Doctoral Programme in Statistics, Optimization and Applied Mathematics

---

Model-based Contributions to Small Area Estimation

María Bugallo Porto

DOCTORAL DISSERTATION

Supervisor

Domingo Carlos Morales González

Co-supervisor

María Dolores Esteban Lefler

Miguel Hernández University of Elche

– 2024 –

Recommended Citation

Bugallo, M. Model-based Contributions to Small Area Estimation, Doctor of Philosophy Thesis. Center of Operations Research, Miguel Hernández University of Elche, Spain, 2024.

This doctoral thesis has been financially supported by

– Grant PROMETEO-2021-063 from the Generalitat Valenciana, Spain.

– Study grant from the Manuel Ventura Figueroa Foundation, Spain.

The Doctoral Thesis entitled "*Model-based Contributions to Small Area Estimation*", carried out by the Doctoral Candidate **María Bugallo Porto**, under the supervision of Prof. **Domingo Carlos Morales González** and Prof. **María Dolores Esteban Lefler**, both from the Department of Statistics, Mathematics and Computer Science and the Center of Operations Research of the Miguel Hernández University of Elche, is presented in the form of a **conventional thesis** with the following indications of quality:

1. M. Bugallo, M. D. Esteban, M. Marey-Pérez, D. Morales (2023). Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain. *Journal of Environmental Management*, **328**. https://doi.org/10.1016/j.jenvman.2022.116788.
2. M. Bugallo, M. D. Esteban, D. Morales (2024). Small area estimation of the proportion of single-person households. Application to the Spanish Household Budget Survey. *SORT-Statistics and Operations Research Transactions*, **48** (1), 125–152. https://doi.org/10.57645/20.8080.02.16.
3. M. Bugallo, M. D. Esteban, T. Hobza, D. Morales, A. Pérez (2024). Small area estimation of labour force indicators under unit-level multinomial mixed models. *Journal of the Royal Statistical Society: Series A*. https://doi.org/10.1093/jrsssa/qnae033.
4. M. Bugallo, M. D. Esteban, M. C. Pagliarella, D. Morales. Model-based estimation of small area dissimilarity indexes: An application to sex occupational segregation in Spain. *Social Indicators Research*, **174**, 473–501. https://doi.org/10.1007/s11205-024-03393-w.

Other publications under review

The following dissertation-related publications of the Doctoral Thesis entitled "*Model-based Contributions to Small Area Estimation*" are being peer-reviewed:

1. M. Bugallo, M. D. Esteban, M. Marey-Pérez, D. Morales. Pattern recognition and modelling of virulent wildfires in Spain. *International Journal of Wildland Fire.*
2. M. Bugallo, D. Morales, N. Salvati, F. Schirripa. Temporal M-quantile models and robust bias-corrected small area predictors. *Journal of Survey Statistics and Methodology.*

Prof. **Domingo Carlos Morales González** and Prof. **María Dolores Esteban Lefler**, both from the Department of Statistics, Mathematics and Computer Science and the Center of Operations Research of the Miguel Hernández University of Elche, certify that the Doctoral Thesis entitled "*Model-based Contributions to Small Area Estimation*" has been carried out under their supervision by the Doctoral Candidate **María Bugallo Porto**.

Having completed the project, and in accordance with the conditions set out in the research plan and the Code of Good Practice of the Miguel Hernández University of Elche, it achieves its intended objectives in a satisfactory manner for public defence as a doctoral thesis.

In Elche, November 4, 2024.

The supervisor: The co-supervisor:

Prof. Domingo Carlos Morales González Prof. María Dolores Esteban Lefler

Prof. **Domingo Carlos Morales González**, Coordinator of the Doctoral Programme in Statistics, Optimization and Applied Mathematics of the Miguel Hernández University of Elche, certifies that the Doctoral Thesis entitled "*Model-based Contributions to Small Area Estimation*" has been carried out under the supervision of the Doctoral Programme by the Doctoral Candidate **María Bugallo Porto**.

Having completed the project, and in accordance with the conditions set out in the research plan and the Code of Good Practice of the Miguel Hernández University of Elche, it achieves its intended objectives in a satisfactory manner for public defence as a doctoral thesis.

In Elche, November 4, 2024.

The coordinator:

Prof. Domingo Carlos Morales González

Acknowledgements

First of all, I am grateful to the Generalitat Valenciana, and in particular, to the Regional Ministry of Innovation, Universities, Science and Digital Society, for its long-term commitment to research, which has provided facilities and resources for this work through the PROMETEO-2021-063 project. I would also like to extend my sincere gratitude to the Manuel Ventura Figueroa Foundation for its generous support. Thanks should also go to the Vice-Rectorate for Research of the Miguel Hernández University of Elche, for funding a research stay abroad to obtain the international mention in the PhD title.

This thesis would not have been possible without the help of my supervisors, Domingo Morales and María Dolores Esteban. Even more, my most special gratitude goes to Domingo, who devoted all his time to help me. He has made it possible to start, advance and finish this project and with it, this lovely stage of my life in Elche. My sincere affection also goes to all my colleagues and fellow PhD students at the Center of Operations Research and, of course, to the professors of the University of Santiago de Compostela, who taught me the beauty of mathematics. Furthermore, I would be remiss if I did not mention my co-authors Tomás Hobza, Manuel Marey Pérez, Maria Chiara Pagliarella and Agustín Pérez, but I am missing two researchers who deserve all my sympathy.

Words cannot express my gratitude for having been accepted at the Department of Economics and Management of the University of Pisa. I cherish all the good times I had during my two stays in Pisa, the hospitality of Nicola Salvati and his enthusiasm for research, shared with Francesco Schirripa, whom I also thank for his warmth and helpfulness. To them and to all the Italian people I met outside academia, I hope that our paths will cross again.

Finally, but no less importantly, I cannot forget to thank my family, my friends, my flatmate Sofía and my partner Eduardo, for all their love, understanding and unconditional support during these intense years.

Table of Contents

Abstract

English abstract

National statistical offices and private institutions are increasingly interested in having information on specific subgroups of the population. The main motivation is to address decision-making more effectively. Survey data are widely used for this purpose and no technical problem arises as long as the sample sizes are large enough to yield direct estimates of acceptable reliability. Otherwise, Small Area Estimation is an effective solution. This thesis contributes to this field using both area-level and unit-level models. First, new zero-inflated mixed models are proposed. Subsequently, the Fay-Herriot model is generalised and the unit-level multinomial logit mixed model is investigated. We predict segregation indexes and unemployment rates, respectively. Finally, the M-quantile regression is generalised to temporal data and the optimal selection of robustness parameters is addressed. In general, fitting algorithms are proposed and model-based predictors and mean squared error estimates are derived. Simulation studies and applications to real data are carried out to analyse the properties and applicability of the new statistical methods.

Keywords: Small Area Estimation; official statistics; indirect estimation; zero-inflated model; Fay-Herriot model; unit-level model; M-quantile model.

Resumen en castellano

Los institutos nacionales de estadística y las instituciones privadas están cada vez más interesadas en disponer de información sobre subgrupos específicos de la población. La principal motivación es abordar la toma de decisiones eficientemente. Las encuestas se utilizan ampliamente con este fin, sin ningún problema técnico si el tamaño muestral es adecuado para producir estimaciones directas veraces. En caso contrario, la estimación en áreas pequeñas es una solución eficaz. Esta tesis contribuye a este campo utilizando modelos tanto a nivel de área como de unidad. En primer lugar, se proponen modelos mixtos inflados en el cero. Además, se generaliza el modelo Fay-Herriot y se estudia el modelo mixto logístico multinomial a nivel de unidad para predecir indicadores no lineales, como índices de segregación y tasas de paro, respectivamente. Finalmente, la regresión M-cuantil se generaliza a datos temporales y se aborda la selección óptima de los parámetros de robustez. En general, se proponen algoritmos de ajuste y se derivan predictores y estimaciones del error cuadrático medio. También se llevan a cabo estudios de simulación y aplicaciones a datos reales para analizar las propiedades y la aplicabilidad de los nuevos métodos estadísticos.

Palabras clave: Estimación en áreas pequeñas; estadística pública; estimación indirecta; modelo inflado en el cero; modelo Fay-Herriot; modelo de unidad; modelo M-cuantil.

List of Abbreviations

Abbreviation	Definition
AR	**A**uto**r**egressive
ARR	**A**ggregated **R**aw **R**esidual
ASR	**A**ggregated **S**tandardized **R**esidual
aZIG	**A**rea-Level **Z**ero-**I**nflated **G**amma
aZINB	**A**rea-Level **Z**ero-**I**nflated **N**egative **B**inomial
aZIP	**A**rea-Level **Z**ero-**I**nflated **P**oisson
BE	**Be**rnoulli
BI	**Bi**nomial
BLUP	**B**est **L**inear **U**nbiased **P**redictor
BMQ	**B**ias-corrected **M**-**q**uantile
BP	**B**est **P**redictor
BTMQ	**B**ias-corrected **T**emporal **M**-**q**uantile
c.d.f.	**C**umulative **d**istribution **f**unction
CI	**C**onfidence **I**nterval
CV	**C**oefficient of **V**ariation
DSI	**D**uncan **S**egregation **I**ndex
EBLUP	**E**mpirical **B**est **L**inear **U**nbiased **P**redictor
EBP	**E**mpirical **B**est **P**redictor
EMP	**E**mpirical **M**arginal **P**redictor
FH	**F**ay-**H**erriot
FH2	Two-fold **F**ay-**H**erriot
FH3	Three-fold **F**ay-**H**erriot
GA	**Ga**mma
GFFS	**G**eneral **F**orest **F**ire **S**tatistics
GLMM	**G**eneralized **L**inear **M**ixed **M**odel
GVF	**G**eneralized **V**ariance **F**unction
HBS	**H**ousehold **B**udget **S**urvey
i.i.d.	**I**ndependent and **i**dentically **d**istributed
IRLS	**I**terative **R**e-weighted **L**east **S**quares
LB	**L**ower **B**ound

Abbreviation	Definition
LCS	**L**iving **C**onditions **S**urvey
LFS	**L**abour **F**orce **S**urvey
LMM	**L**inear **M**ixed **M**odel
LogN	**Log**-**N**ormal
MAD	**M**edian **A**bsolute **D**eviation
ML	**M**aximum **L**ikelihood
MP	**M**arginal **P**redictor
MQ	**M**-**q**uantile
MQ2	Two-fold **M**-**q**uantile
MQ3	Three-fold **M**-**q**uantile
MSE	**M**ean **S**quared **E**rror
NB	**N**egative **B**inomial
NER	**N**ested **E**rror **R**egression
NUTS	**N**omenclature of **T**erritorial **U**nits for **S**tatistics
p.d.f.	**P**robability **d**ensity **f**unction
PI	**P**rediction **I**nterval
PO	**Po**isson
PQL	**P**enalized **Q**uasi-**L**ikelihood
RMSE	**R**oot **M**ean **S**quared **E**rror
RRMSE	**R**elative **R**oot **M**ean **S**quared **E**rror
RR	**R**aw **R**esidual
RSPE	**R**elative **S**quared **P**rediction **E**rror
SAE	**S**mall **A**rea **E**stimation
SHBS	**S**panish **H**ousehold **B**udget **S**urvey
SLCS	**S**panish **L**iving **C**onditions **S**urvey
SLFS	**S**panish **L**abour **F**orce **S**urvey
SR	**S**tandardized **R**esidual
TMQ	**T**emporal **M**-**q**uantile
TWMQ	**T**ime-**W**eighted **M**-**q**uantile
UB	**U**pper **B**ound

Preface

Modern societies are changing faster and faster and the scientific community must respond appropriately to these new dynamics. As of today, it is safe to say that there are two international objectives whose interests overlap.

On the one hand, there is a growing and widespread governmental concern about issues of discrimination, equity and disparity, as well as environmental challenges. It stands to reason that socio-economic indicators, including unemployment rates and household distribution statistics, are no less popular as decision-making aids. Added to this, climate change and its consequences are an increasingly worrying problem and all indications are that this trend will continue, with potentially noticeable changes, e.g. in fire behaviour. While wealthy western countries are better equipped to tackle the problems of global warming, specific solutions need to be provided and implemented extensively. Additionally, developed countries are promoting fair treatment and legal protection for women and minority groups, and governments are looking for places where systemic discrimination occurs.

The main evidence of the current commitment to address all these issues is the 2030 Sustainable Development Agenda (United Nations, 2015), set by the United Nations General Assembly, and its 17 Sustainable Development Goals. At its core, it commits to "*no one left behind*" and called for more granular and better-quality statistical results to measure specific indicators for different population subgroups (areas or domains). Namely, disaggregated by geographical area (e.g. regions, provinces, municipalities, health service areas), sex, age group or citizenship. Disaggregated statistics are also useful for allocating resources and developing programmes that target disadvantaged territories or population subgroups. In a more and more divided world, any project to reduce inequalities, no matter how small, is welcome. No one could deny the existence of stereotypes and their impact on an individual's opportunities and experiences in education, employment and social interactions.

On the other hand, although we live in an age of information, there is still a need for cost-effective survey design. In this context, National Statistical Offices periodically collect survey data to facilitate targeted policy-making. In the case of survey-based estimates, the granularity of the data implies that the survey design, often resource-constrained, accommodates the level of aggregation of the population and adequately represents each population subgroup in the sample. However, the latter is rarely the norm and direct survey estimates –based only on the sample data in the area– for subgroups with small sample sizes lead to unacceptably high variability. In other words, they are designed to estimate indicators of interest in large geographical areas or broad demographic communities, but not for minority subgroups. It is not always feasible to require a certain sample size for many small areas. Budgetary con-

straints and a lack of planning in the initial design are the most common reasons for small or even zero sample sizes, prompting the need for further research.

Fortunately, Small Area Estimation (SAE) methods provide more reliable granular level estimates by "*borrowing strength*" from auxiliary variables, information from other subgroups and underlying dependency structures. As a result, indiscriminate increases in survey sample sizes are avoided. This will allow better use of existing data to, among other advantages, identify smaller geographical areas, where development needs are greatest, and improve resource allocation. The term "*small area*" is commonly used to refer to a geographical area or a subgroup of the population defined according to some combination of socio-demographic characteristics but where, in any case, direct estimation is not accurate enough due to the smallness of the sample size. For instance, if a survey is designed to obtain precise direct estimates at national level and results disaggregated by region, or for a particular minority group, are of interest, these unplanned estimation domains are called small areas. It is therefore desirable, but also necessary, to use more sophisticated prediction tools.

Research in this field can be of great help to policy makers in deciding where to implement effective policies based on factual information. In addition, new approaches to raising awareness of extreme fire events need to be further explored. This thesis contributes to SAE and to the aforementioned concerns from different perspectives, but always guided by the current demands of our societies and the scientific interests of the national statistical offices. All new methods will be extensively studied by simulations, and illustrated by real problem applications.

Chapter 1

Introduction

National statistical offices design surveys to provide a cost-effective way of collecting data and to obtain accurate estimates at a given level of aggregation. However, disaggregated statistics facilitate more effective targeting of decision-making, but obviously require more information to adequately represent each population subgroup in the sample (Rao, 2003). The importance of rich and granular data should therefore not be undervalue (see Rao and Molina (2015), Chapter 1.2). If the aim is to ensure valid inference on specific subgroups of the population, where the portion of available data is large enough, we can accurately estimate domain characteristics using direct estimators, i.e. relying only on data from sample units in the domain of interest. The most commonly used direct estimators are the Horvitz and Thompson (1952) and the Hájek (1971) estimators. Although it is difficult to establish general conditions under which one of them is better than the other, Sarndal et al. (1992) present some evidence in favour of the Hájek estimator. This has motivated our choice of the Hájek estimator, which is why it will be referred to hereafter as either the Hájek estimator or the direct estimator.

Unfortunately, unplanned domains can be of any size, making estimation more difficult. In small areas, indirect estimation techniques based on statistical modelling must be used. SAE addresses this challenge by relying on auxiliary variables, data from other domains and underlying dependency structures. A common feature of model-based predictors is that they do not aim to reproduce the trend of direct estimators, but to smooth them and provide more accurate results. Inference based on information from other domains and auxiliary variables is widely used in the literature, as it is expected to be more efficient (Singh et al., 1994). Nonetheless, it is important to bear in mind that they might add a certain degree of subjectivity to the results. It is therefore of great importance for a good mathematical modelling that the selection of the auxiliary information is done carefully (Humi, 2017). In most cases, data to improve direct estimates are selected from the available sources, such as previous or auxiliary surveys and official census records.

1.1 Literature review

The following is an overview of the state of the art. SAE uses linear mixed models (LMM) and generalized linear mixed models (GLMM), which can be fitted to either unit-level or area-level data, and then derives predictors from them. This is the usual way to incorporate additional information from other domains, auxiliary variables and hierarchical, spatial or temporal dependency structures (Singh et al., 1994). The use of valid statistical models provide small area predictions with greater accuracy, but bias can result from an incorrect model. As a matter of fact, statistical modelling is richer when random effects are also introduced to account for more complex correlation patterns (Jiang and Lahiri, 2006). An excellent review provided by Morales et al. (2021) covers a wide range of statistical models for SAE based on the traditional approach governed by the Gaussian law of errors. Rao and Molina (2015) and Pratesi (2016) are also comprehensive and up-to-date accounts.

Depending on the type of data available, SAE models fall into two broad categories: area-level models, which relate design-based direct estimates to area-specific covariates, and unit-level models, which use individual survey responses as the target variables rather than direct estimates. The former are needed when unit-level census data are not available.

Area-level models have the advantage of being able to easily incorporate auxiliary variables from statistical sources other than the sample. Typically, these models include area-specific random effects to account for the between-area variability that is not explained by the covariates. Among the precursors of area-level SAE models, Fay and Herriot (1979) (FH) suggest estimates based on the best prediction method (Henderson, 1975), i.e., empirical best linear unbiased predictors (EBLUP) of linear domain indicators and empirical best predictors (EBP) of non-linear domain indicators. A basic example is the estimation of domain totals, means and proportions using EBLUPs. It stands to reason, however, that it is crucial to decide whether it is worth using a mixed effects model rather than a simpler fixed effects model. This led Marhuenda et al. (2016) to come up with tests for the variance parameter in the FH model. Also recently, Reluga et al. (2021) propose simultaneous inference methods for unit-level binomial (BI), area-level Poisson (PO)-Gamma (GA) and area-level PO Log-Normal (LogN) mixed models, and Reluga et al. (2023) for mixed parameters assuming a LMM.

An alternative approach to incorporating spatial information into a small area regression model, and which does not necessarily include random effects, is to assume that the model parameters themselves vary spatially across the region of interest. Geographically Weighted Regression (Brundson et al., 1996) models this spatial variation by using local rather than global parameters. Maiti et al. (2016) have proposed a functional mixed-effects model for SAE. Indeed, they fit a linear mixed-effects model with varying coefficients, where the varying coefficients are semi-parametrically modelled by B-splines, to area-level data.

