Small Area Estimation of Unemployment Rates Using Big Data

Integration of Labor Force Surveys with Geospatial Data

Author: **Joseph Oluoch (Date of Birth – 960605)**

Semester:

Spring 2025

Course name, level and credits:

Independent Project II, Second Cycle, 15 Credits

Subject:

Statistics

Supervisor:

Thomas Laitila, Professor of Statistics

Examiner:

Ann-Marie Flygare, PhD

Örebro University School of Business

Abstract

This paper explores the integration of satellite-derived geospatial data and administrative registers into a Small Area Estimation (SAE) framework to generate precise unemployment rate estimates at the county level in Sweden. Using the Fay–Herriot (FH) model combined with Empirical Best Linear Unbiased Prediction (EBLUP), the study assesses how effectively selected geospatial covariates such as nighttime lights, elevation, urban cover, NDVI, and temperature enhance the precision of unemployment estimates compared to traditional survey based estimates alone. Empirical results show that incorporating geospatial indicators reduces the mean squared error of estimates, confirming that satellite-derived data significantly enhances estimation accuracy. Model diagnostics support the validity of the FH approach in integrating geospatial covariates. The study also critically evaluates the practical considerations, including computational complexity and the potential limitations inherent to Big Data integration.

Keywords: Small Area Estimation, Fay–Herriot Model, Unemployment Rates, Geospatial Data, Official Statistics, Empirical Best Linear Unbiased Prediction, Big Data.

Preface

This thesis marks the culmination of my Master's studies in Statistics, with a specialization in Statistical Modeling and Data Science, at Örebro University. My interest in the production and communication of official statistics, especially at more granular geographical levels, led me to explore the use of satellite-derived geospatial data in enhancing traditional survey-based unemployment estimates.

The idea for this research emerged from my participation in the European Big Data Hackathon 2025 organized by Eurostat, where I got a first interaction on a larger scale with geospatial data. I then developed an interest in the potential of small area estimation techniques. Working with data from the Swedish Labour Force Survey and remote sensing sources has allowed me to bridge theoretical knowledge with applied statistical modeling in a meaningful way.

Conducting this research has been both intellectually rewarding and personally challenging. From grappling with model assumptions and diagnostics to processing and harmonizing large scale geospatial indicators, the process has deepened my appreciation for the complexities of producing reliable statistics that inform policy.

This thesis is intended for students, researchers, and practitioners interested in small area estimation, labour market statistics, or the integration of geospatial data in official statistics. It also serves as a modest contribution to the ongoing discussions about enhancing the relevance and precision of subnational indicators through data innovation.

Joseph Nyajuoga

Örebro City, June 2025

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Professor Thomas Laitila for his invaluable guidance throughout the development of my thesis. His support has been instrumental in shaping my research and enhanced my understanding of small area estimation. I am also greatful to the Örebro School of Business for allowing me to pursue my interest in official statistics. I thank my examiner, Anne-Marie Flygare for her expert recommendations and thorough scrutiny of my work. Her insights and expertise in survey sampling have significantly enhanced the quality of this thesis. I would like to thank my lecturers and fellow classmates for their cooperation, support, and engaging discussions throughout my academic journey.

Special thanks to the Swedish Institute for funding my studies in Sweden; without their financial support, this endeavor would not have been possible. Finally, heartfelt thanks to my family and friends for their support and encouragement.

Table of contents

Table of contents **IV**

List of Figures **V**

List of Tables **VI**

List of Abbreviations **VII**

1 Introduction **1**

- 1.1 Background & Motivation . . . 1
- 1.2 Literature Review . . . 2
- 1.3 Contribution . . . 4
- 1.4 Thesis Outline . . . 4

2 Methodology **5**

- 2.1 The FH Model Framework . . . 5
  - 2.1.1 Model Specifications . . . 5
  - 2.1.2 Role of Auxiliary Variables . . . 7
- 2.2 Parameter Estimation . . . 7
  - 2.2.1 Estimation of Regression Coefficients (β) . . . 7
  - 2.2.2 Estimation of Random Effects Variance (σ_u^2) . . . 8
- 2.3 Prediction: EBLUP . . . 8
- 2.4 Transformations . . . 9
- 2.5 Mean Squared Prediction Error (MSPE) Estimation . . . 10
  - 2.5.1 Analytical MSPE Approximation (Prasad-Rao) . . . 10
  - 2.5.2 Bootstrap MSPE Estimation . . . 11
- 2.6 Model Selection and Diagnostics . . . 12
  - 2.6.1 Model Selection Criteria . . . 12
  - 2.6.2 Model Diagnostics . . . 12

3 Data **13**

- 3.1 Direct Estimates . . . 13
- 3.2 Auxiliary Data . . . 13
  - 3.2.1 Geospatial Covariates . . . 13
  - 3.2.2 Extra Covariates . . . 13
- 3.3 Data Concerns . . . 14

4 Empirical Results and Analysis **15**
4.1 Software Setup 15
4.2 Preliminary Analysis 15
  4.2.1 Examining Linear Relationships 15
4.3 Modeling 16
4.4 Model Diagnostic Checks 18
4.5 Comparison of Direct Estimates vs FH Estimates 19
4.6 Discussions 21

5 Conclusions & Recommendations **23**
5.1 Conclusions 23
5.2 Limitations 23
5.3 Recommendations 23

References **24**

Appendix **26**
Appendix A: Visual of the Direct Estimates by Swedish Counties 26
Appendix B: Direct Vs Model-based Estimates 27
Appendix C: Brown's Correlation Tests 28

List of Figures

4.1 Correlation Matrix . . . 15
4.2 Residual Plots of the Reduced Original Scale Model . . . 18
4.3 MSE Comparison Plots of the Model Estimates vs Direct Estimates . . . 19
4.4 CV Comparison Plots of the Model Estimates vs Direct Estimates . . . 20
4.5 ScatterPlot of the Model Estimates vs Direct Estimates . . . 20
4.6 CV Comparison of the Direct vs Model-based Estimates . . . 21

5.1 Direct Estimates by County . . . 26
5.2 Maps of the Direct Estimators vs FH-based estimators . . . 27

List of Tables

3.1 Geospatial Covariates 14

4.1 Resulting Models 16
4.2 Residuals Tests for the Reduced Original Scale Model 18

5.1 Direct vs FH Estimates (Reduced Original Scale Model) 27
5.2 Brown's Test for the Reduced Original Scale Model 28

List of Abbreviations

Abbreviation	Definition
FH	Fay-Herriot
EBLUP	Empirical Best Linear Unbiased Prediction
SAE	Small Area Estimation
NSI	National Statistical Institute
LFS	Labor Force Survey
GIS	Geographical Information System
ONS	Office for National Statistics
StatsCAN	Statistics Canada
SCB	Statistics Sweden
LST	Land Surface Temperature
NO_2	Nitrogen dioxide
SDG	Sustainable Development Goals

Chapter 1

Introduction

1.1 Background & Motivation

Reliable labor market statistics at fine geographic scales are crucial for informed policy-making and regional planning. Local unemployment rates can vary markedly from the national estimates, reflecting regional economic structures and disparities. The huge labor force surveys (LFS)[1] tend to produce accurate estimates at the national or at broad regional levels but less precise estimates for small sub-regions. As a result, when one attempts to estimate unemployment for smaller domains like counties or municipalities, the direct survey estimates are often unreliable or unavailable due to insufficient sample sizes. This poses a significant challenge for agencies and stakeholders who need granular labor market data.

Granular labor market indicators are effective tools for policy development, resource allocation, and the evaluation of initiatives at the local level. Statistics Canada (2024) recently highlighted that improving the precision of sub-provincial employment estimates via modeling supports better policy planning in smaller communities and rural areas. Similarly, in the UK, the Office for National Statistics (ONS) moved to a model-based approach for local unemployment estimates when it became clear that the direct survey estimates for many districts were too unreliable for publication (Chambers, Salvati and Tzavidis, 2016). These are clear demonstrations of the trend in stakeholders growing interest in the labor statistics at the disaggregated levels that address the geographic inequalities and the impacts of local labor programs.

Small Area Estimation (SAE) has emerged as a powerful solution to this problem. SAE techniques "borrow strength" from larger areas or auxiliary information by using statistical models, thereby enabling more precise and reliable estimates for small domains even when direct survey data in those domains are sparse[2]. By integrating survey data with auxiliary covariates, SAE can substantially reduce the estimation error for counties or other smaller domains. The fundamental premise is that a well-specified model with good covariates can predict the "true" domain value more accurately than the survey alone, especially for domains with small samples. SAE has become a standard tool in the NSIs toolkit to meet needs for disaggregated statistics. Eurostat and other NSIs have published guidelines and case studies promoting SAE methods for regional labor market indicators (European Commission, 2019a; Statistics Canada, 2024).

The traditional sources of auxiliary information like the administrative data and registers have been the ideal sources of auxiliary information used when carrying out SAE modeling. However, recently the proliferation of the Big Data[3] sources has brought the attention to exploring the use of this readily available data as a possible source of the auxiliary information. Specifically, with the official statistics community there have been commendable efforts to address this new found mine field. Researchers and national statistical agencies have made use of Big Data in implementation of poverty mapping and poverty estimation methodologies (Marchetti *et al.*, 2015). To keep up with the growth of Big Data, the processing power of computers has also been beefed up and now the geospatial data as one of the sources of Big Data is available at fine resolutions. Geospatial data through the use of satellite-derived measures of night-time luminosity, indicators of accessibility like the travel times to urban centers, and land use are now increasingly available at a fine spatial resolution and can proxy the economic activity at the disaggregated levels (Mellander *et al.*, 2015). Such indicators, when appropriately incorporated, can help explain spatial variation in labor market outcomes that surveys alone cannot capture. Recent methodological guidance by The United Nations Statistics Division encourages leveraging geospatial data in SAE to enhance the precision and relevance of small-area estimates[4].

---

[1] See more about LFS - EU labour force survey - Statistics Explained

[2] More background from Eurostat's Statistics Explained City Statistics - small area estimation

[3] European Parliament describes Big Data as large datasets that are large and complex and may require newer technologies such as AI to process - Source link.

[4] The guide can be accessed here Small Area Estimation with Geospatial Data: A Primer.

Building up on the SAE context, this study will integrate both geospatial and administrative auxiliary information in a model-based framework to estimate county-level unemployment rates in Sweden. These auxiliary inputs are expected to be correlated with local unemployment rates, and by incorporating them into the estimation model, we leverage additional information beyond the survey itself to explain between-county differences. The fundamental premise is that a well-specified model with useful covariates can predict the "true" unemployment rate in each county more accurately than the noisy survey alone, especially for counties where the survey sample is small. If successful, this approach will yield estimates that are more reliable at disaggregated geographic levels.

1.2 Literature Review

The seminal work of Fay and Herriot (1979) introduced the classic area-level SAE model, originally to estimate per-capita income for small places in the US. The FH model treats direct survey estimates for each area as observations in a hierarchical model, combining them with area-level auxiliary variables through random effects. This approach was a breakthrough in producing reliable estimates for subpopulations, and it laid the groundwork for widespread adoption of SAE. Since then, comprehensive overviews of SAE methods have been provided with theoretical advances and practical applications. Comprehensive treatments are given by Rao (2003) and Rao and Molina (2015), who detail both the theoretical advances and practical applications since FH. Two broad classes of models are distinguished: area-level models (like FH) which use aggregated outcomes and covariates for each area, and unit-level models (like the Battese–Harter–Fuller model) which operate on micro-data. An early unit-level application by Battese, Harter and Fuller (1988) used a mixed model to predict county crop yields, incorporating satellite-based land cover data as covariates. In official statistics, area-level models remain especially popular because researchers often have access only to published area aggregates and reliable auxiliary totals by area, rather than unit records.

A variety of extensions to the basic SAE models have been proposed to handle different data structures and to improve estimation. For estimating proportions or counts, researchers have extended generalized linear mixed models to the small-area context. Malec *et al.* (1997) developed a Bayesian hierarchical model for small-area proportions in a health survey, and Nandram and Choi (2002) applied Bayesian methods to estimate domain-level binary outcomes accounting for survey weights and non-response. These early Bayesian approaches demonstrated how logistic or other non-linear link functions could be combined with random area effects to model labor force status. Empirical best prediction (EBP) techniques were introduced for small-area proportions under generalized linear mixed models by Pfeffermann *et al.* (1998), paralleling the empirical best linear unbiased prediction (EBLUP) used for continuous variables. The frequentist vs. Bayesian dichotomy in SAE has since evolved into a complementary set of tools with empirical Bayes (EBLUP/EBP) providing point estimates and mean squared error estimates under assumed models, and fully Bayesian methods enabling flexibility in model hierarchies and straightforward uncertainty propagation at the cost of intensive computation, both approaches are widely used in practice. For instance, You and Rao (2003) employed a pseudo–Bayesian (EB) time-series modeling for Canadian unemployment rates.

