* (pp. 247–270). Language Science Press. https://doi.org/10.5281/zenodo.1181801

Krug, M., Vetter, F., & Sönning, L. (2025). Background data for: Ordinal random forests in language data analysis. https://doi.org/10.5281/zenodo.17762245

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces*, *46*(2), 151–160. https://doi.org/10.2307/2574595

Lahiri, P., & Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, *77*(2), 312–320. https://doi.org/10.1007/s13571-015-0096-0

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, Y., Rojas-Perilla, N., Runge, M., & Schmid, T. (2023). Variable selection using conditional aic for linear mixed models with data-driven transformations. *Statistics and Computing*, *33*(1), 27. https://doi.org/10.1007/s11222-022-10198-9

Lenth, R. V. (2025). *Emmeans: Estimated marginal means, aka least-squares means*. manual. https://CRAN.R-project.org/package=emmeans

Levshina, N. (2020). Conditional inference trees and random forests. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 611–643). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_25

Li, S., Cai, T. T., & Li, H. (2022). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association*, *117*(540), 1835–1846. https://doi.org/10.1080/01621459.2021.1888740

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lin, Y.-P., Chu, H.-J., Wu, C.-F., Chang, T.-K., & Chen, C.-Y. (2010). Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques. *International Journal of Environmental Research and Public Health*, *8*(1), 75–88. https://doi.org/10.3390/ijerph8010075

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata, 2nd edition* (2nd ed.). StataCorp LLC. https://EconPapers.repec.org/RePEc:tsj:spbook:long2

Long, L., Lin, J., & Pu, W. (2009). Model-based synthesis of household travel survey data in small and midsize metropolitan areas. *Transportation Research Record*, *2105*(1), 64–70. https://doi.org/10.3141/2105-09

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(8), 1–19. https://doi.org/10.18637/jss.v009.i08

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. John Wiley. https://doi.org/10.1002/9780470580066

Marbach, M. (2026). *Bonn: Access inkar database*. manual. https://github.com/sumtxt/bonn

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, *38*(3), 1035–1051. https://doi.org/10.3233/SJI-210902

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *42*(2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

McGuckin, N., & Fucci, A. (2018). *Summary of travel trends: 2017 National Household Travel Survey* (tech. rep. No. FHWA-PL-18-019). Federal Highway Administration. Washington DC. https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons. https://doi.org/10.1002/9780470191613

Molnar, C. (2025). *Interpretable machine learning. a guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book

Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer International Publishing. https://doi.org/10.1007/978-3-030-63757-6

Mühlbauer, M. (2026). Estimating disaggregated mobility indicators using area-level models on survey data.

Mühlbauer, M., & Schmid, T. (2025). Density estimation from aggregated data with integrated auxiliary information: Estimating population densities with geospatial data. https://doi.org/10.48550/arXiv.2508.03610

Mühlbauer, M., & Sönning, L. (2025). Ordinal random forests in language data analysis. https://doi.org/10.31234/osf.io/s4bv2_v1

Neves, A., do Nascimento Silva, D. B., & Correa, S. (2014). Small domain estimation for the Brazilian service sector survey. *Estadistica. Revista Semestral del Instituto Interamericano de Estadistica*, *65*(185), 13–37. https://eprints.soton.ac.uk/362740/

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *6*(3), 457. https://doi.org/10.2307/3318671

Nobis, C., & Köhler, K. (2018). *Mobilität in Deutschland- MiD: Nutzerhandbuch* (tech. rep.). Infas, DLR, IVT and infas 360. https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Nutzerhandbuch.pdf

Nobis, C., & Kuhnimhof, T. (2018). *Mobilität in Deutschland- MiD: Ergebnisbericht* (tech. rep.). Infas, DLR, IVT and infas 360. https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Ergebnisbericht.pdf

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Peterson, R. A., & Cavanaugh, J. E. (2022). Ranked sparsity: A cogent regularization framework for selecting and estimating feature interactions and polynomials. *AStA Advances in Statistical Analysis*, *106*(3), 427–454. https://doi.org/10.1007/s10182-021-00431-7