Regarding generalizations of the FH model, some temporal extensions have been given by Pfeffermann and Burck (1990), Rao and Yu (1994), Ghosh et al. (1996), Datta et al. (2002) and Singh et al. (2005). For estimating proportions, Esteban et al. (2012), Marhuenda et al. (2013, 2014) and Morales et al. (2015) have proposed predictors based on variants of the FH model. Chapter 19 of Morales et al. (2021) describes the bivariate FH model. The extension to multivariate FH models, with unstructured random effect covariance matrix,

increases the number of variance component parameters. Indeed, multivariate FH models have been studied by Huang and Bell (2004), González-Manteiga et al. (2008), Porter et al. (2015) and Benavent and Morales (2021), to cite just but a few. Benavent and Morales (2016) introduce multivariate FH models with covariance patterns between the components of the vector of random effects, but they only considered the estimation of univariate indexes. Arima et al. (2017) and Burgard et al. (2021) study multivariate FH models with error-measured covariates. Krause et al. (2022b) propose penalized multivariate FH models and present an application to alcohol consumption data. Esteban et al. (2020) have adapted a trivariate FH model for the estimation of small area compositions. Cabello et al. (2024) applies a multivariate FH model to log-ratio transformations for the small area prediction of divergence indexes. On the other hand, some extensions of the FH model that ensure estimates in the range $[0, 1]$ have been proposed in the literature. For example, LMMs with appropriate transformations have been proposed by González-Manteiga et al. (2002) and Beta regression models by Janicki (2020).

As for the nesting level, the two-fold FH (FH2) model has been introduced by Rao and Yu (1994) and studied by Esteban et al. (2012), Marhuenda et al. (2013) and Morales et al. (2015), among others. The model is adapted to area-level data indexed by areas and subareas. The three-fold FH (FH3) model (Marcis et al., 2023) can further describe data structured in areas, subareas and time periods or subsubareas. This is the case of the employment data used to estimate sex segregation by province, occupational sector and time period. Under the FH3 model, Krenzke et al. (2020) have estimated adult literacy of US counties and Cai and Rao (2022) have studied some variable selection methods.

Based on area-level multivariate LMMs, Erciulescu and Fuller (2013) derive small area predictors for the mean of a BI random variable. Chambers et al. (2016) develop semiparametric SAE for binary outcomes with application to UK unemployment data. Under area-level PO, BI, Negative Binomial (NB) and multinomial mixed models, count and proportion predictors have been introduced by Boubeta et al. (2016a, 2017, 2023), Burgard et al. (2021, 2022), Krause et al. (2022a,b) and Díz-Rosales et al. (2023), among others. As for the computational limitations of the PO-GLMMs, but with a unit-level approach, Berg (2022) has shown that the conjugate form of the GA-PO model allows for computationally light estimation and prediction procedures. Faltys et al. (2022) introduce a general area-level model-based approach based on GLMMs. Overall, most of the contributions in the above non-exhaustive collection of relevant papers have in common that they propose area-level SAE methods for predicting domain proportions, totals and counts. However, none of the papers cited above deal with data with excess zeros. This is a partial step that we address in this thesis.

One possible solution is to fit a FH model after a transformation and then apply the methodology of Berg and Fuller (2012) to obtain a non-zero variance estimate when the observed value is zero. Another idea is to consider models in which the probability of the target variable is modified from that which would correspond to a given probability distribution. Zero-inflated models play an important role because of their flexibility.

As a precursor, Hall (2000) studies zero-inflated BI and PO regression with random effects: a case study. The focus is on modelling, not on deriving predictors with desired theoretical properties or estimating the mean squared error (MSE). Pfeffermann et al. (2008) consider situations where the target response value is either zero or an observation from a continuous

distribution. A typical example analyzed in the paper is the assessment of literacy proficiency, where the possible outcome is either zero, indicating illiteracy, or a positive score measuring the level of literacy. They apply a unit-level mixture between zero and a multi-level LMM. Chandra and Chambers (2011) address the modelling of skewed data in the presence of zeros in small areas. Chandra and Sud (2012) propose a unit-level mixture between zero and a LMM on the log-scale. The aim is to estimate the domain mean of a continuous variable y when the census y-vector contains a substantial proportion of zeros. However, the EBP methodology is not applied. Anggreyani et al. (2015) address the estimation of infant mortality in small areas using a zero-inflated area-level PO mixed-effects model, but they do not propose EBPs either. Krieg et al. (2016) and Santi et al. (2019) have conducted simulation experiments for unit-level mixtures between zero and a nested error regression (NER) model under a Bayesian approach. Hartono et al. (2017) deal with area-level zero-inflated BI models, with an application to unemployment data in Indonesia. Last but not least, Datta and Mandal (2015) and Sugasawa et al. (2017) propose uncertain random effects, which are expressed as mixtures of a normal distribution and a one-point-at-zero distribution.

Although the scope of this thesis is mainly in the field of SAE, many methods are potentially applicable to environmental studies and, in particular, to the modelling of forest fire data. Without being exhaustive, some recent contributions related to our research are listed below. To provide predictions of fire counts by forest area, Boubeta et al. (2019, 2016a) and Ríos-Pena et al. (2017) have proposed PO mixed models and binary structured additive regression models, respectively. In a related vein, Rodrígues et al. (2014) and Ríos-Pena et al. (2015) have applied logistic regression models to address the presence or absence of forest fires. In this context, it is common to deal with target variables that contain more zeros than would be expected if the data-generating process came purely from a standard probability distribution (Feng and Dean, 2012; Feng et al., 2020). In fact, the zeros are sometimes not really zeros at all, but very small values that were not observed. Some engineers and statisticians have therefore modelled this peculiarity of the data.

For count data in medical and environmental studies, but without the inclusion of random effects, recent research on zero-inflated models is quite extensive. Namely, fixed effect zero-inflated regression models describe the effects of air pollutants on hospital admissions (Cengiz and Terzim, 2012), analyse the occurrence of fires, which are likely to be scattered and often have an excess of zero counts, or investigate the occurrence and burned area of forest fires using a zero-one-inflated structured additive beta regression model. Specifically, Ríos-Pena et al. (2018) and Viedma et al. (2018) for the forest fires in Spain and Tan et al. (2021) for the Indonesian scenario. Eklund et al. (2022) have used the same technique to investigate the effect of Covid-19 on the increase in arson activity in Madagascar's protected areas. Finally, it has been found that spatial correlation models and Pattern Recognition techniques are used to model burned areas (Moanga et al., 2020; Pereira et al., 2015). Hand in hand with this, as it also concerns the prediction of burned area, Boubeta et al. (2016b) propose two semi-parametric time series models. Part of this research has been motivated by applications to forest fire data, proposing new predictors and risk measures derived from zero-inflated NB and GA mixed models.

On a slightly different topic, most models use cross-sectional data, but a large number of surveys are repeated over time. LMMs and GLMMs using temporal information are therefore

quite useful, as the recent past is often very informative in explaining current patterns. This issue has been explored extensively in longitudinal studies using biological or medical data. However, the use of temporal models in SAE is more recent, as data are often available for many small areas simultaneously, but possibly only for a few time periods. The main idea is to jointly use data from all domains in a given time period together with relevant historical information. The task could be to introduce vector autoregressive (AR) multivariate distributions for the domain-time random effects. You and Rao (2000) and Datta et al. (2002) used the Rao-Yu model (Rao and Yu, 1994), but replaced the AR(1) process with a random walk. Datta et al. (1999) envisaged a similar model but added extra terms to the linking models to reflect seasonal variation in their case study. Pfeffermann and Burck (1990) considered a model with AR(1) time-varying random slopes. A second challenge is to incorporate multivariate spatial correlation structures. For example, random effects may follow a multivariate spatial or conditional autoregressive model. Benavent and Morales (2021) have taken the first steps in this direction. A minor criticism is that these models rely on strong distributional assumptions and it is also necessary to formally specify the dependence structure of the random effects. They are inflexible, fully parametric models.

The scientific literature also offers many contributions to SAE based on unit-level models. Chapter 7 of Rao and Molina (2015) reviews some commonly used unit-level small area models, although the most popular techniques are based on generalisations of the NER model. As a matter of fact, unit-level models are very promising and have a high predictive capability if auxiliary information is available (Parker et al., 2023a,b). In practice, however, it is common not to have any supporting census files, having to limit ourselves to ANOVA-type models. It is expected that the methodology loses strength but, as we will discuss below, they are certainly no less important. In addition, the lack of administrative records or supporting files is in itself a major challenge. It must be said that national statistical offices have censuses and/or administrative files, so they are able to use auxiliary variables measured without error. However, the access is often restricted, so the scientific community is forced to fit measurement error models (Battese et al., 1988) or to use estimates of the auxiliary information as population values. As for the former, measurement error models are rather sophisticated because they are not LMMs and their study should be investigated elsewhere. Hariyanto et al. (2018) provides a comprehensive and up-to-date account of these models in the context of SAE. On the other hand, Marchetti et al. (2018) mention the possibility of using area-level auxiliary variables (i.e. contextual variables) to fit unit-level models.

Among the most outstanding unit-level contributions, Rao and Molina (2010) have proposed EBPs based on NER models. Herrador et al. (2011), Marhuenda et al. (2017), Guadarrama et al. (2021) and Esteban et al. (2022a,b) have modified NER models and extended the EBP approach to two-fold, temporal and multivariate regression. Hobza and Morales (2016) and Hobza et al. (2018) apply unit-level logit mixed models to estimate small area poverty rates; and Morales and Santamaría (2019) propose unit-level temporal LMMs with an AR(1) type time correlation structure for the random effects. Ranalli et al. (2018) study benchmarking procedures for unit-level logit models. In addition, Marino et al. (2019) propose a semiparametric approach for unit-level models and Lombardía et al. (2021) define a new unit-level nested structure adapted to model the gender pay gap by economic activity.

In short, the EBP methodology can be applied to predict domain indicators defined by

non-linear transformations of means, totals and proportions. They represent alternative methods to the area-level model-based approach. As a foretaste, EBPs based on unit-level models are also part of one of the research lines of this thesis. They will be used to predict, among other labour indicators, unemployment rates. At this regard, the estimation of rates involves the estimation of numerators and denominators, so it is advisable to use predictors based on multivariate models. In what follows, we will cite some contributions about multivariate model-based prediction of labour indicators in small areas. Datta et al. (1999) provide hierarchical Bayes estimates of unemployment rates for USA states. Molina et al. (2007) consider area-level multinomial logit mixed models with a common random effect for $q = 3$ categories, and propose model-based predictors of labour proportions and rates. Saei and Taylor (2012) treat the same problem by using a multinomial model for $q = 3$ categories and two dependent random effects. López-Vizcaíno et al. (2013) consider multinomial models for $q \geqslant 3$ categories and $q - 1$ independent random effects. López-Vizcaíno et al. (2015) propose multinomial mixed models for temporal data. They derive algorithms to compute penalized quasi-likelihood (PQL) estimators, opening the way for research on maximum likelihood (ML) and PQL estimators of small area ratio indicators. These authors only present plug-in predictions. Esteban et al. (2020) introduce an area-level compositional mixed model and give predictors of domain proportions of people in the four categories of the variable labour status: under 16 years of age, employed, unemployed and inactive.

Nevertheless, the modelling of unit-level multi-category outcomes, and the subsequent construction of small area predictors, has been scarcely studied. Dawber et al. (2022) derive robust predictors based on multinomial M-quantile (MQ) and expectile regression models. Esteban et al. (2023) deal with unit-level compositional data and derive predictors of small area average compositions under multivariate NER models. The development of EBPs or plug-in predictors based on unit-level multinomial logit mixed models is still to be done.

Moving on to a different approach, the demand for results unaffected by outliers in small areas has encouraged the development of robust inference techniques for SAE. It stands to reason that outlier observations can significantly affect the estimation of population parameters, even more so in the context of SAE. Sinha and Rao (2009) have addressed this issue from the perspective of LMMs and Ghosh et al. (2009) studied robust procedures using Bayesian methods. A recent book by Yi and Nordhausen (2023) sheds new light on robust statistics and, among other aspects, pays special attention to bias calibration for robust estimation in small areas (Ranjbar et al. (2023), pp. 365–394).

On the other hand, both quantiles (Koenker and Bassett, 1978) and expectiles (Newey and Powell, 1987) have been extended to conditional distributions to provide quantile and expectile generalizations of the conventional regression models based on the Gaussian law of errors. As an alternative to the frequentist approach, the pioneering paper by Chambers and Tzavidis (2006) is a tipping point for research in new SAE unit-level models and predictors. The idea proposed by these authors is to non-parametrically capture the variability of the population, beyond what is explained by the covariates, using the so-called MQ coefficients (Breckling and Chambers, 1988). Generally speaking, the new approach avoids distributional assumptions as well as problems associated with the specification of the random effects, allowing differences between areas to be characterised by the variation in area-specific MQ coefficients. For a more in-depth understanding of this methodology, see a review by Dawber and Chambers (2019).

Notable contributions following this idea include Tzavidis et al. (2008), Salvati et al. (2012), Tzavidis et al. (2014), Marchetti et al. (2018) and Schirripa Spagnolo et al. (2021). The methodology has been applied to predict, among other indicators, poverty rates and labour indicators. MQ models have also been used to estimate acidity in northeastern US lakes (Pratesi et al., 2008) and in an analysis of temporal gene expression data (Vinciotti and Keming, 2009). In addition, Chambers et al. (2012, 2016) propose MQ regression models adapted to binary data. On top of that, Tzavidis et al. (2014), Chambers et al. (2014b) and Chandra et al. (2017) derive robust small area predictors for counts. Tzavidis et al. (2010) studied robust prediction of small area means and distributions based on MQ models. One of the most prevalent and shared features of all previous studies is robustness.

Apart from applied contributions, theoretical developments of MQ models have been made by Bianchi and Salvati (2015), Alfo et al. (2017) and Bianchi et al. (2018).

1.2 Objectives and scope

The scientific proposal of this thesis is motivated by the need to map, with a sufficient level of detail, complex socio-economic indicators derived from variables measured in public databases. This will allow a better understanding of our societies and, ultimately, the identification of the most vulnerable areas. The scope of application is limited to SAE techniques, but the methodological developments are applicable to other fields of statistics and involve several branches. Namely, survey sampling, finite population inference, asymptotic theory and simulation and bootstrap methods. The mathematical research develops statistical models, derives fitting algorithms for estimating model parameters, studies asymptotic and inferential related issues, builds model-based predictors of indicators at population and small area-level and calculates error measures. It also implements the software for the production, mapping and interpretation of complex socio-economic variables.

A variety of methods are available for SAE, but there are still many avenues to be investigated. The first line of research pursues the development of model-based statistical techniques for the small area prediction of indicators dependent on zero-inflated variables based on probabilistic frameworks. This is motivated by the fact that it is common in scientific and technical studies to find count data with many zeros (Chapter 9 of Zuur et al. (2009); Michael and Thomas (2016)). It should be stressed that zero-inflated outcomes are quite common in forestry databases, where excess zeros indicate undetectable events. The second line of research is guided by the seminal paper by Fay and Herriot (1979), generalising it into three levels of nesting and addressing the prediction of Duncan Segregation Indexes (DSI) (Duncan and Duncan, 1955). The third line of research proposes unit-level multinomial logit mixed models to analyse employment data in small areas. The task is to derive fitting algorithms and develop EBPs of multivariate linear and non-linear indicators, such as unemployment rates. To estimate MSEs in all the above situations, we have introduced parametric bootstrap algorithms by following Hall and Maiti (2006) and González-Manteiga et al. (2008, 2010). In a guide for practitioners and researchers, Chernick (2007) provide a detailed, multidisciplinary coverage of bootstrap methods.

The thesis concludes with the study of new models and predictors based on the MQ re-

gression approach to SAE, developing for the first time in the literature temporal MQ linear models. As usual, our proposal can be considered as a robust-projective approach based on plug-in robust prediction, i.e. the optimal, but outlier-sensitive, parameter estimates are replaced by outlier-robust versions. Unfortunately, although these methods usually lead to a low prediction variance, they may also introduce an unacceptable prediction bias (Chambers, 1986; Chambers et al., 2014a; Dongmo-Jiongo et al., 2013). To address this issue, we derive robust bias-corrected predictors, although the introduction of a bias correction term may increase the variability of the corrected versions. In general, the robust bias-corrected predictor is obtained by incorporating a second influence function that depends on a tuning constant, usually called the robustness parameter. The selection of this parameter is crucial as it allows a trade-off between bias and variance. Against this background, the optimal selection of the robustness parameter for bias correction in MQ models is a theoretical contribution of this thesis, exploring its applicability in outlier detection.

1.3 Materials and methods

1.3.1 Computing resources

The experimental part of the project was developed in the Cluster of Scientific Computing of the Miguel Hernández University of Elche (http://ccc.umh.es; accessed on: November 4, 2024), which provides a high level of availability and performance, both in terms of intensive computation and storage. In addition, hardware and software resources of the University Research Institute "Center of Operations Research" of the Miguel Hernández University of Elche have been used. The computer code has been implemented in the programming language `R` (R Development Core Team, 2024) and may be provided on request. Unfortunately, the methods are not yet available in a `R` package to the user. In any case, the scientific publications that are part of this doctoral thesis contain links to online and open access repositories, where the corresponding code and databases are stored.

1.3.2 Sources of information and databases

In addition to the theoretical foundations that must underpin any statistical method, its success in simulation experiments and its application to real data must also prove its worth. Although the applicability of the proposed methods is of a general nature, the case studies focus on the Spanish context. In particular, our contributions have been applied to forest fire data and to Household Budget Surveys (HBS), Living Conditions Surveys (LCS) and Labour Force Surveys (LFS). Supporting information from official population censuses has also been used. Regarding the former, data were obtained from (i) the General Forest Fire Statistics (GFFS), the national agency of the European Forest Fire Information System, which provides services related to forest fires in Spain (https://effis.jrc.ec.europa.eu/; accessed on: November 4, 2024). GFFS data after 2015 are being cleaned and are not available for research purposes; (ii) AEMET: the Spanish Meteorological Agency (https://www.aemet.es; accessed on: November 4, 2024), state agency of the Government of Spain.

For the analysis of the forest fire data, it is useful to describe the study region. Spain is a country located in the Iberian Peninsula, in southwestern Europe, bordering North Africa and partly surrounded by the Mediterranean Sea. It has two archipelagos, the Canary Islands and the Balearic Islands, and two autonomous cities, Ceuta and Melilla (Figure 1.1). In view of their respective insular and urban conditions, it is customary to treat the archipelagos separately and to omit both cities from forestry studies. In territorial terms, the province is an administrative demarcation with competences in environment and forest fire management. According to the Spanish Ministry for Ecological Transition and the Demographic Challenge, the provinces can be grouped based on their fire regime into three main regions: Northwest Spain, Mediterranean Coast and Peninsular Center (MITECO, 2023).