Socio-economic indicators often exhibit correlations over time and space, researchers have extended SAE models to exploit these dimensions. Rao and Yu (1994) Rao proposed combining time-series and cross-sectional data in an area-level model, introducing auto-correlated random effects to borrow strength from past periods. Their approach was applied to U.S. state poverty estimates and later to Canadian labor force data (You and Rao, 2003), showing significant gains in stability for month-to-month unemployment rates. Subsequent work have refined these ideas, Brakel and Krieg (2015) developed integrated multilevel time-series models for official labor force statistics in the Netherlands. Spatial extensions introduce correlation among area effects to utilize geographic information. Besag, York and Mollié (1991) were early proponents of spatial random effects in regional estimation; more recently Singh, Shukla and Kundu (2005) proposed an EBLUP that includes a spatially correlated area effect. A general finding is that spatio-temporal SAE models can further reduce estimation error by borrowing information across multiple dimensions, though they require careful modeling to avoid bias.

With the expanding toolkit of SAE models, there has been increasing emphasis on model diagnostics and validation. As with any model-based method, SAE estimates rely on assumptions and violations can lead to biased or unreliable estimates. Researchers therefore recommend routine diagnostics such as checking residuals and comparing model predictions to direct estimates for signs of lack-of-fit (Rao and

Molina, 2015, chap.7). Techniques like cross-validation on held-out areas, or comparing different models via AIC/BIC, are used for model selection in SAE applications. Another important consideration in official statistics is benchmarking, adjusting model-based small-area estimates so that they aggregate to known totals at higher levels. Benchmarking ensures consistency with official national figures and is often required for official releases. Methods for benchmarking SAE estimates include ratio adjustments or model-based approaches that constrain the sum of estimates. These adjustments typically inflate small-area estimates slightly to remove any discrepancy with the reliable larger-area totals, trading off a minor increase in variance for adherence to published national figures. Modern software implementations like the *emdi*[5] package in R facilitate such calibration of SAE results. Current best practice in SAE involves not only fitting sophisticated models but also validating them and aligning them with known aggregates to maintain credibility for official use.

SAE methods have been increasingly adopted by national and international statistical agencies to improve the granularity of official data. Statistics Canada has invested in SAE for labor force and other surveys in 2024 it released experimental small area estimates of employment and unemployment for metropolitan areas and newly defined self-contained labor areas, using SAE to augment the LFS sample (Statistics Canada, 2024). Other NSIs, such as Statistics New Zealand, Statistics Italy (ISTAT) and Statistics Spain (INE), have applied SAE for indicators ranging from poverty rates to health outcomes, often in collaboration with academic researchers (Zhang and Bryant, 2020). Eurostat has actively facilitated this diffusion of SAE knowledge through projects like EURAREA and an ESSnet[6] on SAE, and also through publishing manuals on guidelines on SAE to help NSIs produce city and regional indicators with modeling techniques (European Commission, 2019b). In the domain of labor statistics, several academic-government partnerships in Europe have demonstrated SAE's utility. Ugarte *et al.* (2009) analyzed regional unemployment in Navarre, Spain, comparing design-based and model-based estimators; their study for a Spanish regional government led to recommendations on the best performing methods given available auxiliary data. In Italy, SAE methods have been tested for provincial employment rates, and the German Federal Statistical Office has explored SAE for district-level unemployment as well. These efforts collectively show that SAE is transitioning from research to practice in official statistics, addressing the common problem of disseminating reliable small-domain estimates without enlarging surveys.

Recent trend in the SAE literature is the incorporation of Big Data or alternative data as auxiliary information. The motivation is that traditional sources like censuses and administrative registers may not fully capture real-time or granular variation in socio-economic conditions, whereas new data streams could provide additional predictive power. Hadam *et al.* (2024) integrate mobile phone network data into a FH model to estimate unemployment for Functional Urban Areas in Germany. By using aggregated mobile phone activity as covariates, they improved estimates of unemployment in urban centers versus hinterlands, especially when considering alternative unemployment definitions based on workplace. This study illustrated how high-frequency big data can complement surveys by reflecting mobility and local economic activity. In poverty mapping, efforts have been made to use satellite imagery and machine learning to predict poverty in areas with sparse survey data (Jean *et al.*, 2016; Steele *et al.*, 2017), essentially as a form of SAE with very rich auxiliary data. Within official statistics, these approaches are at an experimental stage, but there is growing evidence of their value: for instance, the World Bank and UN Global Pulse have piloted combining survey microdata with Earth Observation (EO) indicators to improve estimates of welfare and population at the local level (Newhouse *et al.*, 2022). Specifically focusing on geospatial auxiliary data, researchers have begun evaluating indicators like nighttime lights, NDVI (vegetation index), land cover types, climate and pollution measures as predictors in SAE models. These variables can proxy the level of infrastructure development, economic intensity, or environmental factors that correlate with employment. Early findings are encouraging. Besides the previously mentioned result that satellite-measured night lights correlate with economic activity (Mellander *et al.*, 2015), studies have shown that vegetation and agricultural indices can help explain rural poverty or employment in agrarian regions (Tang, Liu and Matteson, 2022), and remote-sensing based accessibility metrics like travel time to cities and road network density can serve as socio-economic proxies (Weiss *et al.*, 2018). The United Nations Statistics Division (2023) recently released a primer on SAE with geospatial data that showcases an end-to-end example of using satellite-derived covariates to improve sub-national SDG[7] indicators. The literature is thus converging on the idea that leveraging geospatial technology can enhance small area statistics. However, best practices are still being established, and each application provides new insights. This thesis situates itself in this emerging body of work, by testing geospatial auxiliaries in

---

[5] The emdi package documention - (Kreutzmann *et al.*, 2019)

[6] More in the Eurostat's Innovative Projects in Statistics - ESS Innovation in Statistics

[7] More information on the SDG indicators as listed by the UN Statistics Division (SDG Indicators).

the context of Swedish labor force estimation an environment with high-quality survey and administrative data where the incremental benefit of satellite data can be assessed.

1.3 Contribution

This research contributes to the field of SAE and its application in official statistics in several ways. First, it provides an application and evaluation of the FH area-level model for estimating county-level unemployment rates in Sweden. While SAE methods are well-established in principle, there have been limited implementations for Swedish labor force data at fine geographic scales. By developing a model-based estimation procedure for all 21 counties ("län") using the Swedish LFS, we demonstrate how SAE can be utilized in practice to augment official statistics. We pay particular attention to model diagnostics, uncertainty estimation, and the comparability of model-based estimates with the direct survey estimates that users are familiar with. In doing so, the study offers a template for how other labor market indicators might be estimated for small domains under a similar approach.

Second, the thesis assesses the added value of integrating geospatial data as auxiliary variables in the SAE model for unemployment. This is a novel contribution both for Swedish official statistics and for the SAE literature in developed countries. We compile a rich set of environmental and infrastructural indicators at the county level including nighttime light intensity, vegetation indices (NDVI), land surface temperature, precipitation, and others by leveraging Google Earth Engine and other open data sources. These variables are used alongside more traditional predictors for demographic and economic indicators from administrative registers in the modeling. The analysis tests whether these satellite-derived proxies can improve the model's explanatory power and predictive accuracy for unemployment rates. To our knowledge, this is the first study to evaluate such a suite of geospatial covariates in official statistics on labor force SAE for Sweden. The findings will shed light on which types of auxiliary data are most informative for capturing regional labor market differences. By quantifying the impact on mean squared error reduction and bias, the study contributes empirical evidence to the growing literature on Big Data in official statistics, demonstrating the practical potential or limitations of these novel data sources in a realistic estimation setting.

Third, this study offers practical insights for NSIs and policy analysts in Sweden and similar countries. We evaluate the resulting county unemployment estimates against existing published figures and discuss their reliability. The thesis highlights the gains in precision achieved by the model-based estimates which can enable more counties to have publishable figures and it addresses how to communicate the added uncertainty from modeling assumptions. By doing so, we contribute to the body of knowledge on implementing SAE in official statistics. This contribution is particularly relevant as official statistics moves toward integrating multiple data sources. Our experience can inform future projects by NSIs that aim to augment sample surveys with satellite or other big data. In summary, the research not only answers a specific analytical question but also serves as a case study advancing the practice of evidence-based regional statistics. We believe this work will be of interest to both the academic SAE community and to official statistics stakeholders seeking innovative ways to meet the ever-increasing demand for aggregated-level statistics.

1.4 Thesis Outline

The remainder of this thesis is structured as follows: Chapter 2 describes in detail the methodology, including the FH model specifications, transformation techniques, parameter estimation, model selection criteria, and MSPE estimation procedures. Chapter 3 describes the data sources used, including the Swedish LFS, demographic data, and the specific geospatial covariates, along with the data processing steps. Chapter 4 presents the results of the analysis, including model diagnostics, the final unemployment estimates and their precision, and comparisons between different model specifications. Finally, Chapter 5 provides concluding remarks.

Chapter 2

Methodology

This chapter presents the methodological framework employed in this study, with a primary focus on area-level models, particularly the FH model and its extensions. The chapter begins by providing a discussion of the FH model including its formulation , parameter estimation techniques , and the prediction process using Empirical Best Linear Unbiased Prediction (EBLUP). The chapter further addresses the application of variable transformations, particularly for rates and proportions, approaches for quantifying uncertainty through Mean Squared Prediction Error (MSPE) estimation, and concludes with procedures for model selection and diagnostic evaluation.

2.1 The FH Model Framework

2.1.1 Model Specifications

The FH model structure is that of a two-level hierarchical model, which explicitly acknowledges and models two distinct sources of random variation influencing the observed direct estimate (Rao and Molina, 2015, chap.4).

Level 1: Sampling Model

The first level describes the relationship between the direct survey estimate for a small area and the true, underlying characteristic of interest for that area. Let y_i denote the direct survey estimate for area i (e.g., a county-level mean or proportion), where $i = 1, ..., m$, and m represents the total number of small areas. Let θ_i represent the true, unknown value of the characteristic for area i. The sampling model posits that the direct estimate y_i is an unbiased estimator of the true value θ_i, but is subject to sampling error, e_i:

$$y_i = \theta_i + e_i \tag{2.1}$$

The sampling error e_i reflects the uncertainty inherent in the survey strategy used. Key assumptions regarding these errors include (Rao and Molina, 2015, chap.6):

1. Unbiasedness: The expected value of the sampling error, conditional on the true value, is zero, $E[e_i|\theta_i] = 0$. This implies that the direct estimator is unbiased for the true value, $E[y_i|\theta_i] = \theta_i$.
2. Known Sampling Variance: The estimated sampling variance, $\psi_i = \text{Var}(e_i|\theta_i)$, is assumed to be known for each area i. This variance quantifies the precision of the direct estimate y_i; a smaller ψ_i indicates a more reliable direct estimate.
3. Normality: The sampling errors are assumed to follow a normal distribution, $e_i|\theta_i \sim N(0, \psi_i)$. This assumption is often justified by appealing to the Central Limit Theorem, particularly for direct estimators derived from sufficiently large, albeit potentially small for the domain, sample sizes.
4. Independence: Sampling errors e_i are assumed to be independent across different areas i.

In applications, the sampling variances ψ_i are not known. They must be estimated from the survey data, using design-based variance inline with the appropriate sampling design and the estimator. These estimated variances, denoted $\hat{\psi}_i$, are then treated as if they were the true ψ_i in the subsequent model fitting process. Rao and Molina (2015) note that this is a standard simplification in FH applications, generally considered reasonable provided the $\hat{\psi}_i$ are based on sufficient effective sample sizes and exhibit relative stability. However, it is important to acknowledge that ignoring the uncertainty in $\hat{\psi}_i$ can potentially lead to an underestimation of the true variability, particularly if the $\hat{\psi}_i$ themselves are subject to considerable error. While methods exist to account for this uncertainty, the standard FH model proceeds by treating the sampling variances as fixed and known.

Level 2: Linking Model

The second level of the model, termed the linking model, establishes a connection between the true small area means θ_i and a set of p known area-specific auxiliary variables , contained in the vector $\mathrm{x}_i = (x_{i1}, ..., x_{ip})^T$. This model aims to explain the systematic variation in the true values θ_i across areas using these covariates:

$$\theta_i = \mathrm{x}_i^\top \beta + u_i \tag{2.2}$$

Here, $\beta = (\beta_1, ..., \beta_p)^T$ represents a vector of unknown regression coefficients, assumed to be constant across all areas. These coefficients capture the average linear relationship between the auxiliary variables and the true characteristic θ_i. The term u_i denotes the area-specific random effect for area i. These random effects are a crucial component, capturing the residual spatial heterogeneity or variation in θ_i that remains unexplained by the linear predictor $\mathrm{x}_i^T\beta$. They account for unique characteristics specific to area i or the influence of unobserved covariates that affect θ_i.