Pratesi, M. (Ed.). (2016, January). *Analysis of poverty data by small area estimation*. Wiley. https://doi.org/10.1002/9781118814963

R Core Team. (2026). R: A language and environment for statistical computing. https://www.R-project.org/

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118735855

Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241. https://doi.org/10.1080/01621459.1988.10478591

Rao, J. N. K., Wu, C. F. J., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey methodology*, *18*(2), 209–217. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199200214486

Ren, W., Li, J., Erciulescu, A., Krenzke, T., & Mohadjer, L. (2022). A variable selection method for small area estimation modeling of the proficiency of adult competency. *Stats*, *5*(3), 689–713. https://doi.org/10.3390/stats5030041

Rendtel, U., Neudecker, A., & Fuchs, L. (2021). Ein neues Web-basiertes Verfahren zur Darstellung der Corona-Inzidenzen in Raum und Zeit. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *15*(2), 93–106. https://doi.org/10.1007/s11943-021-00288-x

Rendtel, U., & Ruhanen, M. (2018). Die Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarfs an Kinderbetreuung in Berlin. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *12*(3–4), 271–284. https://doi.org/10.1007/s11943-018-0235-y

Rieck, T., Steffen, A., Schmid-Küpke, N., Feig, M., Wichmann, O., & Siedler, A. (2020). Impfquoten bei Erwachsenen in Deutschland – Aktuelles aus der KV-Impfsurveillance und der Onlinebefragung von Krankenhauspersonal OKaPII. *Epidemiologisches Bulletin*, (47), 3–26. https://doi.org/10.25646/7658

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007368

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148. https://doi.org/10.1111/rssa.12488

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Salon, D. (2016). Estimating pedestrian and cyclist activity at the neighborhood scale. *Journal of Transport Geography*, *55*, 11–21. https://doi.org/10.1016/j.jtrangeo.2016.06.023

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer. https://link.springer.com/book/9780387406206#bibliographic-information

Schai-Braun, S. C., Weber, D., & Hackländer, K. (2012). Spring and autumn habitat preferences of active European hares (Lepus europaeus) in an agricultural area with low hare density. *European Journal of Wildlife Research*, *59*(3), 387–397. https://doi.org/10.1007/s10344-012-0684-5

Schütze, C. T., & Sprouse, J. (2014). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). Cambridge University Press. https://doi.org/10.1017/cbo9781139013734.004

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silverman, B. (2018, February). *Density estimation for statistics and data analysis*. Routledge. https://doi.org/10.1201/9781315140919

Sliwinski, K., Ronnenberg, K., Jung, K., Strauß, E., & Siebert, U. (2019). Habitat requirements of the European brown hare (Lepus europaeus Pallas 1778) in an intensively used agriculture region (Lower Saxony, Germany). *BMC Ecology*, *19*(1). https://doi.org/10.1186/s12898-019-0247-7

Slud, E. V., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 239–257. https://doi.org/10.1111/j.1467-9868.2006.00542.x

Smith, R. K., Vaughan Jennings, N., & Harris, S. (2005). A quantitative analysis of the abundance and demography of European hares Lepus europaeus in relation to habitat type,

intensity of agriculture and climate. *Mammal Review*, *35*(1), 1–24. https://doi.org/10.1111/j.1365-2907.2005.00057.x

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Sönning, L. (2024). Ordinal response scales: Psychometric grounding for design and analysis. *Research Methods in Applied Linguistics*, *3*(3), 100156. https://doi.org/10.1016/j.rmal.2024.100156

Sönning, L., Krug, M., Vetter, F., Schmid, T., Leucht, A., & Messer, P. (2024). Latent-variable modelling of ordinal outcomes in language data analysis. *Journal of Quantitative Linguistics*, *31*(2), 77–106. https://doi.org/10.1080/09296174.2024.2329448

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data*, *2*(1), 150045. https://doi.org/10.1038/sdata.2015.45

Statistische Ämter des Bundes und der Länder. (2025). Regionaldatenbank Deutschland. Retrieved December 1, 2025, from https://www.regionalstatistik.de/genesis/online