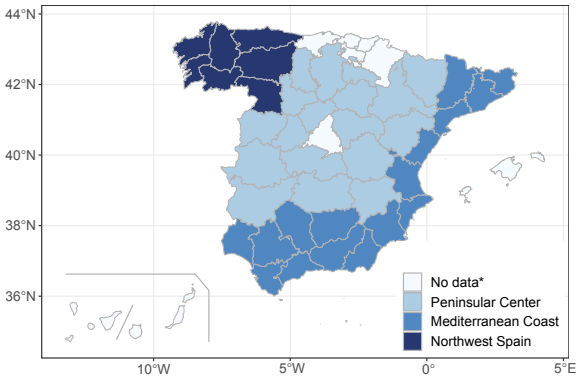


Figure 1.1: Political map of Spain at provincial level (NUTS3). Ceuta and Melilla are not included. Territorial divisions are taken from MITECO (2023). The "No data" category refers to provinces that do not provide (or do not have reliable) data on the distance between fires and human buildings for our forest fire studies (see Section 2.4.3). The northern ones belong to Northwest Spain and Madrid to the Peninsular Center.

As for data recording, the reporting of a forest fire is located in the province where it originated and, therefore, the ignition point is crucial for the distribution of events between neighbouring provinces. For some years now, the term "*megafire*" has been widely used to refer to forest fires that are extremely intense and difficult to control. In Spain, a megafire is considered to be any forest fire that reaches 500 hectares (Ha) of burned forest area in the Iberian Peninsula and 250 Ha in the archipelagos (MITECO, 2023). Reasons for the selection of the area-level auxiliary variables in our forest fire research are presented below. In general, weather conditions and, to some extent, the socio-economic situation of a country and its investment in firefighting resources are crucial factors in identifying patterns in the study of forest fires. In addition, the presence of human settlements leads to biases in firefighting guidelines, as fires that are more dangerous to urban areas are given priority over those in more remote regions. The simultaneity of events is also crucial for the allocation of firefighting equipment. The latter is discussed in more detail in Section 2.3.3 and Section 2.4.3.

The following is a description of the sources of information for the survey microdata. The HBS, LCS and LFSs are periodically statistics harmonized at the European level, carried out by the European National Statistical Offices under the supervision of EUROSTAT, the statistical office of the European Union. According to the Nomenclature of Territorial Units for Statistics (NUTS, 2016), the HBS and LCS are designed to obtain reliable direct estimates at NUTS2 level and the LFS is designed to obtain precise direct estimates at NUTS3 level. The accuracy of the results at a lower level of aggregation than that established in the sample

design is not guaranteed in either survey.

We have worked with microdata from the Spanish HBS (SHBS), the Spanish LFS (SLFS) and the Spanish LCS (SLCS), being in charge of the sampling, storage and validation of this information the Spanish National Statistical Office (INE, *Instituto Nacional de Estadística*). The anonymized data files can be downloaded free of charge from the INE website (https://www.ine.es/; accessed on: November 4, 2024). While the INE publishes the SLFS quarterly, the periodicity of the SHBS and SLCS is annual. Official census data are updated every 10 years, with the latest available in 2021. As for the territorial geocoding in Spain, the NUTS2 geocode corresponds to the autonomous community level, with 19 subdivisions, and the NUTS3 geocode to the province level, with 52 subdivisions. At this regard, the province is the territorial division we use in this research. It should be stressed that if open access data had been available at a lower level of aggregation, we would have chosen it.

The following is a brief description of the surveys. Firstly, the SHBS is published annually to study the nature and destination of household expenditure on goods and services. It includes nearly 24,000 dwellings in its sample, selected by means of a two-stage stratified random sampling carried out independently in each NUTS2 region. Secondly, the SLCS is an annual household survey that focuses on providing comparable and harmonised information on living standards, living conditions and social cohesion. A sample of approximately 13,000 dwellings is taken for the SLCS, selecting 2,000 census sections throughout the national territory. Finally, the SLFS is published quarterly, includes nearly 65,000 dwellings, equivalent to approximately 160,000 people, and collects data on the labour force and its various categories, as well as on the population outside the labour market. Its sampling is two-stage with stratification in the census sections, which are geographical areas with around 500 dwellings or approximately 3,000 people. Census sections are grouped into strata according to the size of the municipality to which they belong. Secondary sampling units are dwellings, and all individuals aged 16 or over in the selected dwelling are interviewed.

1.4 Structure

The thesis is divided into six chapters and four appendices, with mathematical results and additional information and methods, key to achieving a self-contained memory and a reproducible research. Chapter 1 is the introduction, i.e. a comprehensive review of the state of the art, the main objectives and scope, the materials and methods and the structure of the document. Chapter 2 provides a fairly detailed description of the new area-level zero-inflated models and small area predictors, including zero-inflated PO, NB and GA mixed models. Chapter 3 fits a FH3 model to predict DSIs of sex occupational segregation by province and time period. Chapter 4 presents a self-contained account of a unit-level multinomial logit mixed model for the small area prediction of employed, unemployed and inactive proportions and unemployment rates. Chapter 5 departs from the rest in terms of mathematical modelling, but not in its purpose, as it focuses on robust small area prediction using MQ regression. Contributions to this field include the extension of the MQ linear regression to the modelling of time-dependent data, through a Time-Weighted MQ (TWMQ) model. Also, the pioneering proposal of data-driven criteria for the selection of nuisance parameters is included. Chapter

6 summarises the main findings and the lines of future research.

The thesis has four appendices. Appendix A develops the ML-Laplace algorithm to compute the ML estimators of the model parameters and the modal predictors of the random effects for the area-level zero-inflated mixed models in Chapter 2. Appendix B describes the K-means algorithm used in Section 2.4. Appendix C provides an adaptation of the Iterative Re-weighted Least Squares (IRLS) algorithm used to estimate the model parameters of the TWMQ linear models formulated in Chapter 5. Appendix D provides technical specifications and step-by-step proofs of Theorems 1 and 2 in Section 5.4.

Chapter 2

Area-level zero-inflated mixed models

This chapter contains the contributions based on area-level zero-inflated mixed models. It is divided in four self-contained sections. Section 2.1 provides a brief introduction to area-level mixed models. Thereafter, each section corresponds to a paper, to which reference is made in the text itself, and which contains supplementary material available online on the website of the journal in which it has been published. Appendix A goes hand in hand with this chapter. It develops the ML-Laplace algorithm to compute the ML estimators of the model parameters and the modal predictors of the random effects for the area-level zero-inflated mixed models presented here. Appendix B describes the K-means algorithm used in Section 2.4.

2.1 Brief introduction

First of all, we should have a look at the motivation for the problem we want to investigate. In this respect, the topic of modelling zero-inflated outcomes has received less attention than deserves in the SAE literature. Nevertheless, zero-inflated data are almost inevitably complicated by some form of non-observation or inaccurate measurement. Prompted by the need to model variables with an implausible number of zeros, but from a probabilistic framework, we first propose mixtures of mixed models for the small area prediction of indicators dependent on zero-inflated outcomes. For count variables, zero-inflated PO mixed models (Bugallo et al., 2024b) and zero-inflated NB mixed models (Bugallo et al., 2023) are proposed in Section 2.2 and Section 2.3, respectively. For continuous positive variables, Section 2.4 proposes zero-inflated GA mixed models (Bugallo et al., 2024c). The strategy is to tackle the problem in a similar way to that used from standard regression problems, now extended to a much more complex prediction situation when using mixed models. In light of what has been said, all mathematical steps are detailed to justify the soundness of what is presented. In addition, the new methods are illustrated with applications to socio-economic data, modelling the proportion of single-person households by province, sex and age group of the main breadwinner in Section 2.2.5; and to environmental data, modelling the number of provincial forest fires in Section 2.3.4 and the total and the average burned area in Section 2.4.4.

Parametric resampling methods are used to compute bootstrap confidence intervals (CI)

for the model parameters and to estimate the MSEs of the proposed predictors by following Hall and Maiti (2006) and González-Manteiga et al. (2008, 2010). Predictors derived from zero-inflated PO mixed models have also been the subject of several simulation studies in Section 2.2.4. The formulation of the models is always motivated by the case studies, guided by the promise of finding accurate but simple models. Even so, they are easily adaptable to more sophisticated situations, related to more general zero inflation problems.

2.2 Area-level zero-inflated Poisson mixed model

This section describes an area-level zero-inflated PO mixed model aimed at deriving predictors of proportions and counts in small areas. Let U be a finite population of size N which can be divided into mutually disjoint subpopulations U_{ijk} of size N_{ijk}, $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$. Let y_{ijk} be a count variable taking values in $\{0, 1, 2, \dots\}$. Let $D = IJK$ be the total number of y-values. Let z_{ijk}, $\boldsymbol{x}_{1,ijk} = (x_{1,ijk1}, \dots, x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk} = (x_{2,ijk1}, \dots, x_{2,ijkq_2})$ be latent (non observable) variables and $1 \times q_1$, $q_1 \geqslant 1$, and $1 \times q_2$, $q_2 \geqslant 1$, row vectors containing area-level auxiliary variables, respectively. The $\mathrm{col}(\cdot)$ operator stores the data by indexing the observations according to k, then j and finally i. In this way, we define the target and latent vectors and matrices as follows

$$\boldsymbol{y}_{ij} = \underset{1\leqslant k\leqslant K}{\mathrm{col}}(y_{ijk}),\ \boldsymbol{y} = \underset{1\leqslant i\leqslant I}{\mathrm{col}}(\underset{1\leqslant j\leqslant J}{\mathrm{col}}(\boldsymbol{y}_{ij})); \quad \boldsymbol{z}_{ij} = \underset{1\leqslant k\leqslant K}{\mathrm{col}}(z_{ijk}),\ \boldsymbol{z} = \underset{1\leqslant i\leqslant I}{\mathrm{col}}(\underset{1\leqslant j\leqslant J}{\mathrm{col}}(\boldsymbol{z}_{ij})).$$

For the area-level auxiliary variables, we define

$$\boldsymbol{X}_{1,ij} = \underset{1\leqslant k\leqslant K}{\mathrm{col}}(\boldsymbol{x}_{1,ijk}),\ \boldsymbol{X}_1 = \underset{1\leqslant i\leqslant I}{\mathrm{col}}(\underset{1\leqslant j\leqslant J}{\mathrm{col}}(\boldsymbol{X}_{1,ij}));$$
$$\boldsymbol{X}_{2,ij} = \underset{1\leqslant k\leqslant K}{\mathrm{col}}(\boldsymbol{x}_{2,ijk}),\ \boldsymbol{X}_2 = \underset{1\leqslant i\leqslant I}{\mathrm{col}}(\underset{1\leqslant j\leqslant J}{\mathrm{col}}(\boldsymbol{X}_{2,ij})).$$

Let $u_{1,k}$, $u_{2,ijk}$ be independent $N(0,1)$ random effects, $\boldsymbol{u}_{ijk} = (u_{1,k}, u_{2,ijk})'$, and define

$$\boldsymbol{u}_1 = \underset{1\leqslant k\leqslant K}{\mathrm{col}}(u_{1,k}) \sim N_K(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{u}_2 = \underset{1\leqslant i\leqslant I}{\mathrm{col}}(\underset{1\leqslant j\leqslant J}{\mathrm{col}}(\underset{1\leqslant k\leqslant K}{\mathrm{col}}(u_{2,ijk}))) \sim N_D(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'.$$

The bivariate vectors (y_{ijk}, z_{ijk}) follow an area-level zero-inflated PO (aZIP) mixed model if

$$z_{ijk} \overset{ind}{\sim} \mathrm{BE}(p_{ijk}),\ P(y_{ijk} = 0/z_{ijk} = 1) = 1,\ P(y_{ijk} = t/z_{ijk} = 0) = \frac{e^{-\mu_{ijk}}\mu_{ijk}^t}{t!},\ t \in \{0, 1, 2, \dots\}, \tag{2.1}$$

where the probabilities $p_{ijk} \in (0,1)$, $\mu_{ijk} = m_{ijk}\lambda_{ijk}$, $m_{ijk} \in \mathbb{N}$ is known, and $\lambda_{ijk} > 0$. In addition, p_{ijk} and λ_{ijk} depend on the area-level auxiliary variables $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, on the model parameters $\boldsymbol{\beta}_1 = (\beta_{11}, \dots, \beta_{1q_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \dots, \beta_{2q_2})'$, and on the standard deviations $\phi_1 > 0$ and $\phi_2 > 0$ by means of the link functions

$$\mathrm{logit}(p_{ijk}) = \log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}, \qquad \log(\lambda_{ijk}) = \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}.$$

Inverting the above functions, it follows that

$$p_{ijk} = \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}{1 + \exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}, \qquad \lambda_{ijk} = \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}. \tag{2.2}$$

The proposed model is a mixture of two mixed submodels. The BE submodel drives the mixture and incorporates the information derived from the excess of zeros. The PO submodel deals with the modelling of the count variables. It is assumed that the vectors $(y_{ijk}, z_{ijk})'$ are independent conditioned to the random effects. So as to recap, the vectors $(y_{ijk}, z_{ijk})'$ follow an aZIP13 mixed model (Bugallo et al., 2024b), where the terminology "13" is added to specify that the BE model has random effects in only one component, k, and the PO model in the three, i.e. at each crossing ijk.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \phi_1, \phi_2)'$ be the vector of model parameters and define $\xi_{ijk} = I(y_{ijk} = 0)$. The indicator function is denoted by $I(\cdot)$. From the properties of the PO distribution,

$$\begin{aligned} P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}) &= \xi_{ijk}\left[p_{ijk} + (1-p_{ijk})e^{-\mu_{ijk}}\right] + (1-\xi_{ijk})\left[(1-p_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}\right] \\ &= (1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\})^{-1}\Big\{\xi_{ijk}\Big[\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\} \\ &+ \exp\Big\{-m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}\Big\}\Big] + (1-\xi_{ijk})\exp\Big\{y_{ijk}(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}) \\ &- m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\} + y_{ijk}\log m_{ijk} - \log y_{ijk}!\Big\}\Big\}. \end{aligned}$$

The calculation of a factorial number is denoted by !. By the independence assumptions,

$$P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}).$$

Therefore, the likelihood function of the aZIP13 mixed model is

$$\begin{aligned} P(\boldsymbol{y};\boldsymbol{\theta}) &= \int_{\mathbb{R}^{K(1+IJ)}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} \qquad (2.3) \\ &= \prod_{k=1}^{K}\int_{\mathbb{R}^{1+IJ}} \Big(\prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk}\Big) f_{N(0,1)}(u_{1,k})\, du_{1,k}, \end{aligned}$$

and the respective log-likelihood function is

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{k=1}^{K}\log\int_{\mathbb{R}^{1+IJ}} \Big(\prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk}\Big) f_{N(0,1)}(u_{1,k})\, du_{1,k}.$$

Given $\boldsymbol{y}$, the ML parameter estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{\theta};\boldsymbol{y}), \quad \Theta = \mathbb{R}^{q_1+q_2}\times\mathbb{R}_+^2, \quad \mathbb{R}_+ = (0,\infty).$$

To maximize $\ell(\boldsymbol{\theta};\boldsymbol{y})$ in $\boldsymbol{\theta}$, two functions can be used sequentially. The first one would compute the integral on $\mathbb{R}^{1+IJ}$ and the second one would perform the maximization on $\boldsymbol{\theta}$. Since this approach is not efficient, Appendix A describes the ML-Laplace algorithm as an alternative and preferable maximization method. As for the inference procedures of the ML estimators, we rely on both asymptotic (Appendix A) and resampling methods (Section 2.2.2).

2.2.1 Small area prediction of totals and proportions

This section is devoted to the development of new small area predictors based on the aZIP13 mixed model (2.1)-(2.2). Let $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$. The inference focuses on the expected values

$$\mu_{yijk} \stackrel{\Delta}{=} E[y_{ijk}|\boldsymbol{u}_{ijk}] = m_{ijk}(1 - p_{ijk})\lambda_{ijk}, \tag{2.4}$$

where $p_{ijk} \stackrel{\Delta}{=} p_{ijk}(u_{1,k})$ and $\lambda_{ijk} \stackrel{\Delta}{=} \lambda_{ijk}(u_{2,ijk})$ are defined in (2.2).

Firstly, by plugging ML estimators and modal predictors, the population-based quantities given by (2.4) can be predicted using the plug-in (IN) predictor, defined as

$$\widehat{\mu}^{in}_{yijk} = m_{ijk}\big(1 + \exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1\widehat{u}_{1,k}\}\big)^{-1}\exp\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2\widehat{u}_{2,ijk}\}.$$

Of the various predictors that can be mentioned, this is the simplest to understand and the easiest to calculate. Indeed, its ease of interpretation and calculation, as well as its computational performance and execution times, are unsurpassed (Bugallo et al., 2024b). However, there are other potentially competitive alternatives. Let us define

$$\boldsymbol{y}_k = \underset{1\leqslant i\leqslant I}{\text{col}}\big(\underset{1\leqslant j\leqslant J}{\text{col}}(y_{ijk})\big), \quad \boldsymbol{u}_{2,k} = \underset{1\leqslant i\leqslant I}{\text{col}}\big(\underset{1\leqslant j\leqslant J}{\text{col}}(u_{2,ijk})\big), \quad \boldsymbol{v}_k = (u_{1,k}, \boldsymbol{u}'_{2,k})'.$$

The best predictor (BP) of (2.4) is $\widehat{\mu}^{bp}_{yijk}(\boldsymbol{\theta}) = m_{ijk}E[(1 - p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$. The conditional expectation $E_{ijk} = E[(1 - p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$ is

$$E_{ijk} = \frac{\int_{\mathbb{R}^{1+IJ}}\big(1 + \exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\big)^{-1}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}P(\boldsymbol{y}_k|\boldsymbol{v}_k)f(\boldsymbol{v}_k)\,d\boldsymbol{v}_k}{\int_{\mathbb{R}^{1+IJ}}P(\boldsymbol{y}_k|\boldsymbol{v}_k)f(\boldsymbol{v}_k)\,d\boldsymbol{v}_k}.$$

We denote the numerator and denominator of E_{ijk} by $A_{ijk} = A_{ijk}(\boldsymbol{y}_k, \boldsymbol{\theta})$ and $B_k = B_k(\boldsymbol{y}_k, \boldsymbol{\theta})$, respectively. Then, we define $\xi_{rtk} = I_{\{0\}}(y_{rtk})$, $r = 1, \ldots, I$, $t = 1, \ldots, J$, $k = 1, \ldots, K$.