Key assumptions regarding these random effects are Rao and Molina (2015):

1. Zero Mean: The expected value of the random effects is zero, $E[u_i] = 0$.
2. Homoscedasticity: The random effects are assumed to possess a constant variance across all areas, $\mathrm{Var}(u_i) = \sigma_u^2$. This variance, σ_u^2, often referred to as the model variance, is a critical parameter representing the degree of unexplained variability between the true area means after accounting for the influence of the covariates. A value of $\sigma_u^2 = 0$ would imply that the linking model perfectly explains all systematic variation in θ_i across areas, leaving only sampling error e_i.
3. Normality: The random effects are typically assumed to follow a normal distribution, $u_i \sim N(0, \sigma_u^2)$.
4. Independence: The random effects u_i are assumed to be independent across different areas i.

Furthermore, a fundamental assumption connecting the two levels is the independence of the sampling errors e_i and the random effects u_i. This reflects the conceptual distinction between variability arising from the sampling design (e_i) and variability inherent in the underlying population structure (u_i).

Combined Model

By substituting the linking model Equation 2.2 into the sampling model Equation 2.1, we obtain the combined FH model specification for a single area i:

$$y_i = \mathrm{x}_i^T\beta + u_i + e_i \tag{2.3}$$

This equation explicitly represents the observed direct estimate y_i as the sum of three components: a fixed part determined by the covariates ($\mathrm{x}_i^T\beta$), a random area-specific deviation (u_i), and sampling noise (e_i). Under the stated assumptions, the model implies that the direct estimates y_i are independent across areas and follow a normal distribution:

$$y_i \sim N(\mathrm{x}_i^T\beta, \sigma_u^2 + \psi_i) \tag{2.4}$$

The total variance of the observation y_i is the sum of the model variance and the sampling variance, $\mathrm{Var}(y_i) = \mathrm{Var}(u_i) + \mathrm{Var}(e_i) = \sigma_u^2 + \psi_i$, owing to the independence of u_i and e_i.

For mathematical convenience and implementation, it is often useful to express the model for all m areas simultaneously using matrix notation, as detailed by Rao & Molina (2015). Let $\mathrm{y} = (y_1, ..., y_m)^T$ be the vector of direct estimates, X be the $m \times p$ matrix of auxiliary variables with the i-th row being x_i^T, β be the $p \times 1$ vector of regression coefficients, $\mathrm{u} = (u_1, ..., u_m)^T$ be the vector of random effects, and $\mathrm{e} = (e_1, ..., e_m)^T$ be the vector of sampling errors. The combined model in matrix form is:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{u} + \mathrm{e} \tag{2.5}$$

Under the model assumptions, the distributions of the random vectors are:

$$\mathrm{u} \sim N(0, \mathrm{G}) \quad \text{where} \quad \mathrm{G} = \sigma_u^2 \mathrm{I}_m \tag{2.6}$$

$$\mathrm{e} \sim N(0, \mathit{\Psi}) \quad \text{where} \quad \mathit{\Psi} = \mathrm{diag}(\psi_1, ..., \psi_m) \tag{2.7}$$

Here, I_m denotes the $m \times m$ identity matrix, and $\mathit{\Psi}$ is the $m \times m$ diagonal matrix containing the known sampling variances. Given the assumed independence of u and e, the variance-covariance matrix of the observation vector y is:

$$\mathrm{V} = \mathrm{Var}(\mathrm{y}) = \mathrm{Var}(\mathrm{u} + \mathrm{e}) = \mathrm{Var}(\mathrm{u}) + \mathrm{Var}(\mathrm{e}) = \mathrm{G} + \mathit{\Psi} = \sigma_u^2 \mathrm{I}_m + \mathit{\Psi} \tag{2.8}$$

Consequently, the marginal distribution of the observation vector y, after integrating out the random effects u, is multivariate normal:

$$\mathrm{y} \sim N(\mathrm{X}\beta, \mathrm{V}) \tag{2.9}$$

This formulation clearly shows that the FH model is a specific instance of a linear mixed model. The variance structure V is crucial, depending on the unknown model variance parameter σ_u^2 and the known sampling variances ψ_i. This structure forms the basis for estimating the model parameters β and σ_u^2, and subsequently for predicting the small area means θ_i.

2.1.2 Role of Auxiliary Variables

The practical success of the FH model hinges critically on the availability and predictive power of the auxiliary variables x_i. Ideally, these variables should exhibit a strong correlation with the variable of interest (θ_i) and must be available for all small areas within the population scope. The careful selection, processing, and validation of relevant and high quality auxiliary variables represent a critical preliminary step in any successful application of the FH model.

2.2 Parameter Estimation

Applying the FH model requires estimating the unknown parameters: the vector of regression coefficients β and the random effects variance σ_u^2. The estimation of σ_u^2 is particularly pivotal, as its value directly influences the degree of shrinkage applied to the direct estimates when forming the final predictions.

2.2.1 Estimation of Regression Coefficients (β)

If the variance components matrix $\mathrm{V} = \sigma_u^2 \mathrm{I}_m + \mathit{\Psi}$ were fully known (specifically, if σ_u^2 were known), the Best Linear Unbiased Estimator (BLUE), also known as the Generalized Least Squares (GLS) estimator, of β could be calculated directly. As shown by Rao and Molina (2015), this estimator is:

$$\tilde{\beta}(\sigma_u^2) = (\mathrm{X}^T \mathrm{V}^{-1} \mathrm{X})^{-1} \mathrm{X}^T \mathrm{V}^{-1} \mathrm{y} \tag{2.10}$$

However, in practice, σ_u^2 is unknown and must be estimated from the data. Let $\hat{\sigma}_u^2$ denote a suitable estimator of σ_u^2. By substituting this estimate into the expression for V, we obtain an estimated variance matrix $\hat{\mathrm{V}} = \hat{\sigma}_u^2 \mathrm{I}_m + \mathit{\Psi}$. The two-stage or Empirical Best Linear Unbiased Estimator (EBLUE) of β is then obtained by plugging $\hat{\mathrm{V}}$ into the GLS formula:

$$\hat{\beta} = \tilde{\beta}(\hat{\sigma}_u^2) = (\mathrm{X}^T \hat{\mathrm{V}}^{-1} \mathrm{X})^{-1} \mathrm{X}^T \hat{\mathrm{V}}^{-1} \mathrm{y} \tag{2.11}$$

The statistical properties of this estimator $\hat{\beta}$ are inherently linked to the properties of the chosen estimator $\hat{\sigma}_u^2$.

2.2.2 Estimation of Random Effects Variance (σ_u^2)

Several methods have been proposed and are used for estimating the single variance component σ_u^2 in the FH model including (Rao and Molina, 2015, chap.6):

1. Method of Moments (FH Method): This iterative approach, originally proposed alongside the model by Fay and Herriot (1979), is based on equating weighted sums of squares of residuals to their expected values. While relatively simple to implement, it does not guarantee non-negative estimates, and resulting negative values are typically truncated to zero. This truncation can introduce bias.

2. Maximum Likelihood (ML): This method is based on maximizing the marginal likelihood of the observed data y, derived from the multivariate normal distribution specified in Equation 2.9, under the assumption of normality for both error terms. The log-likelihood function (ignoring constant terms) is given by:
$$l_{ML}(\beta, \sigma_u^2) = -\frac{1}{2}\log|\mathrm{V}| - \frac{1}{2}(\mathrm{y} - \mathrm{X}\beta)^T\mathrm{V}^{-1}(\mathrm{y} - \mathrm{X}\beta) \tag{2.12}$$
ML estimators $\hat{\beta}_{ML}$ and $\hat{\sigma}_{u,ML}^2$ are obtained by simultaneously maximizing this function. It is well-known that ML estimators of variance components can be biased downwards, particularly when the number of areas (m) is small.

3. Restricted (or Residual) Maximum Likelihood (REML): REML estimation aims to mitigate the bias observed in ML variance component estimators. It achieves this by maximizing a modified likelihood function derived from residual contrasts, which effectively accounts for the degrees of freedom lost in estimating the fixed effects β. The REML log-likelihood function (ignoring constants) depends only on σ_u^2:
$$l_{REML}(\sigma_u^2) = -\frac{1}{2}\log|\mathrm{V}| - \frac{1}{2}\log|\mathrm{X}^T\mathrm{V}^{-1}\mathrm{X}| - \frac{1}{2}\mathrm{y}^T\mathrm{P}\mathrm{y} \tag{2.13}$$
where $\mathrm{P} = \mathrm{V}^{-1} - \mathrm{V}^{-1}\mathrm{X}(\mathrm{X}^T\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}^T\mathrm{V}^{-1}$. Maximizing $l_{REML}(\sigma_u^2)$ yields the REML estimator $\hat{\sigma}_{u,REML}^2$. This maximization typically requires iterative numerical methods, such as the Newton-Raphson or Fisher scoring algorithms. Due to their reduced bias compared to ML estimators, REML estimators of σ_u^2 are generally preferred in practice, especially in official statistics applications (Rao and Molina, 2015, chap.6).

Both ML and REML methods inherently produce non-negative estimates of σ_u^2. However, it is still possible, particularly if the true variance is small or the number of areas m is limited, for the estimate to be $\hat{\sigma}_u^2 = 0$. This outcome implies that the model finds no significant variation between areas beyond that explained by the covariates and sampling error.

2.3 Prediction: EBLUP

The ultimate objective in SAE is typically the prediction of the small area means $\theta_i = \mathrm{x}_i^T\beta + u_i$. Within the framework of the FH model Equation 2.3, Henderson (1975) showed that if the parameters β and σ_u^2 were known, the Best Linear Unbiased Predictor (BLUP) of θ_i takes the form of a weighted average. It optimally combines the direct estimate y_i and the regression-synthetic predictor $\mathrm{x}_i^T\beta$ (Rao and Molina, 2015, chap.6):

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i)\mathrm{x}_i^T\beta \tag{2.14}$$

The weight γ_i, often referred to as the shrinkage factor, is defined as:

$$\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i} \tag{2.15}$$

This shrinkage factor γ_i lies in the interval [0, 1] and represents the proportion of the total variance (model variance plus sampling variance, $\sigma_u^2 + \psi_i$) that is attributable to the model variance σ_u^2. The BLUP $\tilde{\theta}_i$ thus represents a composite estimator that optimally balances the information from the direct estimate and the regression prediction based on their relative precisions:

1. When the sampling variance ψ_i is small relative to the model variance σ_u^2 (indicating a precise direct estimate y_i), γ_i approaches 1. In this case, the BLUP gives substantial weight to the direct estimate y_i.
2. Conversely, when the sampling variance ψ_i is large relative to σ_u^2 (indicating an imprecise direct estimate y_i), γ_i approaches 0. The BLUP then shrinks towards the regression prediction $\mathrm{x}_i^T\beta$, effectively borrowing strength from the auxiliary variables and the information pooled across all areas.
3. In the specific case where $\sigma_u^2 = 0$ (implying the linking model perfectly explains the true means), $\gamma_i = 0$, and the BLUP reduces entirely to the synthetic regression predictor $\mathrm{x}_i^T\beta$.

Since β and σ_u^2 are unknown in practical scenarios, they must be replaced by their respective estimates, $\hat{\beta}$ (from Equation 2.11) and $\hat{\sigma}_u^2$ (e.g., the REML estimate). Substituting these estimates into the BLUP formula yields the EBLUP of θ_i:

$$\hat{\theta}_i^E = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\mathrm{x}_i^T\hat{\beta} \tag{2.16}$$

where $\hat{\gamma}_i$ is the estimated shrinkage factor:

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i} \tag{2.17}$$

The EBLUP $\hat{\theta}_i^E$ is the standard predictor utilized in applications of the FH model. It provides a data-driven composite estimate that adaptively balances the direct estimate and the regression prediction based on the estimated model variance and the known sampling variance for each area.

2.4 Transformations

When the target variable of interest represents a proportion or a rate, which is inherently bounded (typically between 0 and 1), directly applying the standard FH model to the estimates y_i can encounter several theoretical and practical difficulties:

1. Violation of Normality: Proportions or rates, especially those close to the boundaries of 0 or 1, often exhibit skewed distributions, potentially violating the normality assumptions made for the sampling errors e_i and the random effects u_i.
2. Heteroscedasticity: The sampling variance ψ_i of a proportion frequently depends on the true proportion θ_i itself (e.g., for a binomial proportion, $\psi_i \approx \theta_i(1 - \theta_i)/n_i$, where n_i is the sample size). This violates the standard FH assumption that ψ_i is independent of θ_i.
3. Predictions Outside Bounds: The EBLUP $\hat{\theta}_i^E$, being a linear combination, is not inherently constrained to fall within the valid [0, 1] range, potentially leading to nonsensical predictions.