Statistisches Bundesamt. (2025). GENESIS-Online: Die Datenbank des Statistischen Bundesamtes. Retrieved December 1, 2025, from https://www-genesis.destatis.de/genesis/online

Stead, D. (2001). Relationships between land use, socioeconomic factors, and travel patterns in Britain. *Environment and Planning B: Planning and Design*, *28*(4), 499–528. https://doi.org/10.1068/b2677

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, *10*(2), 1–22. https://doi.org/10.1371/journal.pone.0107042

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Rothacher, Y., Theiler, S., & Henninger, M. (2026). Detecting interactions with random forests: A comment on gries' words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, *22*(1), 1–26. https://doi.org/10.1515/cllt-2024-0028

Sutton, P., Roberts, D., Elvidge, C., & Baugh, K. (2001). Census from heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, *22*(16), 3061–3076. https://doi.org/10.1080/01431160010007015

Sutton, P. (1997). Modeling population density with night-time satellite imagery and gis. *Computers, Environment and Urban Systems*, *21*(3–4), 227–244. https://doi.org/10.1016/s0198-9715(97)01005-3

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*(1), 1–31. https://doi.org/10.18637/jss.v106.i01

Thüringer Landtag. (2019, October). Gesetz zur freiwilligen Neugliederung des Landkreises Wartburgkreis und der kreisfreien Stadt Eisenach, zur Änderung der Thüringer Kommunalordnung und zur Änderung des Thüringer Finanzausgleichsgesetzes sowie zur Anpassung gerichtsorganisatorischer Vorschriften Vom 16. Oktober 2019. https://parldok.thueringer-landtag.de/parldok/dokument/72572/gesetz_und_verordnungsblatt_nr_12_2019.pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Torabi, M., & Rao, J. N. K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, *127*, 36–55. https://doi.org/10.1016/j.jmva.2014.02.001

Tutz, G. (2021). Ordinal trees and random forests: Score-free recursive partitioning and improved ensembles. *Journal of Classification*, *39*(2), 241–263. https://doi.org/10.1007/s00357-021-09406-4

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society A*, *181*, 927–979. https://doi.org/10.1111/rssa.12364

Vaish, A. K., Chen, S., Sathe, N. S., Folsom, R. E., Chandhok, P., & Guo, K. (2010). Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey. *Transportation*, *37*(6), 825–848. https://doi.org/10.1007/s11116-010-9279-8

Wang, B., & Wang, X. (2007). Bandwidth selection for weighted kernel density estimation. *arXiv e-prints*, arXiv:0709.1616. https://doi.org/10.48550/arXiv.0709.1616

Wei, G. C. G., & Tanner, M. A. (1990). Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704. https://doi.org/10.2307/2290005

Wieditz, J., Miller, C., Scholand, J., & Nemeth, M. (2024). A brief introduction on latent variable based ordinal regression models with an application to survey data. *Statistics in Medicine*, *43*(29), 5618–5634. https://doi.org/10.1002/sim.10208

Winter, B., & Grice, M. (2021). Independence and generalizability in linguistics. *Linguistics*, *59*(5), 1251–1277. https://doi.org/10.1515/ling-2019-0049

Worcester, R. M., & Burns, T. R. (1975). A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society*, *17*(3), 181–197.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Wu, B., Yang, C., Wu, Q., Wang, C., Wu, J., & Yu, B. (2023). A building volume adjusted nighttime light index for characterizing the relationship between urban population and nighttime light intensity. *Computers, Environment and Urban Systems*, *99*, 101911. https://doi.org/10.1016/j.compenvurbsys.2022.101911

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*(1). https://doi.org/10.1214/aos/1176346060

Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, *34*(3), 275–285. https://doi.org/10.1002/gepi.20459

Yang, Z., He, S., Zhang, H., Li, M., & Liang, Y. (2022). Spatiotemporal analysis of gastrointestinal tumor (GI) with kernel density estimation (KDE) based on heterogeneous background. *International Journal of Environmental Research and Public Health*, *19*(13), 7751. https://doi.org/10.3390/ijerph19137751

Ybarra, L. M. R., & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*(4), 919–931. https://doi.org/10.1093/biomet/asn048