It holds that

$$A_{ijk} = \int_{\mathbb{R}^{1+IJ}}\frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}}{1 + \exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}\prod_{r=1}^{I}\prod_{t=1}^{J}\omega_{rtk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk},$$

$$B_k = \int_{\mathbb{R}^{1+IJ}}\prod_{r=1}^{I}\prod_{t=1}^{J}\omega_{rtk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk},$$

$$\begin{aligned}\omega_{rtk} &= (1 + \exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\})^{-1}\Bigg\{\xi_{rtk}\Bigg[\exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\\ &+ \exp\Big\{-m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk}\}\Big\}\Bigg] + (1 - \xi_{rtk})\exp\Big\{y_{rtk}(\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk})\\ &- m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,rtk}\} + y_{rtk}\log m_{rtk} - \sum_{a=1}^{y_{rtk}}\log a\Big\}\Bigg\}.\end{aligned}$$

The EBP is $\widehat{\mu}^{ebp}_{yijk} = \widehat{\mu}^{bp}_{yijk}(\widehat{\boldsymbol{\theta}})$ and can be calculated by a Monte Carlo method using antithetic variables to reduce the variability (Hobza and Morales, 2016) as follows:

1. Calculate the ML parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \widehat{\phi_1}, \widehat{\phi_2})'$.

2. For $s = 1, \ldots, S$, generate $u_{1,k}^{(s)}$, $u_{2,rtk}^{(s)}$ independent and identically distributed (i.i.d.) according to the $N(0,1)$ distribution and set $u_{1,k}^{(S+s)} = -u_{1,k}^{(s)}$, $u_{2,rtk}^{(S+s)} = -u_{2,rtk}^{(s)}$.

3. Calculate $\widehat{\mu}_{yijk}^{ebp} = m_{ijk}\widehat{A}_{ijk}/\widehat{B}_k$, where

$$
\begin{aligned}
\widehat{A}_{ijk} &= \frac{1}{2S}\sum_{s=1}^{2S}\frac{\exp\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,ijk}^{(s)}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{(s)}\}}\prod_{r=1}^{I}\prod_{t=1}^{J}\widehat{\omega}_{rtk},\ \widehat{B}_k = \frac{1}{2S}\sum_{s=1}^{2S}\prod_{r=1}^{I}\prod_{t=1}^{J}\widehat{\omega}_{rtk}, \qquad (2.5)\\
\widehat{\omega}_{rtk} &= \frac{1}{1+\exp\{\boldsymbol{x}_{1,rtk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{(s)}\}}\Bigg\{\xi_{rtk}\Bigg[\exp\{\boldsymbol{x}_{1,rtk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{(s)}\}\\
&+ \exp\Big\{-m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,rtk}^{(s)}\}\Big\}\Bigg] + (1-\xi_{rtk})\exp\Big\{y_{rtk}(\boldsymbol{x}_{2,rtk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,rtk}^{(s)})\\
&- m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,rtk}^{(s)}\} + y_{rtk}\log m_{rtk} - \sum_{a=1}^{y_{rtk}}\log a\Big\}\Bigg\},\ \ \xi_{rtk} = I_{\{0\}}(y_{rtk}).
\end{aligned}
$$

It has been noticed that the terms in (2.5) contain products with IJ terms. At the expense of the theoretical properties, simpler alternatives are finally proposed in search of a better computational performance. The simplified predictor (SP) is defined as

$$\widehat{\mu}_{yijk}^{sp}(\boldsymbol{\theta}) = m_{ijk}E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}].$$

The conditional expectation $E_{ijk}^{sp} = E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}]$ is

$$E_{ijk}^{sp} = \frac{\int_{\mathbb{R}^2}\left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\right)^{-1}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}}{\int_{\mathbb{R}^2}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}}.$$

We denote the numerator and denominator of E_{ijk}^{sp} by $A_{ijk}^{sp} = A_{ijk}^{sp}(y_{ijk}, \boldsymbol{\theta})$ and $B_{ijk}^{sp} = B_{ijk}^{sp}(y_{ijk}, \boldsymbol{\theta})$, respectively. It holds that

$$A_{ijk}^{sp} = \int_{\mathbb{R}^2}\frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\})}\,\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk},$$

and

$$B_{ijk}^{sp} = \int_{\mathbb{R}^2}\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk}.$$

The empirical simplified predictor (ESP) is $\widehat{\mu}_{yijk}^{esp} = \widehat{\mu}_{yijk}^{sp}(\widehat{\boldsymbol{\theta}})$ and can be approximated by numerical approximation of integrals. However, we apply an antithetical Monte Carlo algorithm:

1. Calculate the ML parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \widehat{\phi_1}, \widehat{\phi_2})'$.

2. For $s = 1, \ldots, S$, generate $\boldsymbol{u}_{ij}^{(s)} = \left(u_{1,k}^{(s)}, u_{2,ijk}^{(s)}\right)'$ i.i.d. $N_2(\boldsymbol{0}, \boldsymbol{I}_2)$ and set $\boldsymbol{u}_{ij}^{(S+s)} = -\boldsymbol{u}_{ij}^{(s)}$.

3. Calculate $\widehat{\mu}_{yijk}^{esp} = m_{ijk}\widehat{A}_{ijk}^{sp}/\widehat{B}_{ijk}^{sp}$, where

$$\widehat{A}_{ijk}^{sp} = \frac{1}{2S}\sum_{s=1}^{2S}\frac{\exp\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,ijk}^{(s)}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{(s)}\})}\widehat{\omega}_{ijk}, \quad \widehat{B}_{ijk}^{sp} = \frac{1}{2S}\sum_{s=1}^{2S}\widehat{\omega}_{ijk}.$$

Because of the numerical precision of the programming language `R`, the calculation of exponential functions to predict μ_{yijk} can report very small negative values (Boubeta et al., 2016a), being ω_{ijk} almost zero. By definition, the ESP avoids these major drawbacks to a greater extent, but the EBP does not, so the latter is omitted from the simulation experiments in Section 2.2.4 and in the application to real data in Section 2.2.5.

2.2.2 Bootstrap inference

This section presents bootstrap-based CIs for the model parameters and estimators of the MSEs of the predictors. Let θ_ℓ be a component of $\boldsymbol{\theta}$ and $\alpha \in (0,1)$. The following procedure calculates a $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ and a parametric bootstrap estimator of $MSE(\widehat{\mu}_{yijk})$, where $\widehat{\mu}_{yijk}$ can be the EBP, ESP or plug-in predictors defined in Section 2.2.1.

1. Fit the model and calculate the ML parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \widehat{\phi}_1, \widehat{\phi}_2)'$.
2. Let $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$.
   Repeat B times ($b = 1,\ldots,B$):
   (a) Generate $u_{1,k}^{*(b)} \sim N(0,1)$, $u_{2,ijk}^{*(b)} \sim N(0,1)$ and calculate

$$\begin{aligned} p_{ijk}^{*(b)} &= \exp\left\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{*(b)}\right\}\left(1+\exp\left\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u_{1,k}^{*(b)}\right\}\right)^{-1}, \\ \lambda_{ijk}^{*(b)} &= \exp\left\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u_{2,ijk}^{*(b)}\right\}. \end{aligned}$$

   (b) Generate $z_{ijk}^{*(b)} \sim \text{BE}(p_{ijk}^{*(b)})$. If $z_{ijk}^{*(b)} = 1$, $y_{ijk}^{*(b)} = 0$. Otherwise, $y_{ijk}^{*(b)} \sim \text{PO}(m_{ijk}\lambda_{ijk}^{*(b)})$.
   (c) Calculate $\mu_{yijk}^{*(b)} = m_{ijk}(1-p_{ijk}^{*(b)})\lambda_{ijk}^{*(b)}$.
   (d) On the basis of the bootstrap sample $(y_{ijk}^{*(b)}, m_{ijk}, \boldsymbol{x}_{ijk})$, calculate the ML parameter estimator $\widehat{\theta}_\ell^{*(b)}$, the bootstrap version of the vector of model parameters, $\widehat{\boldsymbol{\theta}}^{*(b)}$, and the predictor $\widehat{\mu}_{yijk}^{*(b)}$.
3. Sort the values $\widehat{\theta}_\ell^{*(b)}$, $b = 1,\ldots,B$, from smallest to largest. They are $\widehat{\theta}_{\ell(1)}^{*} \leqslant \ldots \leqslant \widehat{\theta}_{\ell(B)}^{*}$. A $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ is $\left(\widehat{\theta}_{\ell(\lfloor(\alpha/2)B\rfloor)}^{*}, \widehat{\theta}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}^{*}\right)$, where $\lfloor\cdot\rfloor$ is the closest integer operator.
4. To estimate error measures, define

$$mse^*(\widehat{\mu}_{yijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mu}_{yijk}^{*(b)} - \mu_{yijk}^{*(b)}\right)^2.$$

2.2.3 Description of the 2016 SHBS data

The application to real data aims to estimate the proportion and total count of single-person households by province, sex (*sex1*: men, *sex2*: women) and age group (*age1*: less than 45 years; *age2*: between 46 and 55 years; *age3*: between 56 and 64 years; *age4*: 65 years or older) of the main breadwinner, which is particularly noteworthy (Cho and Shim, 2019). Data are from the 2016 Spanish Household Budget Survey (SHBS), so U is the finite population of Spanish households in 2016. As a result, the $D = 416$ domains are defined at NUTS 3 level by Spanish province ($I = 52$) crossed by sex ($J = 2$) and age group ($K = 4$). The quartiles of the domain sample sizes reveal that this is a SAE problem such that $q_0 = 1$, $q_{0.25} = 17$, $q_{0.5} = 34$, $q_{0.75} = 72$ and $q_1 = 367$.

At unit-level, the variable of interest is dichotomic, i.e. $y_{ijkl} = 1$ if the household $u_{ijkl} \in U_{ijk}$ is single-person and $y_{ijkl} = 0$, otherwise. Let $s = \bigcup_{i=1}^{I} \bigcup_{j=1}^{J} \bigcup_{k=1}^{K} s_{ijk}$ be the 2016 SHBS sample extracted from U. Let n and n_{ijk} be the sample sizes of s and s_{ijk}, respectively. For ease of exposition, we write $l = 1, \ldots, n_{ijk}$ for the households in s_{ijk} and $l = n_{ijk}+1, \ldots, N_{ijk}$ for the households in $U_{ijk} - s_{ijk}$. Let w_{ijkl} be the household sampling weight of $u_{ijkl} \in U_{ijk}$. The domain parameters of interest are

$$Y_{ijk} = \sum_{l=1}^{N_{ijk}} y_{ijkl}, \quad \overline{Y}_{ijk} = \frac{Y_{ijk}}{N_{ijk}}. \tag{2.6}$$

The Hájek estimator of Y_{ijk}, N_{ijk} and $\overline{Y}_{ijk}$ are, respectively,

$$\widehat{Y}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}, \quad \widehat{N}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}, \quad \widehat{\overline{Y}}_{ijk}^{dir} = \frac{\widehat{Y}_{ijk}^{dir}}{\widehat{N}_{ijk}^{dir}}.$$

Let $\widehat{\mu}_{yijk}$ be a model-based predictor of y_{ijk}. Population sizes and auxiliary information are taken from the four 2016 SLFSs, as the sample size of each quarterly SLFS is more than three times the size of an annual SHBS. The effect of the variances of the covariate means and population sizes on the properties of the prediction procedure is considered negligible. In addition, the elevation factors are the inverse of the inclusion probabilities, which are deterministic, after a calibration process whose randomness is minimal. Therefore, the population sizes estimated as sums of elevation factors have negligible variability.

As for y_{ijk} and m_{ijk}, two options can be considered:

Option 1. Take $y_{ijk} = \lfloor \widehat{Y}_{ijk}^{dir} \rceil$ and $m_{ijk} = \lfloor N_{ijk} \rceil$. The predictors of $\overline{Y}_{ijk}$ and Y_{ijk} are

$$\widehat{\overline{Y}}_{ijk} = \frac{\widehat{\mu}_{yijk}}{m_{ijk}}, \quad \widehat{Y}_{ijk} = \widehat{\mu}_{yijk}.$$

Option 1 reconciles the area-level model-based approach and the design-based approach to inference in finite populations. This is an important argument in favour of Option 1. However, the fitting algorithm or the calculation of predictors may become unstable when the values of the dependent variable are large, requiring more refined programming.

Option 2. Take $y_{ijk} = \sum_{l=1}^{n_{ijk}} y_{ijkl}$ and $m_{ijk} = \sum_{l=1}^{n_{ijk}} y_{ijkl}$. The predictors of $\overline{Y}_{ijk}$ and Y_{ijk} are

$$\widehat{\overline{Y}}_{ijk} = \frac{\widehat{\mu}_{yijk}}{m_{ijk}}, \quad \widehat{Y}_{ijk} = \widehat{N}^{dir}_{ijk}\widehat{\overline{Y}}_{ijk}.$$

Boubeta et al. (2016a) applies Option 2 for area-level PO mixed models because it is computationally more robust, but it does not include the sampling weights. Since the omission of sampling weights is an important issue with Option 2 –as it can lead to biased predictors– our choice of Option 1 is properly justified, even if it makes programming more difficult.

As for the domain-level auxiliary variables, they are obtained by calculating the Hájek estimates of the proportion of people in the following factor categories: *Citizenship*: Spanish (*cit1*) and foreign (*cit2*); *Education*: primary education or less (*edu1*), basic secondary education (*edu2*), advanced secondary education (*edu3*) and higher education, such as university (*edu4*); *Labour situation*: employed (*lab1*), unemployed (*lab2*) and inactive (*lab3*); *Civil status*: unmarried (*civ1*), married (*civ2*), widowed (*civ3*) and separated or divorced (*civ4*); *Dwelling mobility*: more than a year in the same dwelling (*dwe1*) and the opposite (*dwe2*). The aforementioned auxiliary variables are proportions, bounded in [0,1], i.e. they are continuous variables, not binary indicators. Since the sum of the proportions in the categories of each factor is one, and based on their socio-economic meaning, we omit one category from each factor. Namely, we have deleted *cit2, edu2, lab3, civ1, dwe2*.

So far we have discussed the need of auxiliary information, but we have not addressed the problem of excess zeros, or even shown that they exist. Null counts are caused by the difficulty of detecting single-person households due to the low number of respondents in some domains. Table 2.1 displays the distribution of the 28 zeros by sex and age group in the sample. On closer inspection, they are mainly concentrated in certain crosses (*age1*:*sex2*; *age2*:*sex2*) and the number is too high for what would be expected from a PO distribution.

	age group				
sex	*age1*	*age2*	*age3*	*age4*	Total
sex1	3	2	2	3	10
sex2	8	8	2	0	18
Total	11	10	4	3	28

Table 2.1: Unobserved domain-level single-person households in the 2016 SHBS by sex and age group. In other words, total number of zeros per domain by sex and age group.

To test the dependence between the number of zeros/non-zeros and provinces, sex and age groups, we used the Pearson's Chi-Squared test in $2 \times I$, $2 \times J$ and $2 \times K$ contingency tables, where p-values are calculated by Monte Carlo. As a result, p-values close to 0.06 are obtained for province and age group as inputs, rising to 0.18 for sex. Based on Table 2.1 and the results of the above tests, we have decided to consider only age-group randomness to model zero-inflated probabilities. Furthermore, applying the same tests to assess the dependence between the count of single-person households (less/greater than 1, 2 or 3) and provinces, sex and age groups, only the randomness of the age group is significant. Guided by the promise of finding a good, simple model, the aZIP13 mixed model was proposed in Section 2.2.

2.2.4 Simulations based on the 2016 SHBS data

Based on the 2016 SHBS data described in Section 2.2.3, two simulation experiments were run. According to Option 1, the dependent variable y_{ijk} is the direct estimator of the total count of single-person households in province i, with the main breadwinner of sex j and age group k. It has been assumed that y_{ijk} follows the aZIP13 mixed model selected in the statistical analysis of Section 2.2.5. As $q_1 = 1$, the BE submodel contains one auxiliary variable: $x_{1,1}$ = intercept, and the regression parameter is $\beta_{11} = -2.696$. The PO submodel contains $q_2 = 4$ auxiliary variables: $x_{2,1}$ = intercept, $x_{2,2} = edu3$, $x_{2,3} = civ2$ and $x_{2,4} = civ3$, with regression parameters $\beta_{21} = -1.857$, $\beta_{22} = 2.138$, $\beta_{23} = -0.649$ and $\beta_{24} = 3.881$. The standard deviations are $\phi_1 = 0.398$ and $\phi_2 = 0.5171$. Setting the random effects $u_{1,k}$, $k = 1, \ldots, K$, to their theoretical expected value zero, the basic zero-inflated probability is

$$p_0 = p_0(\beta_{11}) = \exp\{\beta_{11}\}\big(1 + \exp\{\beta_{11}\}\big)^{-1} = 0.063.$$

Simulation 1

Simulation 1 aims to evaluate the fitting algorithm, i.e. the ML-Laplace algorithm described in Appendix A, and investigate the performance of the new small area predictors defined in Section 2.2.1. It also examines what happens when the excess zeros are ignored in the prediction. Apart from those derived from the aZIP13 mixed model, i.e., from the SP, ESP and plug-in (IN) predictors, the plug-in predictor with fixed zero-inflated probability (IN1) and the one based on the erroneous non-inflated PO mixed model (IN0) are considered. The model parameters of the IN1 predictor are $\beta_{11}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \phi_2$ and those of the IN0 predictor are $\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \phi_2$. As for the remaining models, the EBLUP of the total count of single-person households based on the FH model (Fay and Herriot, 1979) is included. In addition, a zero-inflated NB mixed model (aZINB13) is fitted to identify the advantages of the proposed procedure for excess zeros and the corresponding IN predictor is derived.

Simulation 1 has the following steps:

1. Let $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Repeat $R = 10^3$ times ($r = 1, \ldots, R$):

   1.1. Generate $u_{1,k}^{(r)}$, $u_{2,ijk}^{(r)}$ i.i.d. $N(0,1)$.

   1.2. Calculate $p_k^{(r)} = \exp\{\beta_{11} + \phi_1 u_{1,k}^{(r)}\}\left(1 + \exp\{\beta_{11} + \phi_1 u_{1,k}^{(r)}\}\right)^{-1}$,

$$\lambda_{ijk}^{(r)} = \exp\left\{\sum_{\ell=1}^{4} x_{2,ijk\ell}\beta_{2\ell} + \phi_2 u_{2,ijk}^{(r)}\right\}, \quad \mu_{yijk}^{(r)} = m_{ijk}\left(1 - p_k^{(r)}\right)\lambda_{ijk}^{(r)}.$$

   1.3. Generate $z_{ijk}^{(r)} \sim \text{BE}\left(\widehat{p}_k^{(r)}\right)$, $y_{ijk}^{(r)} = 0$ if $z_{ijk}^{(r)} = 1$; $y_{ijk}^{(r)} \sim \text{PO}\left(m_{ijk}\lambda_{ijk}^{(r)}\right)$ if $z_{ijk}^{(r)} = 0$.

   1.4 Calculate the ML estimators $\widehat{\tau}^{(r)} \in \{\widehat{\beta}_{11}^{(r)}, \widehat{\beta}_{21}^{(r)}, \widehat{\beta}_{22}^{(r)}, \widehat{\beta}_{23}^{(r)}, \widehat{\beta}_{24}^{(r)}, \widehat{\phi}_1^{(r)}, \widehat{\phi}_2^{(r)}\}$ and the model-based predictors $\widehat{\mu}_{yijk}^{(r)} \in \{\widehat{\mu}_{yijk}^{sp(r)}, \widehat{\mu}_{yijk}^{esp(r)}, \widehat{\mu}_{yijk}^{in(r)}, \widehat{\mu}_{yijk}^{in1(r)}, \widehat{\mu}_{yijk}^{in0(r)}, \widehat{\mu}_{yijk}^{FH(r)}\}$.