To mitigate these issues, we apply a suitable nonlinear transformation $g(\cdot)$ to the direct estimates y_i before fitting the FH model. This approach aims to achieve properties closer to the model assumptions on the transformed scale (Rao and Molina, 2015, chap.6).

Let $y_i^* = g(y_i)$ be the transformed direct estimate. The corresponding sampling variance ψ_i^* on the transformed scale needs to be derived. One approach is to use the delta method, which provides a first-order approximation: $\psi_i^* \approx [g'(\theta_i)]^2\psi_i$. In practice, this is often evaluated by substituting the direct estimate y_i for the unknown θ_i, yielding $\psi_i^* \approx [g'(y_i)]^2\psi_i$.

The FH model is then specified and fitted on the transformed scale:

$$y_i^* = \mathrm{x}_i^T \beta^* + u_i^* + e_i^* \tag{2.18}$$

where it is assumed that $e_i^* \sim N(0, \psi_i^*)$ and $u_i^* \sim N(0, \sigma_{u^*}^2)$. Fitting this model yields estimates $\hat{\beta}^*$ and $\hat{\sigma}_{u^*}^2$, leading to EBLUPs $\hat{\theta}_i^{E*}$ for the transformed true values $\theta_i^* = g(\theta_i)$.

This study considers two transformations for proportions:

1. Logit Transformation: $g(y) = \log(y/(1-y))$. This transforms the [0, 1] interval to the entire real line. The approximate variance on the logit scale is $\psi_i^* \approx \psi_i/[y_i(1-y_i)]^2$. This transformation requires $0 < y_i < 1$, potentially needing adjustments for estimates exactly equal to 0 or 1.
2. Arcsine Square Root Transformation: $g(y) = \arcsin(\sqrt{y})$. This transformation is particularly noted for its variance-stabilizing properties for binomial proportions. If y_i is a sample proportion based on an effective sample size n_i, its variance is approximately $\theta_i(1-\theta_i)/n_i$. On the arcsine scale, the variance becomes approximately constant: $\psi_i^* \approx 1/(4n_i)$ (Rao and Molina, 2015, chap.6). This simplifies the modeling as the sampling variance component becomes independent of the mean.

After obtaining the EBLUP $\hat{\theta}_i^{E*}$ on the transformed scale, a crucial final step is to back-transform it to the original scale of interest using the inverse function $h(\cdot) = g^{-1}(\cdot)$. A simple, naive back-transformation $\hat{\theta}_i^{naive} = h(\hat{\theta}_i^{E*})$ is generally biased because the inverse transformation $h(\cdot)$ is typically nonlinear. Therefore, bias-corrected back-transformations are necessary to obtain more accurate estimates on the original scale. Assuming normality holds approximately on the transformed scale, such that the predictor $\hat{\theta}_i^{E*}$ can be considered an estimate of θ_i^* with mean μ_i^* and variance σ_i^{*2} (approximated by the MSPE on the transformed scale), a second-order Taylor expansion leads to an approximate bias correction. As detailed by Rao and Molina (2015) the bias-corrected estimate can be approximated as:

$$\hat{\theta}_i^{BC} \approx h(\hat{\theta}_i^{E*}) + \frac{1}{2} h''(\hat{\theta}_i^{E*}) \widehat{\mathrm{MSPE}}(\hat{\theta}_i^{E*}) \tag{2.19}$$

where $h''(\cdot)$ is the second derivative of the inverse transformation function, and $\widehat{\mathrm{MSPE}}(\hat{\theta}_i^{E*})$ is an estimate of the MSPE of the EBLUP on the transformed scale. The specific form of the bias correction depends explicitly on the chosen transformation $h(\cdot)$. For the arcsine transformation, where $h(x) = (\sin(x))^2$, the bias correction involves the estimated MSPE on the arcsine scale.

2.5 Mean Squared Prediction Error (MSPE) Estimation

Providing a reliable measure of uncertainty associated with the produced estimates is a fundamental requirement in statistical estimation, particularly within the context of official statistics. For EBLUPs derived from SAE models, the MSPE serves as the standard measure of variability. It quantifies the expected squared difference between the predictor and the true value:

$$\mathrm{MSPE}(\hat{\theta}_i^E) = E[(\hat{\theta}_i^E - \theta_i)^2] \tag{2.20}$$

The expectation here is taken with respect to the underlying model distribution. The MSPE accounts for multiple sources of uncertainty: the inherent variability in predicting the random effect u_i, and the additional variability introduced by the need to estimate the model parameters β and σ_u^2 from the data.

2.5.1 Analytical MSPE Approximation (Prasad-Rao)

For the standard FH model (without transformations), Prasad and Rao (1990) derived the second-order approximation for the MSPE of the EBLUP $\hat{\theta}_i^E$. This approximation holds under certain regularity conditions and assumes that $\hat{\sigma}_u^2$ is obtained via methods like the FH method of moments, ML, or REML. Their derivation decomposes the MSPE into three additive components (Rao and Molina, 2015, chap.6):

$$\mathrm{MSPE}(\hat{\theta}_i^E) \approx g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) \tag{2.21}$$

These components represent distinct sources of error:

1. $g_{1i}(\sigma_u^2) = \gamma_i \psi_i = \frac{\sigma_u^2 \psi_i}{\sigma_u^2 + \psi_i}$: This is the leading term and represents the uncertainty inherent in predicting the random effect u_i, assuming the model parameters β and σ_u^2 were known. It corresponds to the variance of the BLUP.
2. $g_{2i}(\sigma_u^2) = (1-\gamma_i)^2 \mathrm{x}_i^T(\mathrm{X}^T\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{x}_i$: This term accounts for the uncertainty introduced by having to estimate the regression coefficients β.
3. $g_{3i}(\sigma_u^2) = (\psi_i/(\sigma_u^2+\psi_i))^2 \mathrm{Var}(\hat{\sigma}_u^2)/(\sigma_u^2+\psi_i)$: This term accounts for the uncertainty arising from the estimation of the variance component σ_u^2. Its magnitude depends on the variance of the estimator $\hat{\sigma}_u^2$, denoted $\mathrm{Var}(\hat{\sigma}_u^2)$, which can be approximated based on the chosen estimation method (e.g., using the inverse of the Fisher information matrix for ML/REML estimators).

To obtain an estimator of the MSPE, the unknown σ_u^2 in these terms is replaced by its estimate $\hat{\sigma}_u^2$. However, simply substituting $\hat{\sigma}_u^2$ into the sum $g_{1i}+g_{2i}+g_{3i}$ results in an estimator that is biased downwards (typically of order $O(m^{-1})$). Prasad and Rao (1990), and later Datta and Lahiri (2000) for ML/REML, showed that an approximately unbiased estimator of the MSPE (correct to the second order) is obtained by adding a bias correction term, which often simplifies to doubling the g_{3i} term evaluated at $\hat{\sigma}_u^2$:

$$\widehat{\mathrm{MSPE}}(\hat{\theta}_i^E) \approx g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) \tag{2.22}$$

This analytical approximation provides valuable insights into the components of prediction error and performs well under the model assumptions. However, its derivation can become complex, particularly if transformations are applied or if the model structure deviates significantly from the basic FH assumptions (e.g., spatial correlation in random effects).

2.5.2 Bootstrap MSPE Estimation

Resampling methods, particularly the parametric bootstrap, offer a powerful and flexible alternative for estimating the MSPE. Bootstrap approaches are often computationally intensive but can handle complex situations more readily, such as models involving transformations or non-standard error distributions (Rao and Molina, 2015, chap.6).

Parametric bootstrap procedure tailored for the FH model involves simulating data generation under the fitted model. The steps are typically as follows:

1. Fit the chosen FH model (potentially on a transformed scale) to the original data y (or y^*) to obtain estimates $\hat{\beta}$ (or $\hat{\beta}^*$) and $\hat{\sigma}_u^2$ (or $\hat{\sigma}_{u^*}^2$).
2. Repeat the following steps for $b = 1, ..., B$, where B is a large number of bootstrap replicates (e.g., 500 or 1000):
    a. Generate bootstrap random effects $u_i^{*(b)}$ independently from $N(0, \hat{\sigma}_u^2)$ for each area $i = 1, ..., m$.
    b. Generate bootstrap sampling errors $e_i^{*(b)}$ independently from $N(0, \psi_i)$ for each area $i = 1, ..., m$. (If working on a transformed scale, use $\hat{\sigma}_{u^*}^2$ and ψ_i^*).
    c. Construct the b-th bootstrap sample $\mathrm{y}^{*(b)}$ using the fitted model structure: $y_i^{*(b)} = \mathrm{x}_i^T\hat{\beta} + u_i^{*(b)} + e_i^{*(b)}$. (Use starred parameters if on transformed scale).
    d. Fit the same FH model specification to the bootstrap sample $\mathrm{y}^{*(b)}$ to obtain bootstrap parameter estimates $\hat{\beta}^{*(b)}$ and $\hat{\sigma}_u^{2*(b)}$.
    e. Calculate the bootstrap EBLUP for area i based on the b-th replicate: $\hat{\theta}_i^{E*(b)} = \hat{\gamma}_i^{*(b)} y_i^{*(b)} + (1-\hat{\gamma}_i^{*(b)})\mathrm{x}_i^T\hat{\beta}^{*(b)}$, where $\hat{\gamma}_i^{*(b)} = \hat{\sigma}_u^{2*(b)}/(\hat{\sigma}_u^{2*(b)} + \psi_i)$. (Perform on transformed scale and back-transform if necessary, potentially applying bias correction during back-transformation).

f. Define the bootstrap 'true' value for area i in replicate b as $\theta_i^{*(b)} = \mathrm{x}_i^T\hat{\beta} + u_i^{*(b)}$. (Use starred parameters if on transformed scale, then back-transform $\theta_i^{*(b)}$ if needed, usually without bias correction for the 'true' value).

3. The bootstrap MSPE estimator for area i is then calculated as the average squared difference between the bootstrap EBLUPs and the corresponding bootstrap true values over the B replicates:

$$\widehat{\mathrm{MSPE}}_B(\hat{\theta}_i^E) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_i^{E*(b)} - \theta_i^{*(b)})^2 \tag{2.23}$$

This bootstrap approach naturally incorporates the uncertainty arising from the estimation of both β and σ_u^2. When transformations are involved, the bootstrap simulation and prediction steps are performed on the transformed scale, and the squared differences are calculated after appropriate back-transformation of both the predictor and the 'true' value, ensuring the MSPE is estimated on the original scale of interest.

2.6 Model Selection and Diagnostics

2.6.1 Model Selection Criteria

In situations where multiple potential auxiliary variables are available, model selection techniques play a crucial role in identifying a parsimonious model that exhibits good predictive performance while avoiding overfitting. Several information criteria, adapted for or applicable to linear mixed models like the FH model, are employed for this purpose. Akaike Information Criterion (AIC) is defined as $AIC = -2l_{ML}(\hat{\beta}, \hat{\sigma}_u^2) + 2(p+1)$, where l_{ML} is the maximized log-likelihood value, p is the number of regression coefficients (excluding intercept if present), and m is the number of areas (Rao and Molina, 2015, chap.5).

AIC balances model fit (likelihood) with complexity (number of parameters). Bayesian Information Criterion (BIC) is defined as $BIC = -2l_{ML}(\hat{\beta}, \hat{\sigma}_u^2) + (p+1)\log(m)$. BIC imposes a stronger penalty for model complexity than AIC, particularly for larger m, often leading to the selection of more parsimonious models. Models exhibiting lower values for these criteria are generally preferred. This paper uses the automated step-wise selection procedure based on these criteria with attention paid to the variable's substantive relevance and justification for its inclusion.

2.6.2 Model Diagnostics

Checking the validity of the assumptions underlying the FH model is essential for confirming the reliability of the resulting EBLUPs and their associated MSPE estimates. Standard diagnostic procedures, adapted from linear mixed model theory, should be employed (Rao and Molina, 2015). Examination of model residuals is fundamental. Standardized residuals, $\hat{r}_i = (y_i - \mathrm{x}_i^T\hat{\beta})/\sqrt{\hat{\sigma}_u^2 + \psi_i}$, should ideally exhibit no discernible patterns when plotted against fitted values ($\mathrm{x}_i^T\hat{\beta}$) or individual covariates. Such plots can help detect issues like non-linearity in the fixed effects part or heteroscedasticity not accounted for by the model. Q-Q plots and formal normality tests like Shapiro-Wilk test used in this study are applied to these residuals can assess the plausibility of the normality assumption for the combined error term $u_i + e_i$.