2. For each estimator τ and model-based predictor $\widehat{\mu}_{yijk}$, calculate

$$BIAS(\widehat{\tau}) = \tfrac{1}{R}\sum_{r=1}^{R}(\widehat{\tau}^{(r)} - \tau), \quad RMSE(\widehat{\tau}) = \left(\frac{1}{R}\sum_{r=1}^{R}(\widehat{\tau}^{(r)} - \tau)^2\right)^{1/2},$$

$$BIAS_{ijk} = \tfrac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_{yijk}^{(r)} - \mu_{yijk}^{(r)}), \quad RMSE_{ijk} = \left(\frac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_{yijk}^{(r)} - \mu_{yijk}^{(r)})^2\right)^{1/2},$$

and

$$ABIAS = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}|BIAS_{ijk}|, \; RMSE = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}RMSE_{ijk}.$$

3. Calculate the corresponding relative performance measures in %. That is, calculate the relative bias ($RBIAS_{ijk}$), the relative root-MSE ($RRMSE_{ijk}$), the average absolute relative bias (ARBIAS) and the average relative root-MSE (RRMSE):

$$RBIAS(\widehat{\tau}) = 100\frac{BIAS(\widehat{\tau})}{|\tau|}, \quad RRMSE(\widehat{\tau}) = 100\frac{RMSE(\widehat{\tau})}{|\tau|},$$

$$RBIAS_{ijk} = 100\frac{BIAS_{ijk}}{|\overline{\mu}_{yijk}|}, \quad RRMSE_{ijk} = 100\frac{RMSE_{ijk}}{|\overline{\mu}_{yijk}|}, \quad \overline{\mu}_{yijk} = \frac{1}{R}\sum_{r=1}^{R}\mu_{yijk}^{(r)},$$

$$ARBIAS = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}|RBIAS_{ijk}|, \quad RRMSE = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}RRMSE_{ijk}.$$

For the SHB2016 scenario, Table 2.2 (top) shows the results of Simulation 1 for the model parameters. To investigate the effect of the basic zero-inflated probabilities, we also consider the cases $p_0 = 0.200$ (Table 2.2, center) and $p_0 = 0.500$ (Table 2.2, bottom).

For both BE and PO submodels, the relative biases are small but the RRMSEs are not, being the variance the main component of the MSE. This may be due to the fact that the ratio between the number of domains and the number of estimated model parameters, $416/7 \approx 60$, is not large enough to activate the asymptotic properties of the ML estimators. For the basic zero-inflated probabilities $p_0 = 0.2$ and $p_0 = 0.5$, the ML estimators of β_{11} and ϕ_1 have slightly lower values of RBIAS and RRMSE than the corresponding ones under the SHBS2016 scenario, with $p_0 = 0.063$. This suggests that the BE submodel estimators perform slightly better as the basic zero-inflated probability increases. However, the changes are small. There are no notable differences in the remaining coefficients.

Table 2.3 includes the relative performance measures of Simulation 1 for the predictors SP, ESP, IN (of the aZIP13 and aZINB13 mixed models), IN1, IN0 and FH. To better understand the necessity of running this experiment and interpret its results, we emphasize that the predictors SP and ESP are not calculated, but rather are approximated, since the integrals that appear in their formulas cannot be calculated analytically. The approximations are obtained by the antithetical Monte Carlo method, with $S = 2000$, as described in Section

$p = 0.063$	BE submodel		PO submodel				
	β_{11}	ϕ_1	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_2
Estimate	-2.696	0.398	-1.857	2.138	-0.649	3.881	0.517
RBIAS	-0.183	-42.196	0.881	-2.303	-1.289	-0.486	-0.510
RRMSE	11.329	78.693	190.866	121.998	326.656	96.880	3.708

$p = 0.2$	BE submodel		PO submodel				
	β_{11}	ϕ_1	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_2
Estimate	-2.696	0.398	-1.857	2.138	-0.649	3.881	0.517
RBIAS	0.791	-34.480	1.325	-3.198	-2.113	-0.801	-0.659
RRMSE	16.398	60.556	190.449	122.655	345.917	96.770	4.007

$p = 0.5$	BE submodel		PO submodel				
	β_{11}	ϕ_1	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_2
Estimate	-2.696	0.398	-1.857	2.138	-0.649	3.881	0.517
RBIAS	NaN	-29.43	1.576	-3.116	-3.407	-0.731 3	-1.110
RRMSE	NaN	55.028	191.085	123.530	315.871	96.855	4.994

Table 2.2: Relative performance measures of model parameter estimators for $p = 0.063$ (top), $p = 0.2$ (center) and $p = 0.5$ (bottom) for the BE (left) and PO (right) submodels.

2.2.1. Since we approximate integrals in $\mathbb{R}^2$, the theoretical properties are largely missing but increasing S even more in a simulation experiment with $R = 1000$ iterations entails unaffordable computation times in Simulation 1 and even more so in Simulation 2. Therefore, the results are subject to the approximation method and the number of iterations.

		aZIP13					FH	aZINB13
p_0	Measure	SP	ESP	IN	IN1	IN0	EBLUP	IN
0.063	ARBIAS	0.358	0.361	0.790	9.179	3.068	0.780	3.968
	RRMSE	14.429	14.476	14.662	60.759	60.789	30.380	28.143
0.200	ARBIAS	0.727	0.739	2.444	9.286	13.460	1.405	4.492
	RRMSE	26.270	26.155	25.894	60.714	63.759	57.578	34.117
0.500	ARBIAS	2.965	2.130	5.955	10.716	77.754	2.925	5.566
	RRMSE	43.697	43.075	40.926	62.293	104.149	111.921	44.572

Table 2.3: Relative performance measures (in %) for the predictors with $S = 2000$.

The discussion of Table 2.3 starts with the analysis of the predictors proposed for the aZIP13 mixed model. Under all scenarios, the SP has the lowest bias, increasing slightly in its theoretical versions. When substituting true model parameters by ML parameter estimators, the performance of the ESP is almost as good as that of the SP. In nominal terms, the variance has a notable contribution to the RRMSE for all predictors. Since the ESP and the

IN predictor have similar RRMSEs, it has been decided to use the latter in Simulation 2 and in the case study, as it is computationally preferable. As expected, the predictors IN1 and IN0 are biased and have higher RRMSE than the IN predictor.

Under the scenarios with basic zero-inflated probabilities $p_0 = 0.2$ and $p_0 = 0.5$, the predictors IN1 and IN0 perform poorly, with relative biases equal to 9.286 and 13.460 ($p = 0.2$). By increasing p_0 from 0.2 to 0.5, the RRMSE of the IN1 predictor stabilizes, even though the IN predictor is better. The latter indicates that the age-group randomness is less relevant for such high zero-inflated probabilities. In the latter case, the IN0 predictor performs extremely poorly. We conclude that the IN predictor obtained from the aZIP13 mixed model performs much better than the predictor based on the model with constant zero inflation structure. The same applies to the IN0 predictor. Therefore, we do not recommend to use predictors IN0 and IN1 when there is an excess of zeros.

As for the EBLUP-FH, its bias is small for all values of p_0, with results close to the SP and the ESP. However, this is not a zero-inflated model, which has a negative impact on the error through a significant increase in the variance as p_0 increases. Regarding the IN predictor of the aZINB13 mixed model, the bias is greater than that of the IN predictor of the aZIP13 mixed model. Nevertheless, this is compensated to some extent by a lower variance, achieving similar but worse results. To sum up, the main advantages of the aZIP13 mixed model over existing models, and in particular of the IN predictor, are computational performance and reduction in ARBIAS and RRMSE.

Simulation 2

Simulation 2 studies the behaviour of the parametric bootstrap estimator of the MSE of a predictor $\widehat{\mu}_{yijk}$ of μ_{yijk}. The latter is done by comparing $mse^*(\widehat{\mu}_{yijk})$ with the empirical MSE of $\widehat{\mu}_{yijk}$, obtained from Simulation 1. For illustrative purposes and speed of computation, we select $\widehat{\mu}_{yijk} = \widehat{\mu}^{in}_{yijk}$. The aim is to give some advice on which B value to choose.

Simulation 2 has the following steps:

1. Let $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$.
   Take $MSE_{ijk} = RMSE^2_{ijk}$ from Simulation 1.
2. Repeat $R = 500$ times ($r = 1, \ldots, R$):
   2.1. Generate a sample $(y^{(r)}_{ijk}, x_{1,ijk}, \boldsymbol{x}_{2,ijk})$ and calculate the ML estimation $\widehat{\boldsymbol{\theta}}^{(r)}$.
   2.2. Repeat B times ($b = 1, \ldots, B$):
      2.2.1. Generate $u^{*(rb)}_{1,k}$, $u^{*(rb)}_{2,ijk}$ i.i.d. $N(0,1)$ and calculate

$$
\begin{aligned}
p^{*(rb)}_k &= \exp\{\widehat{\beta}_{11} + \widehat{\phi}_1 u^{*(rb)}_{1,k}\} \left(1 + \exp\{\widehat{\beta}_{11} + \widehat{\phi}_1 u^{*(rb)}_{1,k}\}\right)^{-1}, \\
\lambda^{*(rb)}_{ijk} &= \exp\left\{ \sum_{\ell=1}^{4} x_{2,ijk\ell}\widehat{\beta}_{2\ell} + \widehat{\phi}_2 u^{*(rb)}_{2,ijk} \right\}, \quad \mu^{*(rb)}_{yijk} = m_{ijk}\left(1 - p^{*(rb)}_k\right)\lambda^{*(rb)}_{ijk}.
\end{aligned}
$$

2.2.2 Generate $z_{ijk}^{*(rb)} \sim \text{BE}\left(\widehat{p}_k^{*(rb)}\right)$, $y_{ijk}^{*(rb)} = 0$ if $z_{ijk}^{*(rb)} = 1$ and $y_{ijk}^{*(rb)} \sim \text{PO}\left(m_{ijk}\lambda_{ijk}^{*(rb)}\right)$ if $z_{ijk}^{*(rb)} = 0$. Then calculate the predictor $\widehat{\mu}_{yijk}^{*(rb)}$.

2.3 Define

$$mse_{ijk}^{*(r)} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mu}_{yijk}^{*(rb)} - \mu_{yijk}^{*(rb)}\right)^2.$$

3. Calculate

$$B_{ijk} = \frac{1}{R}\sum_{r=1}^{R}\left(mse_{ijk}^{*(r)} - MSE_{ijk}\right), \quad RE_{ijk} = \left(\frac{1}{R}\sum_{r=1}^{R}\left(mse_{ijk}^{(*r)} - MSE_{ijk}\right)^2\right)^{1/2},$$

$$AB = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}|B_{ijk}|, \quad RE = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}RE_{ijk}.$$

4. Calculate the corresponding relative performance measures in %, i.e.

$$RB_{ijk} = 100\frac{B_{ijk}}{MSE_{ijk}}, \quad RRE_{ijk} = 100\frac{RE_{ijk}}{MSE_{ijk}},$$

$$ARB = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}RB_{ijk}, \quad RRE = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}RRE_{ijk}.$$

The non-relative measures depend on the large values of the output. Consequently, Figure 2.1 prints five boxplots of RB_{ijk} and RRE_{ijk}, for $B = 100, 200, 400, 500, 600$.

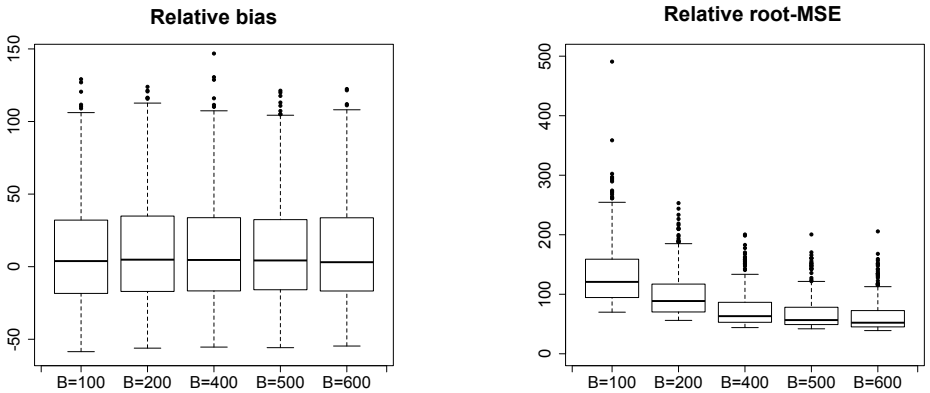


Figure 2.1: Study of the parametric bootstrap estimator of the MSE of $\widehat{\mu}_{yijk}^{in}$. Boxplots of RB_{ijk}'s (left) and RRE_{ijk}'s (right) for $B = 100, 200, 400, 500, 600$.

As can be observed in Figure 2.1 (left), the relative biases do not decrease as the size of B increases, showing an origin-centric behaviour. Moreover, there are few atypical values that correspond to the most conflictive domains, i.e., those with smaller sample sizes or with zero observed single-person households in the 2016 SHBS data. This severely distorts the

symmetry of the ordinate axis. Figure 2.1 (right) illustrates that the RRMSEs decrease as B increases. Table 2.4 confirms this behaviour, with an ARB stabilized around 10.500 and a RRE decreasing as B increases, but suggesting some stabilization around $B = 600$ iterations. It is concluded that the results for the MSE estimator of the IN predictor are reasonable in most domains. However, the low sample size of some of them and the non-observation of single-person households add additional biases.

B	100	200	400	500	600
ARB	10.244	10.566	10.657	10.723	10.594
RRE	134.643	99.255	73.821	68.054	60.089

Table 2.4: Study of the parametric bootstrap estimator of the MSE of $\widehat{\mu}^{in}_{yijk}$. Average relative performance measures for $B = 100, 200, 400, 500, 600$.

2.2.5 Application to the 2016 SHBS data

This study is a pioneering approach for estimating the proportion of single-person households in small areas. Even though, the latter is essential for a more accurate implementation of social policies, as well as for clarifying certain economic aspects related to the housing sector and the private consumption of basic resources (Cohen, 2021). Incidentally, the case study is motivated by the need to map the distribution of single-person households, as it is well-known that household composition reveals vital aspects of the socio-economic situation and major changes in developed countries. Living alone has become a sign of individual autonomy and freedom, even if it is sometimes still stereotyped (Greitemeyer, 2009). Meanwhile, loneliness and its impact on physical and mental health are an increasingly widespread problem (Snell, 2017), accentuating the symptoms of cognitive diseases (Lee and Lee, 2021; Park et al., 2016). Particularly in single-person households inhabited by the elderly. Among the main indicators of loneliness, we can mention the proportion and total count of single-person households by domains defined by territorial and socio-demographic features.

As for the issue at hand, Section 2.2.1 derives predictors of the domain parameters defined in (2.6), based on the aZIP13 mixed model described in Section 2.2. The scenario of the application to real data has been detailed in Section 2.2.3. Population sizes and the area-level auxiliary variables have been obtained from the 2016 SLFS microdata. According to Option 1, the dependent variable, y_{ijk}, is the direct estimator of the total count of single-person households in province i, with main breadwinner of sex j and age group k.

Table 2.5 shows the ML parameter estimators of the model parameters $\boldsymbol{\beta}_1$, ϕ_1 (BE submodel) and $\boldsymbol{\beta}_2$, ϕ_2 (PO submodel), the p-values to test $H_0 : \beta_{t\ell} = 0$, $t = 1, 2$, $\ell = 1, ..., q_t$, and $H_0 : \phi_t = 0$, $t = 1, 2$, and the normal-asymptotic and bootstrap CIs at the 95% confidence level. For convenience, their lower (LB) and upper (UB) bounds are provided. Normal-asymptotic CIs are discussed in Appendix A and bootstrap CIs in Section 2.2.2. The final model incorporates only those variables that are significant at 5%.

The flexibility achieved by making the random effects of the count model domain-dependent allows us to reduce the importance of the set of domain-level variables and incorporate only

		BE submodel		PO submodel				
		β_{11}	ϕ_1	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_2
	Estimate	-2.696	0.398	-1.857	2.138	-0.649	3.881	0.517
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Asymp	LB 95%	-3.270	0.091	-2.319	1.007	-1.057	3.207	0.482
	UB 95%	-2.121	1.752	-1.395	3.269	-0.242	4.554	0.555
Boot	LB 95%	-3.317	0.0002	-2.312	1.051	-1.016	3.215	0.480
	UB 95%	-2.162	0.859	-1.432	3.270	-0.222	4.577	0.554

Table 2.5: Model parameters of the final aZIP13 mixed model for the BE (left) and PO (right) submodels.

those that actually add relevant knowledge. The BE submodel contains one auxiliary variable, $x_{1,1}$ = intercept, and the PO submodel four: $x_{2,1}$ = intercept, $x_{2,2} = edu3$, $x_{2,3} = civ2$, $x_{2,4} = civ3$. The basic zero-inflated probability is $p_0(\widehat{\beta}_{11}) = 0.063$. For further confidence in the model linked to Table 2.5 as the true generator model, a residual analysis is performed. Besides, we are interested in the conciliation of the model-based approach and the design-based approach to SAE. Let us define the raw residuals (RR) as

$$\widehat{e}_{ijk} = y_{ijk} - \widehat{\mu}^{in}_{yijk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

Under Option 1, $y_{ijk} = \lfloor \widehat{Y}^{dir}_{ijk} \rfloor$ and $\widehat{e}_{ijk} = \lfloor \widehat{Y}^{dir}_{ijk} \rfloor - \widehat{\mu}^{in}_{yijk}$. The standardized residuals (SR) are defined by dividing the RRs by its standard deviation, i.e.

$$\widehat{e}_{ijk}\nu^{-1}, \quad \text{where} \quad \nu = \left(\frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\widehat{e}_{ijk} - \widehat{\overline{e}}_{...} \right)^2 \right)^{\frac{1}{2}}, \quad \widehat{\overline{e}}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \widehat{e}_{ijk}.$$

Figure 2.2 plots the SRs of the aZIP13 mixed model versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log-scale (right). In dotted red, the line $y = 0$ is added. As general conclusions, it can be seen that SRs have a pattern of symmetry around zero and are mainly found in $[-3, 3]$. The central plot has a low percentage of domains with large predicted probabilities, which exceed the threshold of 0.7, and correspond to domains with predominantly single-person households inhabited by elderly women. Regarding the right plot, plotting SRs against log-predicted probabilities allows us to detect a conical pattern in the scatterplot. That is, as the log-predicted probabilities increases, so does the variability of the SRs. This phenomenon is in agreement with the theoretical dispersion of the aZIP13 mixed model.

Once the model has been fitted and validated, it is time to provides Hájek estimates and IN predictions of the proportion of single-person households by province, sex and age group. Figure 2.3 shows line charts of these values sorted by domain index (left) and sample size (center), as well as a comparison of both (right). Among the most noteworthy findings, model-based predictors correct the excessively large Hájek estimates. Even more, it is inferred that the IN predictor smoothes the results of the Hájek estimator, although it still presents

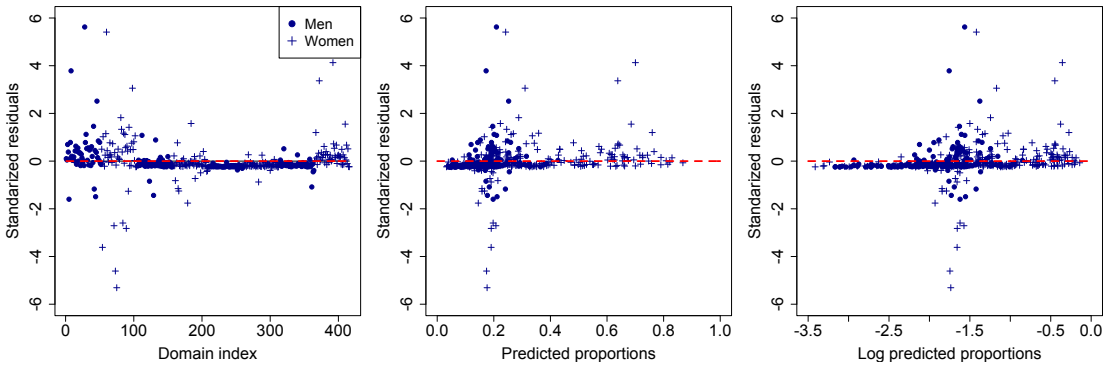


Figure 2.2: SRs versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log-scale (right).

problems when dealing with extreme proportions. On the other hand, provided single-person households are not observed, the Hájek estimator has no margin of error, although the model never comes to such a low proportion. The same is true for values close to one. This can be seen in Figure 2.3 (left). In addition, household composition does not affect all domains equally: as the age group increases, the proportion of single-person households also increases. In this context, we would like to draw the attention to the sudden and noticeable increase in the number of single-person households inhabited by elderly women.

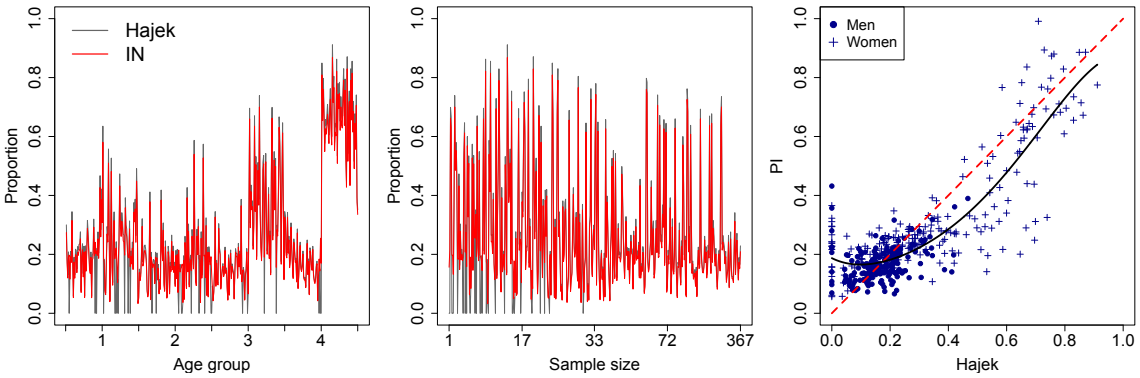


Figure 2.3: IN proportions of single-person households sorted by domain (left) and sample size (center), and Hájek estimates versus IN proportions (right). Sorting by domain corresponds to sorting by age group, showing first the results for men and then for women.

According to Figure 2.3 (center), the IN predictor gets closer to the Hájek estimator as the sample size increases, which is one of the most convincing aspects of the data analysis. Eventually, Figure 2.3 (right) plots the Hájek estimates versus the IN proportions. It can be seen that the dots are evenly distributed around $y = x$. To support this statement, a local polynomial regression of degree 3, with an appropriate bandwidth, is plotted to smoothly represent the relationship between ordinates and abscissas. Consequently, we underline a crucial advantage of our approach: the theoretical properties of the Hájek estimator, such

as asymptotic design-based unbiasedness, are, to some extent, inherited by the IN predictor based on the aZIP13 mixed model.

Table 2.7 (a) reports IN proportions of single-person households by sex and age group. The current trend projects an increase in the proportion of single-person households, with the number of households inhabited by elderly women skyrocketing. The latter is associated with the ageing process, which progressively involves the emancipation of children and widowhood. In addition, the elderly is linked to another factor that alters household composition: mortality. So sex and *age4* are crucial here. The increase in quality of life implies not only an increase in life expectancy but also in the autonomy of the elderly, which results in an increase in the number of single-person households inhabited by retired people. Most men live with their partners until their death. In contrast, women have a longer life expectancy (implying a greater accumulation at the top of the demographic pyramid) and the average age of their partners is higher, so they will live alone to a greater extent. The statements are based on information for 2021 published on the official website of the Ministry of Health of the Government of Spain (https://www.sanidad.gob.es/estadEstudios/estadisticas/inforRecopilaciones/ESPERANZAS_DE_VIDA_2021.pdf; accessed on: November 4, 2024).

	sex		
age group	*sex1*	*sex2*	Total
age1	0.199	0.239	0.219
age2	0.160	0.184	0.174
age3	0.147	0.369	0.261
age4	0.171	0.648	0.439
Total	0.1830	0.337	0.262

(a) IN proportions aggregated by province.

	sex		
age group	*sex1*	*sex2*	Total
age1	20.836	19.825	20.335
age2	20.379	22.094	21.245
age3	20.941	12.680	16.692
age4	20.158	18.115	19.007
Total	20.630	19.313	19.95

(b) IN RRMSEs (%) aggregated by province.

Table 2.7: Results for the predicted proportion of single-person households.

As for the error measures, we calculate the parametric bootstrap estimator of the MSE following Section 2.2.2 with $B = 1000$ resamples. To avoid scale dependencies, and as usual, the script should be focused on RRMSEs. Table 2.7 (b) contains the bootstrap estimates of the RRMSE (in %) for the IN predictor by sex and age group. As a general conclusion, all values are around 20%, with a slightly lower average for women and especially for *age3*.

For illustrative purposes, Figure 2.4 maps the provincial distribution of single-person households for men (left) and women (right) for the first age group of the main breadwinner. The results are expressed as percentages.

It may be suggested that the highest proportions of single-person households are found in central and north-western Spain, with lower proportions in the south and the Canary Islands. Not surprisingly, the distribution between neighboring provinces, or between provinces with similar demographic and socio-economic conditions, is generally homogeneous. This fact justifies how model-based predictors lead to smoother results (and closer to the reality) than direct estimators. In addition, an interesting spatial pattern emerges: an inverse relationship

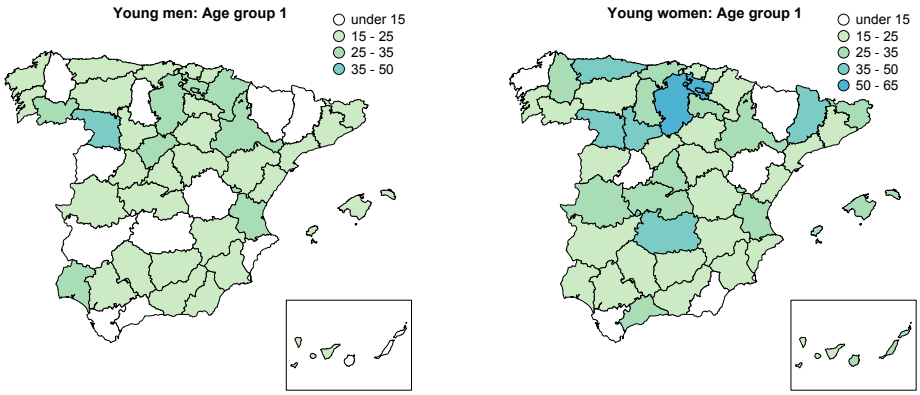


Figure 2.4: Percentages of single-person households for young men (left) and women (right).

between housing prices and the proportion of single-person households (Bugallo et al., 2024b). Thus, lower proportions are estimated for the Catalan Coast, Madrid, Balearic Islands and Málaga. In other words, the Spanish provinces with the highest average prices.

Figure 2.5 maps the corresponding RRMSEs of the predictions in Figure 2.4. Looking at the percentages, the accuracy of our results is statistically reasonable, with RRMSEs below 30% in most domains, and exceeding that only in those where the predicted proportions are rather small, which is a pretty good accuracy for a SAE problem.

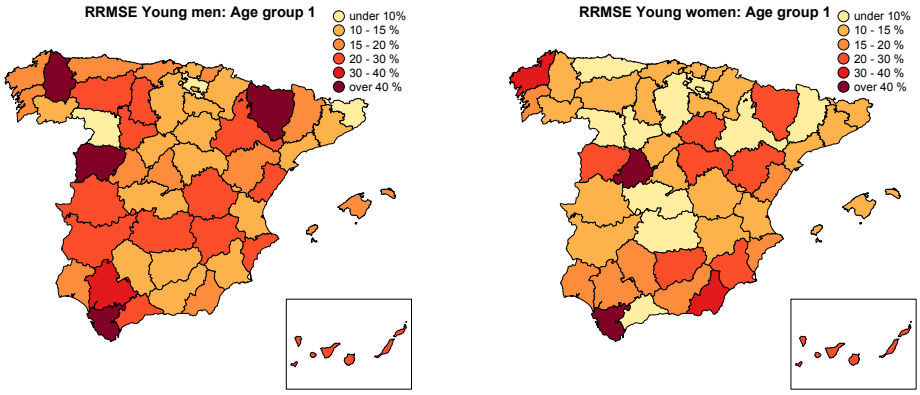


Figure 2.5: RRMSE of the IN predicted proportions for young men (left) and women (right).

Additional results available online at *SORT-Statistics and Operations Research Transactions* [1], include maps of the proportion of single-person households for all age group and sex crossings, and RRMSE estimates. Further charts and analysis for the application to real data are also covered, including a thorough validation of the zero-inflated structure and a

[1] https://www.idescat.cat/sort/sort481/48.1.4.Bugallo-etal.prov.pdf; accessed on: November 4, 2024.

discussion of the need for its inclusion in the case study. On the whole, the evidence suggests that living alone is a common housing choice in all age groups, influenced by marital separation, the emancipation of children, cohabitation and lifestyle in general. Moreover, differences in household composition between men and women are more pronounced among the elderly. Declining fertility and increasing longevity will lead to an ageing population, with an overwhelming increase in the proportion of single-person households.

2.2.6 R codes

As for the `R` codes, the GitHub repository https://github.com/mbugallo/aZIP13 (accessed on: November 4, 2024) contains our dataset and computer code, as well as a detailed description of its contents. It includes a README file that provides basic instructions for the correct execution of the available software.

2.3 Area-level zero-inflated Negative Binomial mixed model

This section describes an area-level zero-inflated NB mixed model aimed at deriving predictors of counts in small areas. Starting from Section 2.2 as the initial reference point, the basic distribution is moved from the PO to the NB to accommodate the overdispersion of the target variable. Let y_{ijk} a count variable taking values in $\{0, 1, 2, \dots\}$, $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$. Let $D = IJK$ be the total number of y-values. Let z_{ijk}, $\boldsymbol{x}_{1,ijk} = (x_{1,ijk1}, \dots, x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk} = (x_{2,ijk1}, \dots, x_{2,ijkq_2})$ be latent (non observable) variables and $1 \times q_1$, $q_1 \geqslant 1$, and $1 \times q_2$, $q_2 \geqslant 1$, row vectors containing area-level auxiliary variables, respectively. Let us define the vectors and matrices

$$\boldsymbol{y}_{jk} = \underset{1\leqslant i\leqslant I}{\text{col}}(y_{ijk}),\ \boldsymbol{z}_{jk} = \underset{1\leqslant i\leqslant I}{\text{col}}(z_{ijk}),\ \boldsymbol{y} = \underset{1\leqslant j\leqslant J}{\text{col}}\Big(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{y}_{jk})\Big),\ \boldsymbol{z} = \underset{1\leqslant j\leqslant J}{\text{col}}\Big(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{z}_{jk})\Big),$$

$$\boldsymbol{X}_{a,jk} = \underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{x}_{a,ijk}),\ \ \boldsymbol{X}_a = \underset{1\leqslant j\leqslant J}{\text{col}}\Big(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{X}_{a,jk})\Big),\ \ a = 1, 2.$$

Let $u_{1,j}$, $u_{1,k}$, $u_{2,j}$, $u_{2,k}$ be independent random effects with standard normal distribution. Define the vectors $\boldsymbol{u}_{1,jk} = (u_{1,j}, u_{1,k})'$, $\boldsymbol{u}_{2,jk} = (u_{2,j}, u_{2,k})'$,$\boldsymbol{u}_{jk} = (\boldsymbol{u}'_{1,jk}, \boldsymbol{u}'_{2,jk})'$ and

$$\boldsymbol{u}_1 = \underset{1\leqslant j\leqslant J}{\text{col}}\Big(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{u}_{1,jk})\Big) \sim N_{2JK}(\boldsymbol{0}, \boldsymbol{I}),\ \boldsymbol{u}_2 = \underset{1\leqslant j\leqslant J}{\text{col}}\Big(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{u}_{2,jk})\Big) \sim N_{2JK}(\boldsymbol{0}, \boldsymbol{I}),\ \boldsymbol{u} = (\boldsymbol{u}'_1, \boldsymbol{u}'_2)'.$$

The vectors (y_{ijk}, z_{ijk}) follow an area-level zero-inflated NB mixed model (aZINB) if

$$\begin{aligned} &z_{ijk} \overset{ind}{\sim} \text{BE}(p_{ijk}),\ \ P(y_{ijk} = 0/z_{ijk} = 1) = 1,\ \ y_{ijk}|_{z_{ijk}=0} \sim \text{NB}(r, \mu_{ijk}),\ \text{ i.e.} \qquad (2.7)\\ &P(y_{ijk} = t/z_{ijk} = 0) = \frac{\Gamma(t+r)}{\Gamma(t+1)\Gamma(r)}\left(\frac{\mu_{ijk}}{r+\mu_{ijk}}\right)^t\left(\frac{r}{r+\mu_{ijk}}\right)^r,\ t \in \{0, 1, 2, \dots\}, \end{aligned}$$

where $p_{ijk} \in (0, 1)$, $r > 0$ and $\mu_{ijk} > 0$. In addition, p_{ijk} and μ_{ijk} depend on the area-level auxiliary variables $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, on the model parameters $\boldsymbol{\beta}_1 = (\beta_{11}, \dots, \beta_{1q_1})'$ and

$\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2q_2})'$, and on the standard deviations $\phi_{11} > 0$, $\phi_{12} > 0$, $\phi_{21} > 0$ and $\phi_{22} > 0$ by means of the link functions

$$\begin{aligned}
\text{logit}(p_{ijk}) &= \log\frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k} = \sum_{\ell=1}^{q_1} x_{1,ijk\ell}\beta_{1\ell} + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}, \\
\log(\mu_{ijk}) &= \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k} = \sum_{\ell=1}^{q_2} x_{2,ijk\ell}\beta_{2\ell} + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}.
\end{aligned}$$

Conversely, we have

$$p_{ijk} = \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}, \quad \mu_{ijk} = \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}\}. \tag{2.8}$$

Finally, it is assumed that the vectors $(y_{ijk}, z_{ijk})'$ are independent conditional on $\boldsymbol{u}$ and it is said that they follow an aZINB11 mixed model (Bugallo et al., 2023). The terminology "11" is added to specify that both the BE and NB models have additive random effects in two components, j and k. The proposed model is a mixture model of two mixed submodels. The BE submodel drives the mixture and incorporates the information derived from the excess of zeros. The NB submodel deals with the modelling of count variables. The overdispersion parameter of the NB submodel is denoted by $\gamma = r^{-1} > 0$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \phi_{11}, \phi_{12}, \phi_{21}, \phi_{22})'$ be the vector of model parameters and define $\xi_{ijk} = I(y_{ijk} = 0)$. From the properties of the NB distribution, it holds that

$$\begin{aligned}
P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) &= \xi_{ijk}\Big[p_{ijk} + (1-p_{ijk})\exp\{r\log r - r\log(r+\mu_i)\}\Big] \\
&\quad + (1-\xi_{ijk})\Bigg[(1-p_{ijk})\exp\Big\{y_{ijk}\log\mu_{ijk} - (y_{ijk}+r)\log(r+\mu_{ijk}) + \\
&\quad r\log r + \log\frac{\Gamma(y_{ijk}+r)}{\Gamma(y_{ijk}+1)\Gamma(r)}\Big\}\Bigg] \\
&= \big(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}\big)^{-1}\Bigg\{\xi_{ijk}\Bigg[\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\} \\
&\quad + \exp\Big\{r\log r - r\log\big(r+\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}\}\big)\Big\}\Bigg] \\
&\quad + (1-\xi_{ijk})\exp\Big\{y_{ijk}(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}) \\
&\quad - (y_{ijk}+r)\log\big(r+\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}\}\big) + r\log r + \log\frac{\Gamma(y_{ijk}+r)}{\Gamma(y_{ijk}+1)\Gamma(r)}\Big\}.
\end{aligned}$$

It follows from the independence assumptions that

$$P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) = \prod_{j=1}^{J}\prod_{k=1}^{K} P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}), \quad P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) = \prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}).$$

Therefore, the likelihood function of the aZINB11 mixed model is

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{4JK}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \prod_{j=1}^{J}\prod_{k=1}^{K}\int_{\mathbb{R}^4}\prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_4(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk}, \quad (2.9)$$

and the respective log-likelihood function is

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\log\int_{\mathbb{R}^4}\prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_4(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk}. \quad (2.10)$$

Given $\boldsymbol{y}$, the ML parameter estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{\theta};\boldsymbol{y}), \quad \Theta = \mathbb{R}^{q_1+q_2}\times\mathbb{R}^4_+.$$

Appendix A describes the ML-Laplace algorithm to maximize $\ell(\boldsymbol{\theta};\boldsymbol{y})$ and calculate the ML estimators of the model parameters. This algorithm also gives modal predictors of random effects. As for the inference procedures for the ML estimators, we rely on both asymptotic (Appendix A) and resampling methods (Section 2.3.2).

2.3.1 Small area prediction of expected counts

This section is devoted to the development of small area predictors of expected counts based on the aZINB11 mixed model (2.7)-(2.8). Let $i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K$ and define

$$\mu_{yijk} \triangleq E[y_{ijk}|\boldsymbol{u}_{jk}] = (1 - p_{ijk}(\boldsymbol{\theta}_1, \boldsymbol{u}_{1,jk}))\mu_{ijk}(\boldsymbol{\theta}_2, \boldsymbol{u}_{2,jk}),$$

where $p_{ijk} \triangleq p_{ijk}(u_{1,k})$ and $\lambda_{ijk} \triangleq \lambda_{ijk}(u_{2,ijk})$ are defined in (2.8).