The predicted random effects, $\hat{u}_i = \hat{\theta}_i^E - \mathrm{x}_i^T\hat{\beta}$, can be examined. Q-Q plots and normality tests applied to $\hat{u}_i$ help assess the normality assumption for the area-specific effects u_i. Plotting $\hat{u}_i$ against covariates can reveal if any systematic variation related to those covariates remains unexplained in the random part of the model. Formal statistical tests can be conducted to compare the fitted model against alternative specifications or to assess the overall fit. For instance, comparing the model likelihood against a simpler or more complex model. If diagnostic checks reveal significant departures from the model assumptions, appropriate remedial actions should be considered. These might include transforming the response variable or covariates, including non-linear terms for covariates, using robust estimation methods less sensitive to outliers, or exploring extensions of the FH model that relax certain assumptions.

Chapter 3

Data

This chapter presents the dataset utilized in this study, beginning with the direct unemployment rate estimates and subsequently detailing the processes employed in acquiring auxiliary variables, including geospatial, administrative, and derived covariates. The chapter also provides preliminary analysis of the data and concludes with a discussion of limitations associated with the dataset.

3.1 Direct Estimates

The direct estimates of the unemployment rates used in this study were sourced from Statistics Sweden (SCB) via their public API[1]. Specifically, this analysis utilizes data from the estimates reported in the first quarter of 2025. The LFS is a reliable, nationally representative survey designed to comprehensively capture the dynamics of the Swedish labor market[2]. It includes detailed metrics such as employment status, unemployment rates, workforce participation rates, and demographic breakdowns both at the national and county level.

Both reported unemployment rate estimates and their margin of errors corresponding to 95% confidence interval were obtained. From these values, the estimated sampling variance[3] and the estimated effective sample sizes[4] were the computed for each county.

3.2 Auxiliary Data

3.2.1 Geospatial Covariates

Geospatial covariates were retrieved using Google Earth Engine (GEE)[5], which provided access to various satellite-derived environmental and infrastructural datasets for the year 2024. The geospatial data was processed through annual averaging. The chosen covariates and their respective resolutions are detailed below:

3.2.2 Extra Covariates

Number of new vacancies per county obtained via API from the Swedish Public Employment Service[6]. The population density[7] per km^2 for the year 2024 is also obtained from SCB[8] to be used as auxiliary variable. It can be observed in Figure 5.1 that the direct estimates for the counties in the Northern Sweden have lower unemployment rates. Therefore, to capture this disparity based on the geographical positioning in the counties this study derives and extra dummy variable that groups counties as either South and North counties[9].

---

[1] The documentation of the API - PxWebApi 1.0 for the Statistical Database

[2] Further info on LFS at SCB - Labour Force Surveys (LFS)

[3] $\hat{\psi} = \left(\frac{\text{Margin of Error}}{1.96}\right)^2$, where $\hat{\psi}$ is the estimated sampling variance.

[4] For a proportion estimate $\hat{p}$, the effective sample size (n_{eff}) is given by: $n_{eff} = \frac{\hat{p}(1-\hat{p})}{\hat{\psi}}$

[5] Source of geospatial data - Google Earth Engine

[6] Data source link from Arbetsförmedlingen.

[7] SCB describes Population Density $= \frac{\text{Total Population}}{\text{Land Area in } km^2}$

[8] Data source link from SCB Statistical Database

[9] Counties defined as North were those originally in the Norrland region of Sweden namely; Gävleborg, Jämtland, Norrbotten, Västerbotten and Västernorrland.

Table 3.1: Geospatial Covariates

Covariate	Data Source	Spatial Resolution	Frequency
Nighttime Lights	NOAA VIIRS DNB monthly	~500 m	Monthly
Urban Cover	Google Dynamic World	10 m	Continuous
NDVI (Normalized Difference Vegetation Index)	MODIS Terra	1 km	16-day
Land-Surface Temperature (LST)	MODIS	1 km	8-day
Precipitation	CHIRPS daily	~5 km	Daily
Tropospheric NO_2	Sentinel-5P	~7 km	Daily
Soil Moisture	NASA GLDAS-2.1 NOAH	~25 km	3-hourly
Elevation and Slope	Copernicus DEM GLO-30	30 m	Static

3.3 Data Concerns

Direct estimates were obtained for 20 of Sweden's 21 counties; Gotland County lacked sufficient data and thus had no reported estimates. According to Statistics Sweden, counties with fewer than 20 observations for monthly estimates or fewer than 40 for quarterly estimates are deemed too uncertain to report[10]. In our SAE framework, we treat Gotland as an out-of-sample unit. Integrating satellite-derived covariates into this framework introduces additional challenges, primarily driven by temporal and spatial misalignment. While SCB publishes unemployment rates on a quarterly and annual basis, our geospatial predictors such as land-surface temperature, vegetation indices, and nighttime lights are compiled at differing temporal resolutions (daily, monthly, or annually, depending on the indicator) and may cover slightly different geographic footprints. To address these mismatches, we aggregate or average the remote-sensing measurements to match the reporting period of each direct estimate and apply spatial overlays to ensure that all covariate values correspond precisely to county boundaries. Even with these harmonization steps, residual timing and boundary effects may introduce additional uncertainty into our model estimates.

---

[10] SCB defines "few observations" as fewer than 20 for monthly and fewer than 40 for quarterly estimates. Source link – SCB

Chapter 4

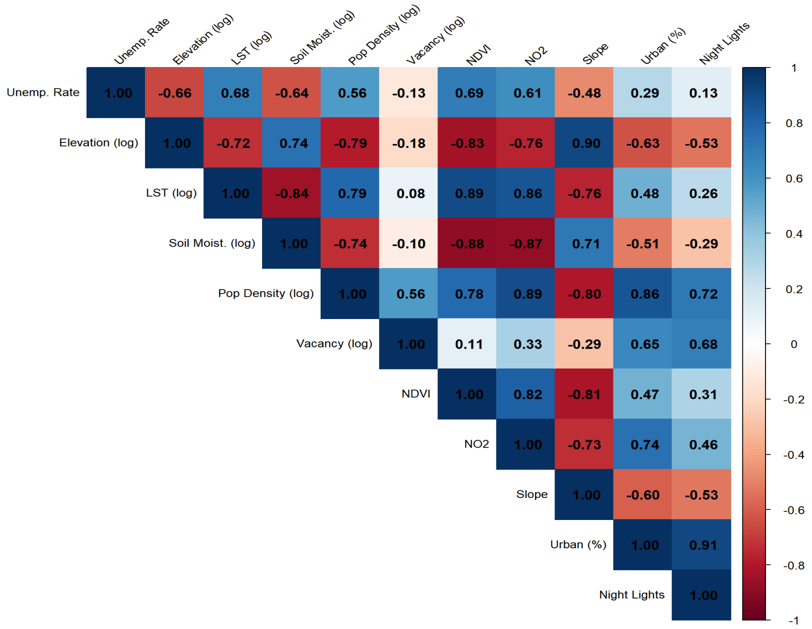
Empirical Results and Analysis

4.1 Software Setup

The thesis analysis was done and authored using R programming language, R version 4.5.0 (2025-04-11 ucrt) via the `Quarto` version 1.8.14. The main package used for the SAE framework is the *emdi* version 2.2.2. The analysis process and the codebase for this project is hosted on Github[1], complete with data collection process, preprocessing of the data and the SAE Modeling.

4.2 Preliminary Analysis

4.2.1 Examining Linear Relationships



	Unemp. Rate	Elevation (log)	LST (log)	Soil Moist. (log)	Pop Density (log)	Vacancy (log)	NDVI	NO2	Slope	Urban (%)	Night Lights
Unemp. Rate	1.00	-0.66	0.68	-0.64	0.56	-0.13	0.69	0.61	-0.48	0.29	0.13
Elevation (log)		1.00	-0.72	0.74	-0.79	-0.18	-0.83	-0.76	0.90	-0.63	-0.53
LST (log)			1.00	-0.84	0.79	0.08	0.89	0.86	-0.76	0.48	0.26
Soil Moist. (log)				1.00	-0.74	-0.10	-0.88	-0.87	0.71	-0.51	-0.29
Pop Density (log)					1.00	0.56	0.78	0.89	-0.80	0.86	0.72
Vacancy (log)						1.00	0.11	0.33	-0.29	0.65	0.68
NDVI							1.00	0.82	-0.81	0.47	0.31
NO2								1.00	-0.73	0.74	0.46
Slope									1.00	-0.60	-0.53
Urban (%)										1.00	0.91
Night Lights											1.00

Figure 4.1: Correlation Matrix

In Figure 4.1 the topographic covariates of log soil moisture, log elevation and slope all display negative correlations with the unemployment rate, with r values of -0.64, -0.66 and -0.48 respectively, indicating that counties characterized by wetter soils, higher altitude and steeper terrain tend to exhibit lower levels of unemployment. By contrast, the surface and vegetation variables of log land surface temperature, NDVI

---

[1] Github link of the analysis - https://github.com/joseph-data/coding

and NO_2 concentration correlate positively with unemployment, with r values of 0.68, 0.69 and 0.61, which suggests that warmer, greener and more polluted regions are also those facing higher unemployment.

The socioeconomic urbanization group, which includes log population density, percent urban and night time light intensity, also shows positive associations with unemployment, with r values of 0.56, 0.29 and 0.13, while log vacancy rates correlate slightly negatively with unemployment, r = -0.13. Within each of these groups the variables are themselves highly collinearity, for example elevation and slope correlate at r = 0.90, land surface temperature and NDVI at r = 0.89 and NDVI and NO_2 at r = 0.82. These high collinearity underscore multicollinearity concerns and argue for selecting a single representative variable from each group or applying variable selection techniques to ensure stable and interpretable parameter estimates.

4.3 Modeling

Table 4.1: Resulting Models

	Initial[2] Model	Log Transformed	Arcsin Transformed	Initial Model (Reduced)	Log Transformed (Reduced)	Arcsin Transformed (Reduced)
Intercept	-0.29 (0.54)	-8.25 (0.08)	-0.64 (0.79)	0.11 (0.10)	-2.36 (0.00)	0.33 (0.24)
Elevation (log)	-0.01 (0.61)	-0.07 (0.63)	-0.01 (0.89)	-0.02 (0.01)	-0.29 (0.00)	-0.05 (0.29)
LST (log)	-0.01 (0.63)	-0.09 (0.64)	-0.02 (0.85)	–	–	–
Soil Moisture (log)	0.10 (0.41)	1.45 (0.22)	0.24 (0.69)	–	–	–
Population Density	0.03 (0.18)	0.37 (0.10)	0.06 (0.60)	0.02 (0.02)	0.30 (0.00)	0.05 (0.29)
New Vacancies (log)	-0.01 (0.23)	-0.16 (0.11)	-0.03 (0.58)	–	–	–
NDVI	0.04 (0.69)	0.47 (0.60)	0.08 (0.85)	–	–	–
NO_2	6102 (0.29)	79960 (0.16)	13156 (0.64)	–	–	–
Slope	0.45 (0.17)	5.62 (0.10)	0.89 (0.60)	0.53 (0.02)	6.84 (0.00)	1.15 (0.29)
Urban Cover (%)	-0.02 (0.24)	-0.29 (0.12)	-0.05 (0.62)	-0.01 (0.00)	-0.11 (0.00)	-0.02 (0.18)
VIIRS	0.03 (0.42)	0.48 (0.25)	0.08 (0.71)	–	–	–
Northern	0.01 (0.55)	0.13 (0.48)	0.02 (0.81)	–	–	–
R^2	0.89	0.93	1.00	0.95	0.97	1.00
Adjusted R^2	0.54	0.61	0.57	0.64	0.70	0.67
Loglike	62.54	17.63	33.69	60.79	14.27	33.37
AIC	-99.08	-9.26	-41.37	-109.58	-16.54	-54.74
BIC	-86.14	3.69	-28.43	-103.60	-10.56	-48.76

The results in Table 4.1 for the fitted FH models show gains including the geospatial covariates to model the unemployement rate. Consistently, the R^2 values in the full models suggest a reasonable fit (0.89, 0.93 and 1.00 for the initial, log transformed and arcsin transformed models). However, looking at the suitability of the included covariates in the full models they appear not to be significant to the model with p-values lower than 5% level of significance. Even the predictors which had shown positive correlation with the unemployment rate (See Figure 4.1), for instance NDVI ($r = 0.69$) included in the full models did not show significance (0.04, $p = 0.69$ (initial model); 0.47, $p = 0.61$ (log transformed -); 0.08, $p = 0.85$ (arcsin transformed model)). Examining the AICs (initial model = -99.08, log transform = -9.26, arcsin transformed = -41.37) and BICs (initial model = -86.14, log transform = 3.69, arcsin transformed = -28.43) for the full models confirm that including all the variables in the model yields marginal improvements at the cost of complexity. The loglikehood values (initial model = 62.54, log transform = 17.63, arcsin transformed = 33.69) further confirm that the more complex transformed fits do not outperform the simpler untransformed case.