The plug-in predictor of $\mu_{y_{ijk}}$ is

$$\widehat{\mu}^{in}_{y_{ijk}} = (1 - p_{ijk}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{u}}_{1,jk}))\mu_{ijk}(\widehat{\boldsymbol{\theta}}_2, \widehat{\boldsymbol{u}}_{2,jk}),$$

where $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ are estimators (e.g. ML parameter estimators) of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, and $\widehat{\boldsymbol{u}}_{1,jk}$ and $\widehat{\boldsymbol{u}}_{2,jk}$ are predictors (e.g., log-likelihood modes) of $\boldsymbol{u}_{1,jk}$ and $\boldsymbol{u}_{2,jk}$, respectively. Depending on its purpose, the plug-in predictor could also be used as a forecasting tool for future average counts. For example, when the time component to be predicted in the future is determined by the i-index. To calculate $\widehat{\mu}^{in}_{yijk}$, we use model parameter estimators and random effect predictions, which only depend on the indexes j and k. Nevertheless, one must specify a prediction scenario determined by the values assumed for $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$.

2.3.2 Bootstrap inference

This section presents bootstrap-based CIs for the model parameters and estimators of the MSE of the plug-in predictor. Let θ_ℓ be a component of $\boldsymbol{\theta}$ and $\alpha \in (0,1)$. The following procedure calculates a $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ and a parametric bootstrap estimator of $MSE(\widehat{\mu}_{yijk})$, where $\widehat{\mu}_{yijk}$ is the plug-in predictor, although the algorithm described below

applies to any model-based predictor derived from the proposed aZINB11 mixed model. The procedure also provides a parametric bootstrap estimator of the marginal variance $\text{var}(y_{ijk})$ and prediction intervals (PI) for expected counts (Bugallo et al., 2023).

1. Fit the model and calculate the ML parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \widehat{\phi}_{11}, \widehat{\phi}_{12}, \widehat{\phi}_{21}, \widehat{\phi}_{22})'$.

2. Let $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$.

   Repeat B times ($b = 1, \ldots, B$):

   (a) Generate $u_{1,j}^{*(b)} \sim N(0,1)$, $u_{1,k}^{*(b)} \sim N(0,1)$, $u_{2,j}^{*(b)} \sim N(0,1)$ and $u_{2,k}^{*(b)} \sim N(0,1)$. Then calculate

   $$p_{ijk}^{*(b)} = \exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_{11}u_{1,j}^{*(b)} + \widehat{\phi}_{12}u_{1,k}^{*(b)}\}\big(1 + \exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_{11}u_{1,j}^{*(b)} + \widehat{\phi}_{12}u_{1,k}^{*(b)}\}\big)^{-1},$$
   $$\mu_{ijk}^{*(b)} = \exp\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_{21}u_{2,j}^{*(b)} + \widehat{\phi}_{22}u_{2,k}^{*(b)}\}.$$

   (b) Generate $z_{ijk}^{*(b)} \sim \text{BE}(p_{ijk}^{*(b)})$. If $z_{ijk}^{*(b)} = 1$, $y_{ijk}^{*(b)} = 0$. Otherwise, $y_{ijk}^{*(b)} \sim \text{NB}(r, \mu_{ijk}^{*(b)})$.

   (c) On the basis of the bootstrap sample, $(y_{ijk}^{*(b)}, \boldsymbol{x}_{ijk})$, calculate the ML parameter estimator $\widehat{\theta}_\ell^{*(b)}$, the bootstrap estimate $\widehat{\boldsymbol{\theta}}^{*(b)}$ and the predictor $\widehat{\mu}_{yijk}^{*(b)}$.

3. Sort the values $\widehat{R}_\ell^{*(b)} = D^{1/2}\big(\widehat{\theta}_\ell^{*(b)} - \widehat{\theta}_\ell\big)$, $b = 1, \ldots, B$, from smallest to largest. They are $\widehat{R}_{\ell(1)}^* \leqslant \ldots \leqslant \widehat{R}_{\ell(B)}^*$. A $(1-\alpha)\%$ basic percentile bootstrap CI for θ_ℓ is

   $$\big(\widehat{\theta}_\ell - D^{-1/2}\widehat{R}_{\ell(\lfloor(\alpha/2)B\rfloor)}^*,\ \widehat{\theta}_\ell + D^{-1/2}\widehat{R}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}^*\big).$$

4. To estimate error measures, define

   $$mse^*(\widehat{\mu}_{y_{ijk}}) = \frac{1}{B}\sum_{b=1}^{B}\big(\widehat{\mu}_{y_{ijk}}^{*(b)} - \mu_{y_{ijk}}^{*(b)}\big)^2,\ rmse^*(\widehat{\mu}_{y_{ijk}}) = \big(mse^*(\widehat{\mu}_{y_{ijk}})\big)^{1/2},$$
   $$rrmse^*(\widehat{\mu}_{y_{ijk}}) = \frac{rmse^*(\widehat{\mu}_{y_{ijk}})}{\widehat{\mu}_{y_{ijk}}},\ \bar{y}_{ijk}^* = \frac{1}{B}\sum_{b=1}^{B} y_{ijk}^{*(b)},\ \text{var}^*(y_{ijk}) = \frac{1}{B-1}\sum_{b=1}^{B}\big(y_{ijk}^{*(b)} - \bar{y}_{ijk}^*\big)^2,$$

   A $(1-\alpha)\%$ PI of y_{ijk} is

   $$PI_{ijk}^{\alpha} = \Big(\widehat{\mu}_{y_{ijk}} - z_{1-\alpha/2}\big(\text{var}^*(y_{ijk})\big)^{1/2},\ \widehat{\mu}_{y_{ijk}} + z_{1-\alpha/2}\big(\text{var}^*(y_{ijk})\big)^{1/2}\Big). \tag{2.11}$$

2.3.3 Description of the 2002-2015 GFFS monthly data

The case study investigates the applicability of the zero-inflated NB mixed model (2.7)-(2.8) to explain the occurrence of forest fires in Spain between 2002 and 2014 by province and month, and to provide forecasts for 2015. Taking Boubeta et al. (2019) as an initial reference point for modelling and predicting forest fires, we propose the following innovations and improvements, which have not yet been considered. First, the basic distribution is shifted

from the PO to the NB to account for the overdispersion of the target variable. Second, excess zeros are modelled using a BE mixed model. Third, both models include random effects that vary with month and province, but not with year. This allows modelling the temporal and spatial variability within each year and predicting the number of fires in future years.

Data are from the General Forest Fire Statistics (GFFS) and contain aggregated records by province and month of all forest fires in Spain from 2002 to 2015, both years included, totalling 216,538 events. The dependent variable y_{ijk} counts the number of Spanish forest fires in year i, month j and province k. Therefore, there are $D = IJK = 8400$ domains, defined by the crosses of years ($I = 14$), months ($J = 12$) and provinces ($K = 50$).

Table 2.8 shows the number of forest fires and zeros per year. An exploratory analysis indicates that there are 951 domains in which no forest fires were recorded and their number is not uniform over the years. In fact, the number of forest fires varies from year to year and a clear change in the pattern of forest fires is observed at the end of 2006 and 2012, which motivates the inclusion of the categorical auxiliary variable *year3* (*year3.1*: time interval [2002, 2006]; *year3.2*: time interval [2007, 2012]; *year3.3*: time interval [2013, 2015]).

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Zeros	107	112	63	51	90	52	60	47	89	44	55	60	77	44
Total	19929	18616	21393	25492	16334	10936	11654	15641	11721	16414	15997	10797	9805	1180

Table 2.8: Annual grouping of domains without observed forest fires (zeros per year). Domains are defined as crossings between years, months and provinces.

Table 2.9 shows the number of forest fires and zeros per month, with late autumn and the whole of winter being a period of low fire activity, in contrast to the summer period.

	Jan.	Feb.	Mar.	April	May	Jun	July	Aug.	Sept.	Oct.	Nov.	Dec.
Zeros	177	112	52	40	22	20	5	6	9	59	206	243
Total	6372	19270	30432	16701	11844	18494	28912	37742	26124	12033	4343	4271

Table 2.9: Monthly grouping of domains without observed forest fires (zeros per month). Domains are defined as crossings between years, months and provinces.

Meteorological variables have been part of the set of auxiliary variables in many fire risk studies. In the current research, the process of selecting domain-level auxiliary variables is divided into two stages and they are described in Table 2.10. A spatial analysis search is used to select those automatic meteorological stations that best represent the climatological conditions of each province, and then data are collected from the selected stations. The auxiliary variables contain aggregated information at monthly level, so their effects on the target variable are, in some cases, limited to the context. In addition, if a fire starts at the end of the month and lasts several days, most of which are in the following month, the count corresponds to the previous month, which distorts the available information. As a result, fitting fire models with aggregated monthly and provincial data is generally a difficult problem. But in spite of this, a model that provides an acceptable solution would allow us to predict the number of future forest fires based on *easy-to-predict* climatological information.

Variable	Description	Units	Variable	Description	Units
e	average vapor pressure	tenths of hPa	*p.mes*	total precipitation	mm
hr	average relative humidity	tenths of mm	*q.mar*	mean sea-level pressure	KPa
n.fog	foggy days	% days	*q.max*	max. absolute pressure	KPa
n.gra	hail days	% days	*q.med*	average pressure	KPa
n.llu	rainy days	% days	*q.min*	max. min. pressure	KPa
n.nie	snowy days	% days	*ta.max*	absolute max. temperature	°C
np.001	precipitation $\geqslant$0.1mm	% days	*ta.min*	absolute min. temperature	°C
np.010	precipitation $\geqslant$1mm	% days	*ti.max*	lowest max. temperature	°C
np.300	precipitation $\geqslant$30mm	% days	*tm.max*	average max. temperature	°C
nt.00	min. temperature $\leqslant$0°C	% days	*tm.mes*	average temperature	°C
nt.30	max. temperature $\geqslant$30°C	% days	*tm.min*	average min. temperature	°C
n.tor	storm days	% days	*ts.min*	highest min. temperature	°C
nw.55	wind speed $\geqslant$55km/h	% days	*w.med*	average speed elaborated from 07, 13, 18 UTC	km/h
nw.91	wind speed $\geqslant$91km/h	% days	*unemp*	unemployment rate	%
p.max	max. daily precipitation	mm	*year3*	year group variable	–

Table 2.10: Two-column description and units of the domain-level auxiliary variables used in the application to the 2002-2015 GFFS monthly data.

2.3.4 Application to the 2002-2015 GFFS monthly data

The problem addressed below is to model the number of forest fires in Spain between 2002 and 2014 by province and month, providing error measures and point and interval forecasts for 2015. Due to seasonality, there are provinces where the number of fires is zero in some months and overdispersed in others. In addition, the Mediterranean countries have a high number of fires, but they are mainly concentrated in the summer months. Fortunately, zero-inflated NB mixed models are well suited to this type of data, as they describe patterns that explain both the number of fires and their non-occurrence. Based on this insight, we fit the aZINB11 mixed model (2.7)-(2.8) to the forest fire data described in Section 2.3.3, with month-dependent random effects $u_{1,j}$, $u_{2,j}$, $j = 1, \ldots, J$, and province-dependent random effects $u_{1,k}$, $u_{2,k}$, $k = 1, \ldots, K$.

Tables 2.11 and 2.12 show the ML parameter estimators of the model parameters $\boldsymbol{\beta}_1$, ϕ_1 (BE submodel) and $\boldsymbol{\beta}_2$, ϕ_2 (NB submodel), the p-values to test $H_0 : \beta_{t\ell} = 0$, $t = 1, 2$, $\ell = 1, ..., q_t$, and $H_0 : \phi_t = 0$, $t = 1, 2$, and the normal-asymptotic and bootstrap CIs at the 95% confidence level. For convenience, their lower (LB) and upper (UB) bounds are provided. Normal-asymptotic CIs are discussed in Appendix A and bootstrap CIs in Section 2.3.2. The models are fitted to data from 2007 to 2014, keeping 2015 to test near future retroconditions.

		BE submodel							
		β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	ϕ_{11}	ϕ_{12}
	Estimate	-14.353	0.192	0.143	-0.132	-1.048	-0.754	0.103	1.872
	p-value	0.000	0.000	0.003	0.001	0.000	0.005	0.000	0.000
Asymp	LB 95%	-19.675	0.138	0.050	-0.208	-1.477	-1.285	0.004	1.424
	UB 95%	-9.030	0.246	0.236	-0.056	-0.619	-0.223	2.940	2.460
Boot	LB 95%	-17.733	0.158	0.070	-0.175	-1.367	-1.253	0.000	1.319
	UB 95%	-11.212	0.232	0.213	-0.094	-0.768	-0.311	0.297	2.305

Table 2.11: Model parameters of the final aZINB11 mixed model for the BE submodel. Model fitted with 2002-2014 data aggregated by province and month.

The final model incorporates only those variables that are significant at 1%. The BE submodel contains $q_1 = 6$ covariables: $x_{1,1}$ = intercept, $x_{1,2} = hr$, $x_{1,3} = np.300$, $x_{1,4} = ta.max$, $x_{1,5} = year3.2$, $x_{1,6} = year3.3$. The NB submodel contains $q_2 = 18$ covariables: $x_{2,1}$ = intercept, $x_{2,2} = e$, $x_{2,3} = hr$, $x_{2,4} = n.llu$, $x_{2,5} = n.nie$, $x_{2,6} = np.300$, $x_{2,7} = nt.00$, $x_{2,8} = nw.55$, $x_{2,9} = nw.91$, $x_{2,10} = q.mar$, $x_{2,11} = q.max$, $x_{2,12} = q.min$, $x_{2,13} = ta.max$, $x_{2,14} = ta.min$, $x_{2,15} = tm.mes$, $x_{2,16} = tm.min$, $x_{2,17} = year3.2$, $x_{2,18} = year3.3$.

		NB submodel									
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}	β_{28}	β_{29}	β_{210}
	Estimate	-10.509	0.009	-0.045	-0.014	-0.032	-0.029	0.016	0.018	-0.025	0.146
	p-value	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000
Asymp	LB 95%	-17.960	0.007	-0.051	-0.016	-0.039	-0.043	0.014	0.015	-0.041	0.072
	UB 95%	-3.058	0.012	-0.038	-0.012	-0.025	-0.015	0.018	0.021	-0.009	0.220
Boot	LB 95%	-12.409	0.009	-0.046	-0.015	-0.034	-0.034	0.016	0.017	-0.030	0.125
	UB 95%	-8.227	0.010	-0.043	-0.014	-0.030	-0.024	0.017	0.019	-0.020	0.165

		β_{211}	β_{212}	β_{213}	β_{214}	β_{215}	β_{216}	β_{217}	β_{218}	ϕ_{21}	ϕ_{22}	γ
	Estimate	-0.057	0.049	0.053	0.025	0.064	-0.150	-0.180	-0.272	0.343	1.156	2.151
	p-value	0.003	0.009	0.000	0.001	0.005	0.000	0.000	0.000	0.000	0.000	0.000
Asymp	LB 95%	-0.095	0.012	0.041	0.011	0.020	-0.191	-0.222	-0.331	0.226	0.948	2.063
	UB 95%	-0.020	0.086	0.064	0.040	0.109	-0.109	-0.139	-0.213	0.519	1.408	2.244
Boot	LB 95%	-0.068	0.039	0.050	0.022	0.055	-0.158	-0.189	-0.286	0.177	0.914	2.007
	UB 95%	-0.046	0.060	0.056	0.029	0.074	-0.140	-0.171	-0.257	0.451	1.370	2.755

Table 2.12: Model parameters of the final aZINB11 mixed model for the NB submodel. Model fitted with 2002-2014 data aggregated by province and month.

The effect of the auxiliary variables derived from Tables 2.11 and 2.12 is consistent with the results obtained in previous studies in which arsonists wait for optimal conditions (a window of opportunity) to start a fire (Marcos et al. (2015); Russo et al. (2017)). In general, the LBs,

UBs and widths of the bootstrap CIs are similar to those of the asymptotic CIs, suggesting that the theoretical distribution is close to the asymptotic one. Furthermore, although the interpretation of the estimated model parameters is reasonable, it is noteworthy that both their sign and magnitude often change, sometimes substantially, so it is important to keep in mind that the results are conditioned by the proposed model. Consequently, the mathematical modelling is based on the actual situation. Other auxiliary variables, time periods or any changes in the database may lead to different results, both in sign and relevance.

To validate the fitted model and detect outliers, we analysed the behaviour of the model residuals. Let us define the raw residuals (RR) as

$$\widehat{e}_{ijk} = y_{ijk} - \widehat{\mu}^{in}_{yijk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

Standardized residuals (SR) are defined by dividing the RRs by its standard deviation, i.e.

$$\widehat{e}_{ijk}\nu^{-1}, \quad \text{where} \quad \nu = \left(\frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\widehat{e}_{ijk} - \widehat{\overline{e}}_{...} \right)^2 \right)^{\frac{1}{2}}, \quad \widehat{\overline{e}}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \widehat{e}_{ijk}.$$

Figure 2.6 plots the SRs of the aZINB11 mixed model against domain indexes (left), plug-in predicted values (center) and plug-in log-predicted values (right). It can be seen that the SRs fluctuate around zero, although there are more positive large residuals than negative ones. The cause of this asymmetric behaviour is the underprediction of the model in provinces where the number of observed forest fires in summer was extremely high, as we will see later. Similarly, in the central plot, a small percentage of domains have large predicted values that exceed the threshold of 400. Finally, plotting the SRs against the log-predicted values allowed us to detect a conical pattern in the scatterplot, maintaining their positive asymmetry, which is accentuated as the abscissa axis increases. As the log-predicted values increase, the variability of the residuals also increases. This phenomenon is consistent with the theoretical overdispersion of the aZINB11 model.

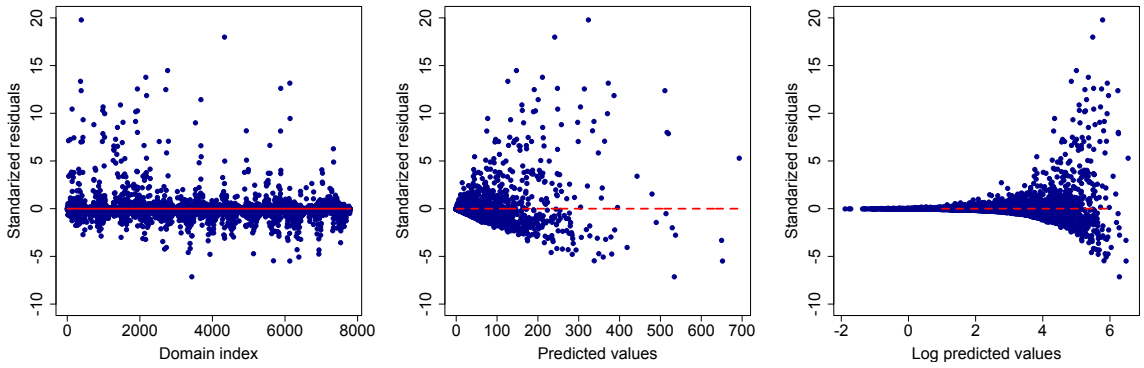


Figure 2.6: SRs versus domain indexes (left) and predicted values of the expected counts of forest fires in 2002-2014 in original (center) and log-scale (right).