Coefficient signs are nevertheless consistent across scales and inline with economic intuition. In every full specification, log-elevation and urban land-cover have negative slopes (-0.01 and -0.03 in the initial model; -0.07 and -0.29 on the log scale), signalling that higher, more urbanized counties tend to have lower unemployment rates. Conversely, population density and slope are positive (0.03 and 0.45 initially; 0.37 and 5.62 on the log scale), implying denser or more rugged terrain is linked to higher unemployment rates.

---

[2] Initial model is the model with the original scale of unemployment rates.

Labor-demand proxy new vacancies is negative (-0.01; -0.16), while environmental greenness (NDVI, 0.04) and industrial activity proxies (NO_2, 6 102; VIIRS lights, 0.03) are positive. The direction of effects are persistent even with the log and arcsine transformations though collinearity inflates some coefficients (NO_2 coefficient is 79 960 in the log model), underscoring that the sign pattern is robust but the full model is unstable.

To improve parsimony and guard against overfitting, this study refined each model by removing covariates that did not contribute significantly. Beginning with the all the covariates in the FH models, we removed the least significant predictor one at a time, retaining each deletion only if it lowered BIC and repeated this process until no further improvement was possible. This systematic pruning recognized that many geospatial metrics were highly correlated (for instance, elevation versus slope (r = -0.48), and NDVI versus NO_2 (r = 0.82)) and that their simultaneous inclusion inflated standard errors without delivering stable inference. The resulting reduced models on each scale converged on four core predictors: log elevation, population density, slope, and percent urban cover. In the untransformed reduced model, this quartet raises R^2 from 0.89 to 0.95, boosts adjusted R^2 from 0.54 to 0.64, and lowers AIC/BIC to –109.58/–103.61. All four coefficients now achieve 5% significance (elevation: –0.03, p = 0.01; density: 0.02, p = 0.02; slope: 0.53, p = 0.02; urban cover: –0.01, p = 0.002). Parallel gains appear in the log-transformed reduced model, though its information-criterion scores (AIC = –16.54; BIC = –10.56) remain far weaker than the untransformed version, and the arcsine-reduced fit, while formally perfect in full model R^2, does not attain comparable parsimony or parameter precision.

Table 4.1 shows that the reduced original scale FH model with log elevation, slope, population density and urban cover percentage is the preferred model. The model, had better overall fit when examining the AIC and BIC values. The transformed models did not outperform the original and introduces uncertainties due to back-transformation. The original scale model strikes a good balance between parsimony and explanatory power and is preferred in this study in the production of county-level estimates.

4.4 Model Diagnostic Checks

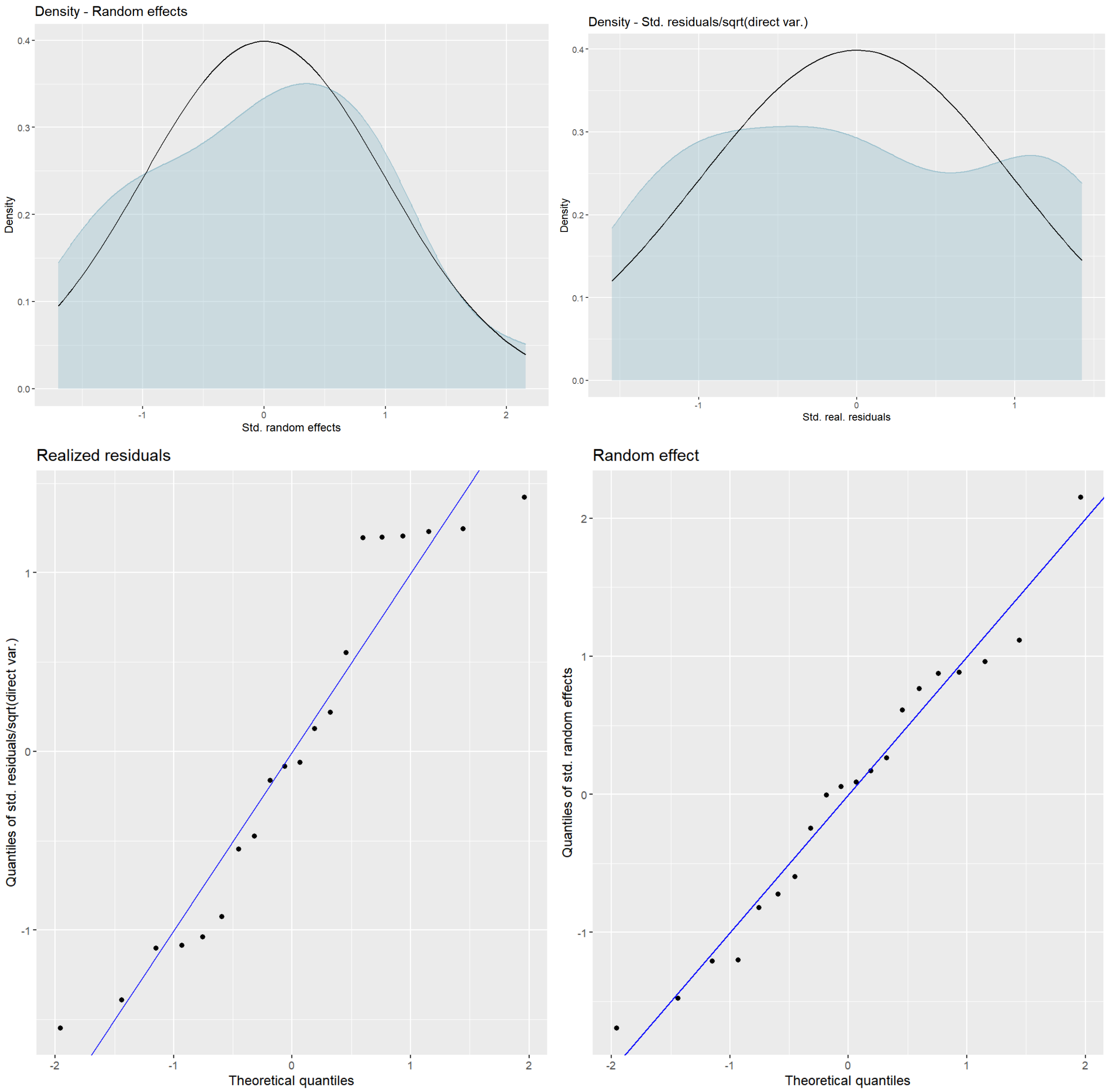


Figure 4.2: Residual Plots of the Reduced Original Scale Model

The original scale model continued to show adherence to the model assumptions of normality. The density plots of the standardized random effects and standardized realized residuals as can be seen in Figure 4.2 suggest a fairly close alignment to with normal distribution. The Q-Q plot supports this by showing a few of the data points veering off the Q-Q line. However, the results of the Shapiro-Wilk test in Table 5.2, on both the standardized ($W = 0.967$, $p = 0.700$) and the predicted random effects ($W = 0.965$, $p = 0.639$) fail to reject normality at 5% level of significance. The results collectively affirm the robustness of the reduced FH model.

Table 4.2: Residuals Tests for the Reduced Original Scale Model

	Standardized Residuals	Random Effects
Shapiro (W)	0.97	0.96
p-value	0.70	0.64

4.5 Comparison of Direct Estimates vs FH Estimates

A key step in evaluating the performance of the SAE models is to compare its model-based estimates with the direct survey estimates. Visually, in Figure 4.3 it can be observed that there are clear gains in precision using the model-based estimates. The boxplots of both the CV and the MSE of the model-based estimates were better in precision compared to the direct unemployment rate estimates. Across the county-level estimates the CV were lower with the model-based estimates.

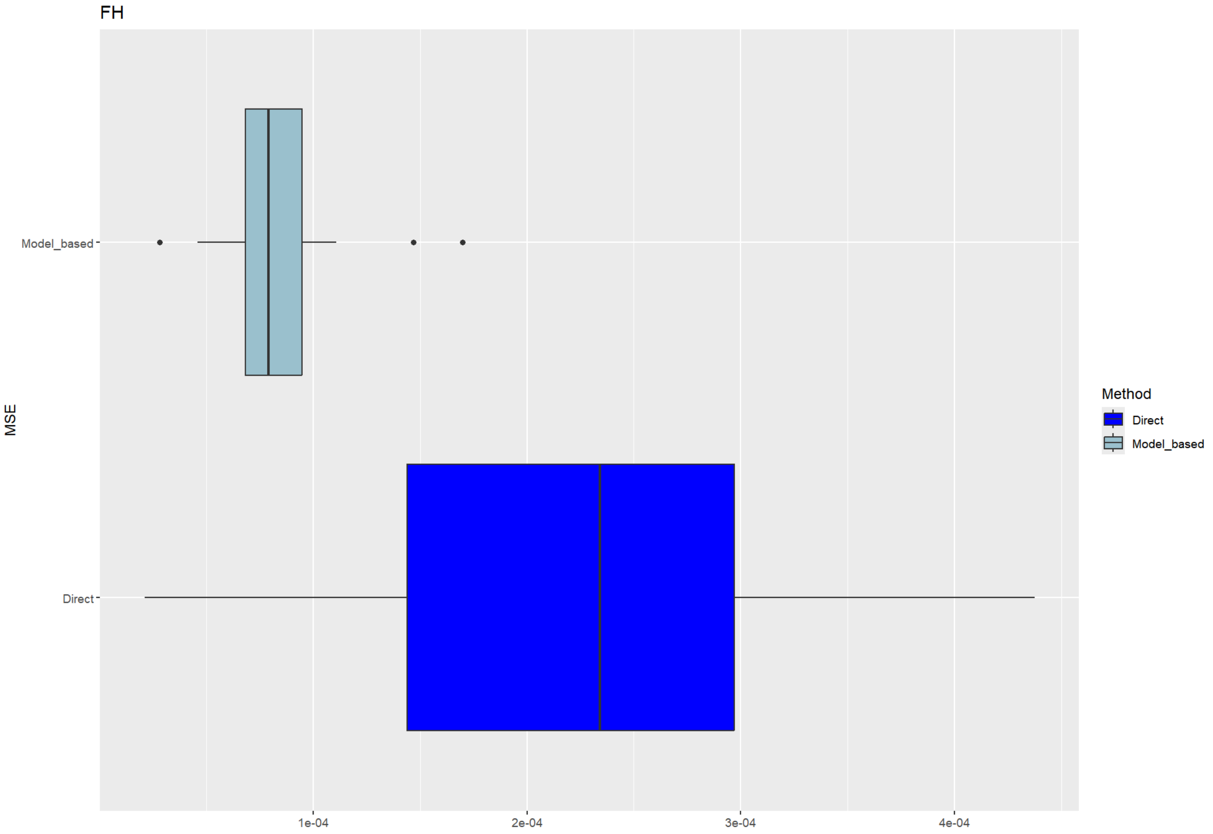


Figure 4.3: MSE Comparison Plots of the Model Estimates vs Direct Estimates

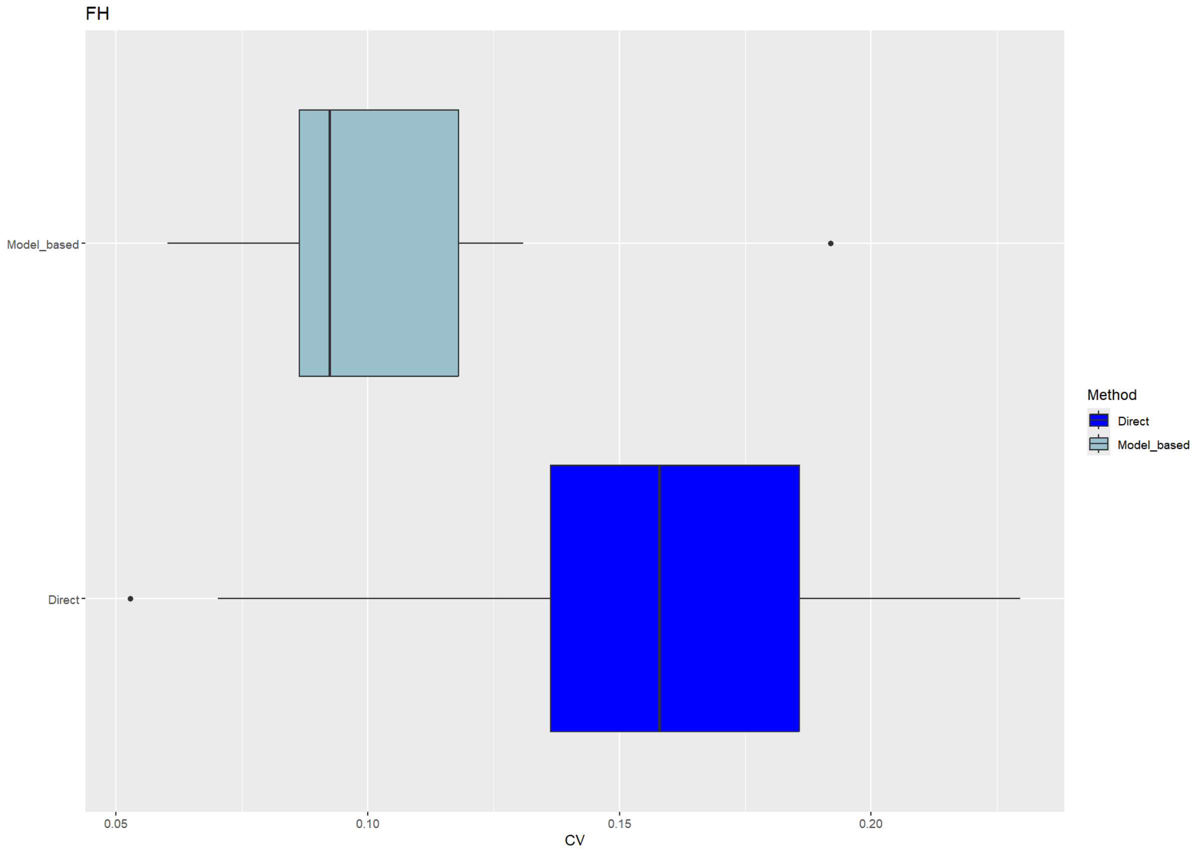


Figure 4.4: CV Comparison Plots of the Model Estimates vs Direct Estimates

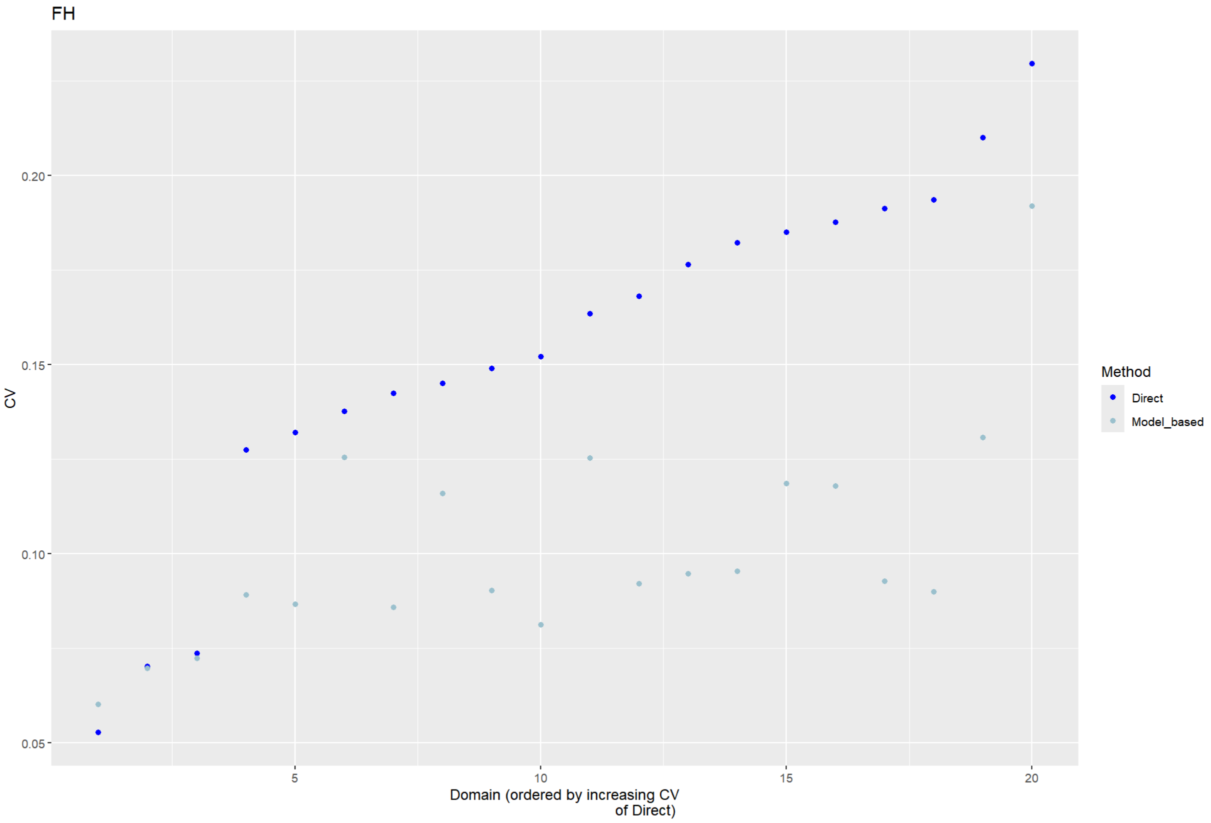


Figure 4.5: ScatterPlot of the Model Estimates vs Direct Estimates

Table 5.1 shows the in the 20 in-sample counties the mean (CV) falls from 0.15 to 0.10 and the median from 0.15 to 0.09, a one-thirds reduction in the relative error. The lowest design-based MSE (Kronoberg, 43.8×10^{-5}) collapses to 11.1×10^{-5} over 70 % efficiency gain. Counties with the least reliable direct numbers benefit most, for instance Blekinge's CV shrinks from 0.14 to 0.09, Västmanland's from 0.15 to 0.08. Where the survey was already precise, the model changes almost nothing; Stockholm's point estimate moves from 8.7 % to 8.8 % and its CV nudges from 0.05 to 0.06, illustrating the FH predictor's data-adaptive shrinkage. The model also extends geographic coverage by providing a fully synthetic estimates for Gotland county (FH rate = 8.2%, MSE = 58.55×10^{-5}) which was dropped due insufficient sample size. While its CV (29.3%) remains high, this demonstrates SAE ability to "borrow strength" for counties with no survey data.

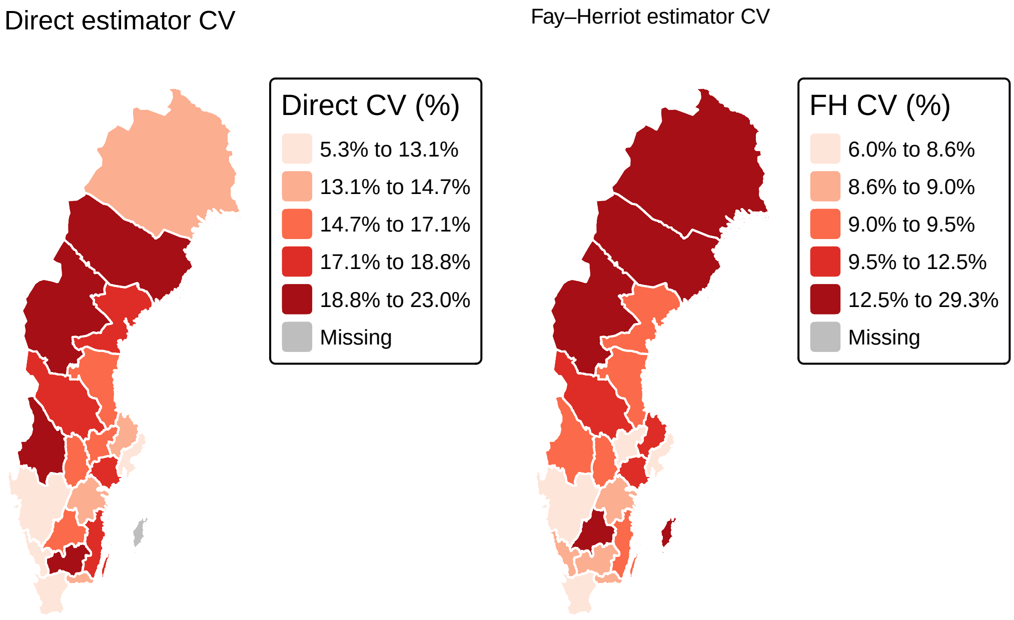


Figure 4.6: CV Comparison of the Direct vs Model-based Estimates

4.6 Discussions

The main objectives of this paper were to provide an application of SAE in estimating county-level unemployment rates in Sweden and to assess the potential of using the geospatial data as auxiliary variables for estimating unemployment rates. The standard FH model, log-transformed and arcsine transformed models were fitted. The information criteria was used in zeroing on the preferred model. The untransformed FH model outperformed the transformed alternatives (AIC = −99.1 versus −9.3 and −41.4), and a stepwise reduction that none significant regressors pushed AIC down to −109.6 while raising R^2 from 0.89 to 0.95. This parsimonious model trimmed the mean coefficient of variation across the 20 in-sample counties from 0.15 to 0.10 and more than halved the widest CVs (Blekinge, Västmanland), while still leaving well-sampled Stockholm unemployment rate estimates, which is data adaptive shrinkage property of the EBLUP estimates. Gotland county received a fully synthetic SAE estimate (8.2% unemployment, CV = 0.29) given that the direct estimates were not reported due to insufficient observations.

An important exploration of this study was to assess the plausibility of the geospatial covariates in SAE modeling. An initial correlation matrix highlighted eight plausible candidates, but step-wise FH estimation distilled that list to four consistently significant covariates; log-elevation ($\beta = -0.03$, p = 0.01), population density ($\beta = 0.02$, p = 0.02), mean slope ($\beta = 0.53$, p = 0.02) and share of urban land cover ($\beta = -0.01$, p = 0.002). Together this quartet captured over 95 % of the between-county variance and produced the higher AIC/BIC improvements, whereas dropping any one of them noticeably weakened the fit. The direction of the effect by the coefficients also mirrored the labor market. Higher altitudes and a more diversified urban footprint coincide with tourism hubs or multi-sector economies that cushion against unemployment, while dense population and rugged terrain can restrict labour mobility or concentrate competition for scarce opportunities. Variables that appeared promising in isolation, NDVI greenness, tropospheric NO_2, nighttime lights fell away once multicollinearity was controlled for, illustrating the importance of iterative modelling when handling rich geospatial stacks. Residual diagnostics (Shapiro–Wilk $p = 0.70$; no heteroskedastic or spatial patterns) confirmed that the final specification met classical FH assumptions, making it ready for production use.

By delivering gains in precision of the unemployment rates, the analysis meet the research objectives. It provides a practical blueprint for NSIs that need reliable small area unemployment figures without enlarging their labour force surveys and demonstrates that carefully curated geospatial data can add measurable value even in a country where administrative sources are already reliable.

Chapter 5

Conclusions & Recommendations

5.1 Conclusions

This thesis demonstrates the possible advantage of integrating geospatial covariates in official statistics via SAE. As demand grows for timely, localized indicators to guide targeted policy interventions, SAE models offer an alternative for producing reliable estimates even where survey samples are sparse. Our findings highlight that a parsimonious selection of readily available geospatial predictors yields simpler, more adoptable models without sacrificing predictive accuracy. By balancing ease of implementation, statistical agencies and policymakers can bridge survey gaps and generate actionable, disaggregated labor-market insights. This thesis thus provides a practical blueprint for enhancing traditional survey methods with SAE techniques and strengthening the evidence base for regional economic policy. However, attention is to be paid to the compromise between the gains in precision and the resources used to obtain these gains. Therefore, this study could be viewed as of high significance to academia in terms of integrating the geospatial sources as auxiliary data with the Swedish LFS and in general the NSIs should conduct a pragmatic cost–benefit assessment before embedding these methods into routine production.

5.2 Limitations

Despite its contributions, the study has two key limitations. First, the model's improvements hinge on the availability and quality of auxiliary data; if important socio-economic or geographic factors are unobserved or only crudely measured, the resulting estimates may be biased or less reliable. Second, the approach was tailored to county-level unemployment in a specific national context, so its performance may not directly generalize to smaller domains or other regions without further adaptation and validation. These limitations suggest that while the model-based methodology is powerful, caution is needed when extending it beyond the study's setting to ensure its assumptions hold and its inputs remain relevant.

5.3 Recommendations

Future studies could explore the effects of spatial correlations using spatial FH models. These models could further make improvements on the precision of the estimates by capturing the spatial dependence. Researchers could also explore the use of spatio-temporal FH models with the geospatial data. One of the challenges here, is the temporal misalignment of the geospatial covariates which leaves the room for future researchers to develop better integration methods while retaining the temporal structure of the geospatial covariates.

References

Battese, G.E., Harter, R.M. and 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), pp. 28–36. Available at: https://doi.org/10.2307/2288915.

Besag, J., York, J. and Mollié, A. (1991) ‘Bayesian image restoration, with two applications in spatial statistics’, *Annals of the Institute of Statistical Mathematics*, 43(1), pp. 1–20. Available at: https://doi.org/10.1007/BF00116466.

Brakel, J.A. van den and Krieg, S. (2015) ‘Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design’, *Survey Methodology*, 41(2), pp. 267–296.

Chambers, R., Salvati, N. and Tzavidis, N. (2016) ‘Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK’, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179(2), pp. 453–479. Available at: https://doi.org/10.1111/rssa.12123.

Datta, G.S. and Lahiri, P. (2000) ‘A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems’, *Statistica Sinica*, 10(2), pp. 613–627. Available at: https://www.jstor.org/stable/24306735.

European Commission (ed.) (2019b) *Guidelines on small area estimation for city statistics and other functional geographies: 2019 edition*. 2019 edition. Luxembourg: Publications Office. Available at: https://doi.org/10.2785/822325.

European Commission (ed.) (2019a) *Guidelines on small area estimation for city statistics and other functional geographies: 2019 edition*. 2019 edition. Luxembourg: Publications Office. Available at: https://doi.org/10.2785/822325.

Fay, R.E. and Herriot, R.A. (1979) ‘Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data’, *Journal of the American Statistical Association*, 74(366a), pp. 269–277. Available at: https://doi.org/10.1080/01621459.1979.10482505.

Hadam, S. *et al.* (2024) ‘Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany’, *Statistical Methods & Applications*, 33(1), pp. 205–233. Available at: https://doi.org/10.1007/s10260-023-00722-0.

Henderson, C.R. (1975) ‘Best linear unbiased estimation and prediction under a selection model’, *Biometrics*, 31(2), pp. 423–447. Available at: https://doi.org/10.2307/2529430.

Jean, N. *et al.* (2016) ‘Combining satellite imagery and machine learning to predict poverty’, *Science*, 353(6301), pp. 790–794. Available at: https://doi.org/10.1126/science.aaf7894.

Kreutzmann, A.-K. *et al.* (2019) ‘The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators’, *Journal of Statistical Software*, 91, pp. 1–33. Available at: https://doi.org/10.18637/jss.v091.i07.

Malec, D. *et al.* (1997) ‘Small area inference for binary variables in the national health interview survey’, *Journal of the American Statistical Association*, 92(439), pp. 815–826. Available at: https://doi.org/10.1080/01621459.1997.10474037.

Marchetti, S. *et al.* (2015) ‘Small Area Model-Based Estimators Using Big Data Sources’, *Journal of Official Statistics*, 31(2), pp. 263–281. Available at: https://doi.org/10.1515/jos-2015-0017.

Mellander, C. *et al.* (2015) ‘Night-Time Light Data: A Good Proxy Measure for Economic Activity?’, *PLOS ONE*, 10(10), p. e0139779. Available at: https://doi.org/10.1371/journal.pone.0139779.

Nandram, B. and Choi, J.W. (2002) ‘A Bayesian analysis of a proportion under non-ignorable non-response’, *Statistics in Medicine*, 21(9), pp. 1189–1212. Available at: https://doi.org/10.1002/sim.1100.

Newhouse, D. *et al.* (2022) ‘Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning’.

Pfeffermann, D. *et al.* (1998) ‘Weighting for unequal selection probabilities in multilevel models’, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1), pp. 23–40. Available at: https://www.jstor.org/stable/2985969.

Prasad, N.G.N. and Rao, J.N.K. (1990) ‘The estimation of the mean squared error of small-area estimators’, *Journal of the American Statistical Association*, 85(409), pp. 163–171. Available at: https://doi.org/10.2307/2289539.

Rao, J.N.K. and Molina, I. (2015) *Small area estimation.* Second edition. Hoboken, New Jersey: John Wiley & Sons, Inc (Wiley series in survey methodology).

Rao, J.N.K. and Yu, M. (1994) ‘Small-area estimation by combining time-series and cross-sectional data’, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 22(4), pp. 511–528. Available at: https://doi.org/10.2307/3315407.

Singh, B., Shukla, G. and Kundu, D. (2005) ‘Spatio-temporal models in small area estimation’, *Survey Methodology*, 31.

Statistics Canada (2024) ‘The Daily — Small area estimates of employment and unemployment rate for self-contained labour areas, November 2021 to June 2024’. Available at: https://www150.statcan.gc.ca/n1/daily-quotidien/240917/dq240917d-eng.htm.

Steele, J.E. *et al.* (2017) ‘Mapping poverty using mobile phone and satellite data’, *Journal of The Royal Society Interface*, 14(127), p. 20160690. Available at: https://doi.org/10.1098/rsif.2016.0690.

Tang, B., Liu, Y. and Matteson, D.S. (2022) ‘Predicting poverty with vegetation index’, *Applied Economic Perspectives and Policy*, 44(2), pp. 930–945. Available at: https://doi.org/10.1002/aepp.13221.

Ugarte, M.D. *et al.* (2009) ‘Estimating unemployment in very small areas’, *SORT-Statistics and Operations Research Transactions*, 33(1), pp. 49–70. Available at: https://raco.cat/index.php/SORT/article/view/144071.

United Nations Statistics Division (2023) *Small area estimation with geospatial data: A primer.* United Nations Statistics Division. Available at: https://unstats.un.org/iswghs/documents/geospatial-data-for-SAE-outline.pdf.

Weiss, D.J. *et al.* (2018) ‘A global map of travel time to cities to assess inequalities in accessibility in 2015’, *Nature*, 553(7688), pp. 333–336. Available at: https://doi.org/10.1038/nature25181.

You, Y. and Rao, J.N.K. (2003) ‘Pseudo hierarchical bayes small area estimation combining unit level models and survey weights’, *Journal of Statistical Planning and Inference*, 111(1), pp. 197–208. Available at: https://doi.org/10.1016/S0378-3758(02)00301-4.

Zhang, J.L. and Bryant, J. (2020) ‘Fully Bayesian Benchmarking of Small Area Estimation Models’, *Journal of Official Statistics*, 36(1), pp. 197–223. Available at: https://doi.org/10.2478/jos-2020-0010.

Appendix

Appendix A: Visual of the Direct Estimates by Swedish Counties

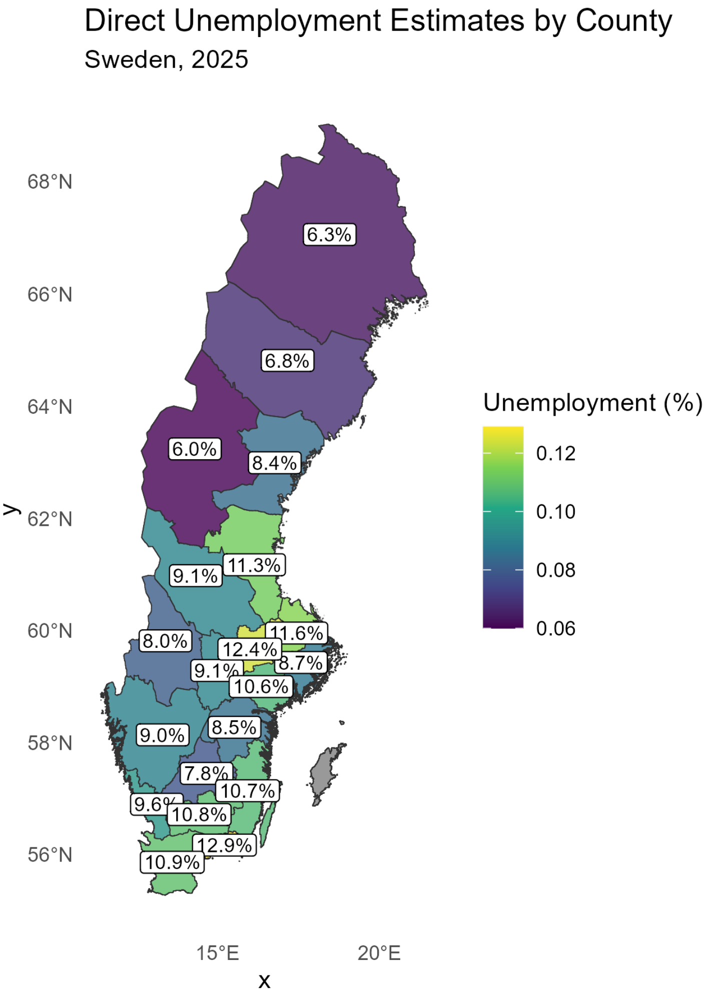
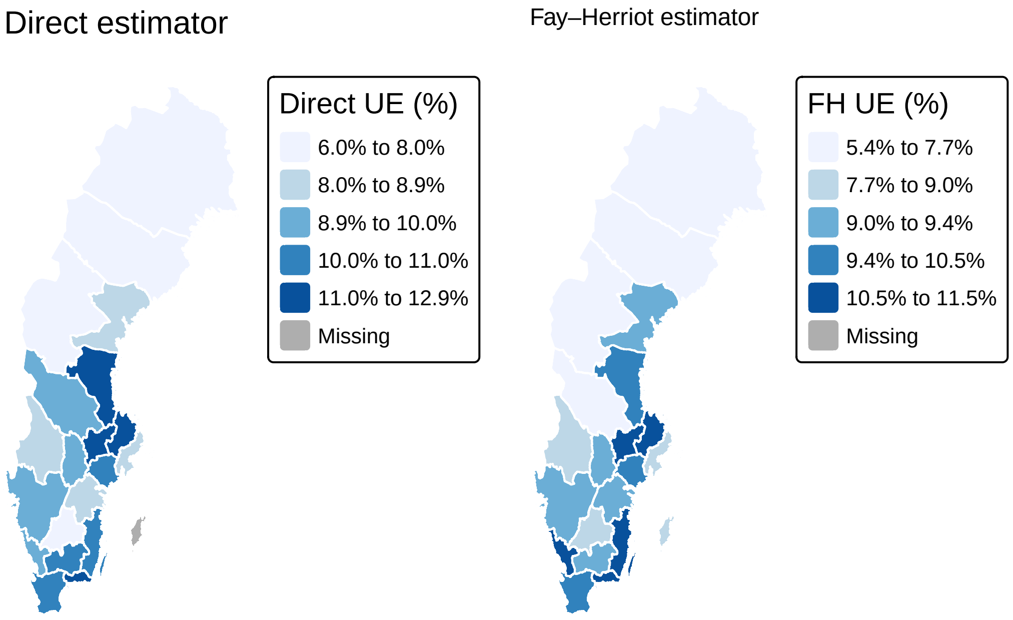


Figure 5.1: Direct Estimates by County

Appendix B: Direct Vs Model-based Estimates

Table 5.1: Direct vs FH Estimates (Reduced Original Scale Model)

County	Direct Rate	Direct MSE (×10)	Direct CV	FH Rate	FH MSE (×10)	FH CV
Stockholm	0.09	2.11	0.05	0.09	2.80	0.06
Skåne	0.11	5.86	0.07	0.10	5.17	0.07
Västra Götaland	0.09	4.40	0.07	0.09	4.58	0.07
Västmanland	0.12	35.64	0.15	0.11	7.86	0.08
Östergötland	0.09	12.60	0.13	0.09	6.61	0.09
Kronoberg	0.11	43.76	0.19	0.09	6.88	0.09
Kalmar	0.11	35.64	0.18	0.11	11.08	0.09
Blekinge	0.13	33.74	0.14	0.11	9.73	0.09
Halland	0.10	14.99	0.13	0.10	8.68	0.09
Värmland	0.08	23.43	0.19	0.09	6.66	0.09
Örebro	0.09	23.43	0.17	0.09	6.94	0.09
Gävleborg	0.11	28.35	0.15	0.10	8.18	0.09
Västernorrland	0.08	23.43	0.18	0.09	7.47	0.10
Uppsala	0.12	28.35	0.15	0.11	16.99	0.12
Södermanland	0.11	39.59	0.19	0.10	14.70	0.12
Dalarna	0.09	28.35	0.19	0.08	8.12	0.12
Jönköping	0.08	16.27	0.16	0.08	9.40	0.13
Västerbotten	0.07	20.41	0.21	0.07	7.93	0.13
Norrbotten	0.06	7.52	0.14	0.07	7.14	0.13
Jämtland	0.06	18.98	0.23	0.05	10.94	0.19
Gotland	NA	NA	NA	0.08	58.55	0.29



Direct estimator

Fay–Herriot estimator

Figure 5.2: Maps of the Direct Estimators vs FH-based estimators

Appendix C: Brown's Correlation Tests

The Brown's test has the following set up:

$$H_0 : \text{EBLUP estimates do not differ significantly from the direct estimates.}$$

Table 5.2: Brown's Test for the Reduced Original Scale Model

W Statistic	Df	p-value
4.39	20	0.9999