Figure 2.7 shows boxplots of the SRs by year, month and province. They fluctuate around zero, mostly in the interval $[-3, 3]$. However, there are more large positive SRs than negative

ones, suggesting underprediction in some provinces. In this sense, there are six provinces with absolute SRs greater than 3. This gives a total of 82 domains. These are the four provinces of Galicia and the two neighbouring autonomous communities in north-western Spain: A Coruña (18), Lugo (5), Ourense (22), Pontevedra (17), Asturias (14) and Cantabria (6).

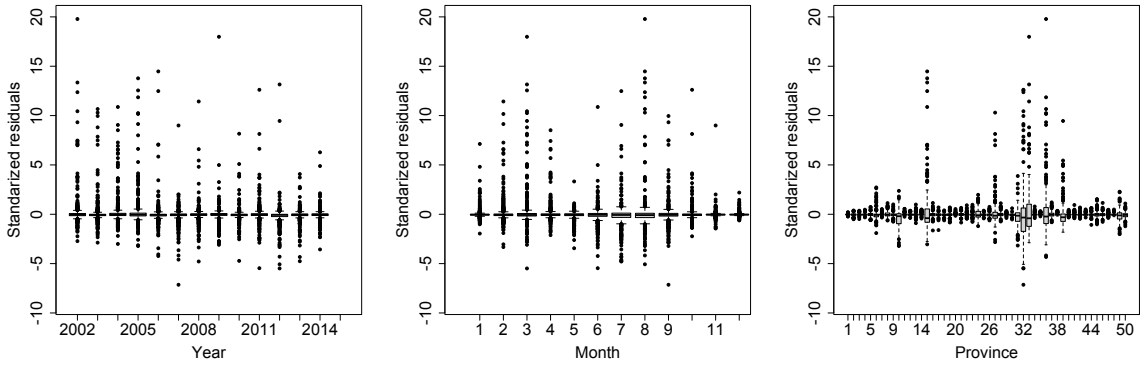


Figure 2.7: Boxplot of SRs by year (left), month (center) and province (right).

The prediction of the number of forest fires in Spain for 2015 ($I = 14$) is discussed below. The idea is to calculate predictions for future horizons in order to optimise forest fire prevention tasks and allocate available resources efficiently. The prediction scenario is the actual 2015 scenario, i.e. the recorded covariates $\boldsymbol{x}_{1,Ijk}$ and $\boldsymbol{x}_{2,Ijk}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, are used, so that the results are actually retroconditions. As the observed forest fire counts y_{Ijk} are available for 2015, the accuracy of the retroconditions can be tested.

For the sake of illustration, Figure 2.8 maps the observed values (left) and plug-in forecasts (center) for July 2015 to analyze the discrepancies in the provincial distribution of forest fires and evaluate the predictive performance. Specifically, the map on the left shows the recorded values of the count variable and the one in the center shows the plug-in predictions. The map on the right allows to widecheck the accuracy of the retroconditions, displaying RRMSE estimates that have been calculated with the algorithm proposed in Section 2.3.2. In order to strike a balance between the approximation capability of the Monte Carlo method and its computational cost, $B = 500$ bootstrap replicates have been used.

Comparing the left and center maps in Figure 2.8, the aZIBN11 mixed model accurately detects the provinces with the highest fire probabilities and reproduces the pattern of fire spread along the Iberian Peninsula. An artificial diagonal line divides the maps into two zones: the north-west, with many fires and low RRMSEs, and the south-east, with few fires and high RRMSEs. The lower the number of forest fires, the higher the RRMSE, because it is challenging to fit a model when the distribution of data across provinces is so uneven, leading to inaccuracies in domains with few events. Nevertheless, it is desirable to predict better in those domains that are more conflictive and with a higher number of forest fires, given that the severity of the environmental problem is greater in those areas. To overcome this challenge, PIs defined in equation (2.11) are calculated.

In addition, let us define the relative squared prediction error (RSPE) for provinces and

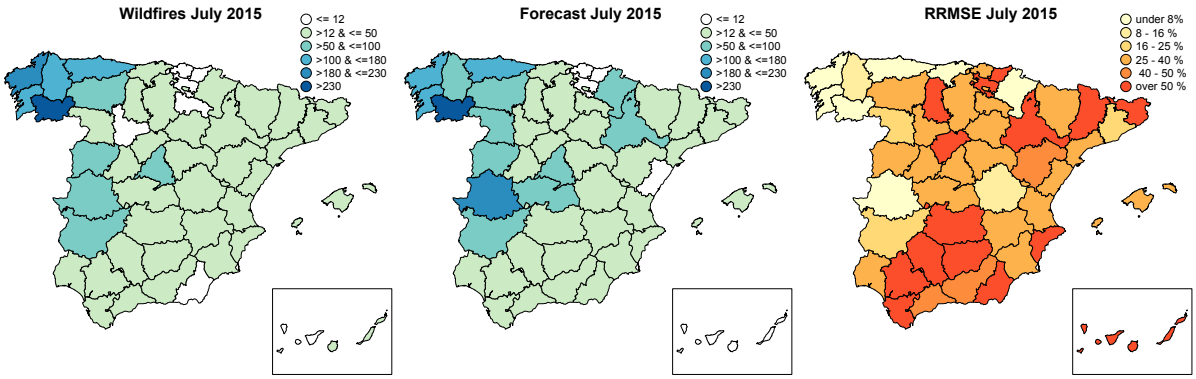


Figure 2.8: Observed (left) and predicted (center) plug-in forest fires and RRMSE estimates (right) in July 2015. The prediction scenario is the actual scenario for 2015.

the province coverage probabilities for 2015 as

$$RSPE_{I.k} = \frac{\sqrt{\sum_{j=1}^{12}(y_{Ijk} - \widehat{\mu}_{Ijk}^{in})^2}}{\sum_{j=1}^{12} y_{Ijk}},\ C_{I.k}^{\alpha} = \frac{1}{12}\sum_{j=1}^{12} C_{Ijk}^{\alpha},\ C_{Ijk}^{\alpha} = I\big(y_{ijk} \in PI_{Ijk}^{\alpha}\big),\ k = 1, \ldots, K,$$

and the RSPE for months and the month coverage probabilities for 2015 as

$$RSPE_{Ij.} = \frac{\sqrt{\sum_{k=1}^{50}(y_{Ijk} - \widehat{\mu}_{Ijk}^{in})^2}}{\sum_{k=1}^{50} y_{Ijk}},\ C_{Ij.}^{\alpha} = \frac{1}{50}\sum_{k=1}^{K} C_{Ijk}^{\alpha},\ C_{Ijk}^{\alpha} = I\big(y_{Ijk} \in PI_{Ijk}^{\alpha}\big),\ j = 1, \ldots, J.$$

Figure 2.9 presents data on provincial $RSPE_{I.k}$ values (left) and provincial coverage probabilities (right) in 2015. As a result, it can be seen that the RSPE values are high in the north of Spain, where the number of recorded forest fires is unusually high in winter 2015, and low in the south (Andalucía), where not many events have been observed or predicted. The opposite applies to the provincial coverage probabilities in Figure 2.9 (right). To provide further evidence for the previous point, Table 2.13 summarizes the $RSPE_{Ij.}$ values (top) and the monthly coverage probabilities (bottom) in 2015. The percentage RSPE values are low in the months with the highest forest fire probabilities (July - September), and rise in spring and autumn. For winter, the anomalous observations reported in December 2015 justify the high value of this relative discrepancy measure, greater than 45%. This is an atypical value as it has been, in fact, an atypical month. In short, the average percentage RSPE for 2015 is 20.34%, so that its four quarterly averages are 17.22%, 21.11%, 11.86% and 31.17%, respectively. In terms of coverage probabilities, the same results as in Figure 2.9 can be extracted.

The coverage probabilities provide encouraging results at provincial and monthly level. For the provinces, the maps show that coverage is 100% for most of them. On a monthly basis, the situation changes slightly due to the anomalous behaviour of Northwest Spain (see Figure 1.1), but coverage is around 85-90% in almost all months. This case study is intended to serve as an example for future applications of forest fire modelling. Current and future

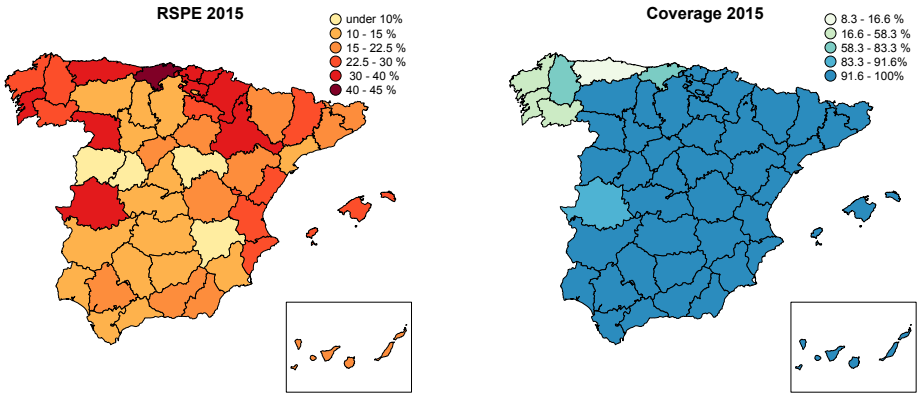


Figure 2.9: RSPE values and coverage probabilities for Spanish provinces in 2015, both in %.

	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
$RSPE_{Ij.}$	16.96	20.30	14.42	20.57	18.43	24.32	10.96	9.70	14.94	27.61	19.04	46.86
$C^{\alpha}_{Ij.}$	94	88	88	90	92	94	98	94	92	94	92	94

Table 2.13: RSPE values and monthly coverage probabilities in 2015, both in %.

fire management in Mediterranean countries requires a paradigm shift. Cooperation between countries is increasingly necessary to face moments of crisis in certain regions. In this respect, old planning systems are no longer effective and a change of scale and of mechanical and human means of extinguishing fires is urgently needed.

It is interesting to note that zero-inflated NB mixed models proved to be flexible tools for describing the behaviour and predicting the number of fires in a region over time. The chosen model has a reasonable interpretation from a forestry point of view, showing dependence and correlation relationships consistent with those published in the scientific literature on fire occurrence modelling. The developed forecasting tool is also useful when applied to the forecasting period. Regarding the improvements of the current research on the statistical methodology for the analysis of forest fires, it provides a forecasting tool that is able to identify values with a 95% confidence in the real data analysed (i.e. retroconditions). In addition, a zero-inflated structure is added, providing a means to deal with areas with very different climatological and socio-economic conditions in relation to their arson activity.

2.3.5 R codes

As for the `R` codes, the GitHub repository https://github.com/mbugallo/aZINB11Fires (accessed on: November 4, 2024) contains our dataset and computer code, as well as a detailed description of its contents. It includes a README file that provides basic instructions for the correct execution of the available software.

2.4 Area-level zero-inflated Gamma mixed model

This section describes an area-level zero-inflated GA mixed model aimed at deriving predictors of averages and totals in small areas for non-negative continuous variables. Let us consider a continuous random variable y_{ijk} taking values on $[0,\infty)$, where $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$. Let $D = IJK$ be the total number of y-values. Let z_{ijk}, $\boldsymbol{x}_{1,ijk} = (x_{1,ijk1},\ldots,x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk} = (x_{2,ijk1},\ldots,x_{2,ijkq_2})$ be latent (non observable) variables and $1\times q_1$, $q_1 \geqslant 1$, and $1\times q_2$, $q_2 \geqslant 1$, row vectors of area-level auxiliary variables, respectively. Let us define the vectors and matrices

$$\boldsymbol{y}_{jk} = \underset{1\leqslant i\leqslant I}{\text{col}}(y_{ijk}),\ \boldsymbol{z}_{jk} = \underset{1\leqslant i\leqslant I}{\text{col}}(z_{ijk}),\ \boldsymbol{y} = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{y}_{jk})),\ \boldsymbol{z} = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{z}_{jk})),$$

$$\boldsymbol{X}_{1,jk} = \underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{x}_{1,ijk}),\ \boldsymbol{X}_{2,jk} = \underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{x}_{2,ijk}),$$

$$\boldsymbol{X}_1 = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{X}_{1,jk})),\ \boldsymbol{X}_2 = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{X}_{2,jk})).$$

Let be $\boldsymbol{u}_{jk} = (u_{1,jk}, u_{2,jk})'$, with $u_{1,jk}$, $u_{2,jk}$ independent $N(0,1)$ random effects, and

$$\boldsymbol{u}_1 = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(u_{1,jk})) \sim N_{JK}(\boldsymbol{0},\boldsymbol{I}),\ \boldsymbol{u}_2 = \underset{1\leqslant j\leqslant J}{\text{col}}(\underset{1\leqslant k\leqslant K}{\text{col}}(u_{2,jk})) \sim N_{JK}(\boldsymbol{0},\boldsymbol{I}),\ \boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'.$$

The vectors (y_{ijk}, z_{ijk}) follow an area-level zero-inflated GA mixed model (aZIG) if

$$\begin{aligned} &z_{ijk} \sim \text{BE}(p_{ijk}),\ \ P(y_{ijk} = 0/z_{ijk} = 1) = 1, \\ &f(y_{ijk} = t/z_{ijk} = 0) = \exp\left\{-\nu\mu_{ijk}^{-1} y_{ijk} - \nu\log\mu_{ijk} + (\nu-1)\log y_{ijk} + \nu\log\nu - \log\gamma(\nu)\right\}, \end{aligned} \tag{2.12}$$

where $p_{ijk} \in (0,1)$, $\nu > 0$, $t > 0$ and $\mu_{ijk} > 0$. In addition, p_{ijk} and μ_{ijk} depend on the area-level auxiliary variables $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, on the model parameters $\boldsymbol{\beta}_1 = (\beta_{11},\ldots,\beta_{1q_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{21},\ldots,\beta_{2q_2})'$, and on the standard deviations $\phi_1 > 0$ and $\phi_2 > 0$ by means of the link functions

$$\begin{aligned} \text{logit}(p_{ijk}) &= \log\frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,jk} = \sum_{\ell=1}^{q_1} x_{1,ijk\ell}\beta_{1\ell} + \phi_1 u_{1,jk}, \\ \log(\mu_{ijk}) &= \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,jk} = \sum_{\ell=1}^{q_2} x_{2,ijk\ell}\beta_{2\ell} + \phi_2 u_{2,jk}. \end{aligned}$$

Inverting the above functions, it follows that

$$p_{ijk} = \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,jk}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,jk}\}},\quad \mu_{ijk} = \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,jk}\}. \tag{2.13}$$

Conditioned on $\boldsymbol{u}$, it is assumed that the vectors $(y_{ijk}, z_{ijk})'$ are independent and it is said that they follow an aZIG22 mixed model (Bugallo et al., 2024c). The terminology "22" is added to specify that both the BE and GA models have multiplicative random effects in two components, j and k. The proposed model is a mixture model of two mixed submodels. The BE submodel drives the mixture and incorporates the information derived from the excess of zeros. The GA submodel deals with strictly positive target values using the GA distribution

with means $\mu_{ijk} > 0$ and constant shape $\nu > 0$, like the normal linear model assumes a constant variance.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \phi_1, \phi_2)'$ be the vector of model parameters and define $\xi_{ijk} = I(y_{ijk} = 0)$. The components of the (continuous) marginal distribution are

$$P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) = \xi_{ijk}p_{ijk} + (1-\xi_{ijk})\Big[(1-p_{ijk})\exp\Big\{-\nu\mu_{ijk}^{-1}y_{ijk} - \nu\log\mu_{ijk} + (\nu-1)\log y_{ijk} + \nu\log\nu - \log\gamma(\nu)\Big\}\Big] = \big(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,jk}\}\big)^{-1}\Big\{\xi_{ijk}\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,jk}\} + (1-\xi_{ijk})\exp\Big\{-\nu y_{ijk}\exp\{-\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 - \phi_2 u_{2,jk}\} - \nu(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,jk}) + (\nu-1)\log y_{ijk} + \nu\log\nu - \log\gamma(\nu)\Big\}\Big\}.$$

By the independence assumptions, it follows that

$$P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) = \prod_{j=1}^{J}\prod_{k=1}^{K} P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}), \quad P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) = \prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}).$$

The likelihood and log-likelihood functions of the aZIG22 mixed model are, respectively,

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{2JK}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \prod_{j=1}^{J}\prod_{k=1}^{K}\int_{\mathbb{R}^2}\prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_2(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk},$$

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\log\int_{\mathbb{R}^2}\prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_2(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk}.$$

Given $\boldsymbol{y}$, the ML parameter estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\,\ell(\boldsymbol{\theta};\boldsymbol{y}), \quad \Theta = \mathbb{R}^{q_1+q_2}\times\mathbb{R}_+^2.$$

Appendix A describes the ML-Laplace algorithm to maximize $\ell(\boldsymbol{\theta};\boldsymbol{y})$ and calculate the ML estimators of the model parameters. This algorithm also gives modal predictors of random effects. As for the inference procedures for the ML estimators, we rely on both asymptotic (Appendix A) and resampling methods (Section 2.4.2).

2.4.1 Small area prediction of expected averages

This section is devoted to the development of small area predictors of expected avegares based on the aZIG22 mixed model (2.12)-(2.13). Let $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$ and define

$$\mu_{yijk} \triangleq E[y_{ijk}|\boldsymbol{u}_{jk}] = (1-p_{ijk}(u_{1,jk}))\mu_{ijk}(u_{2,jk}),$$

where $p_{ijk} \triangleq p_{ijk}(u_{1,jk})$ and $\mu_{ijk} \triangleq \mu_{ijk}(u_{2,jk})$ are defined in (2.13).

By plugging ML estimators and modal predictors, the plug-in predictor of μ_{yijk} is

$$\widehat{\mu}^{in}_{yijk} = \left(1 + \exp\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1\widehat{u}_{1,jk}\}\right)^{-1} \exp\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2\widehat{u}_{2,jk}\}.$$

Following the more applied cut-off of this section, where we will look again at wildfire modelling and prediction (see Sections 2.4.3 and 2.4.4), the plug-in predictor is sufficient for our practical purposes. In fact, it is the most convenient choice as it is unrivalled in terms of ease of interpretation and execution time.

2.4.2 Bootstrap inference

In this section we formalise how to compute bootstrap-based CIs for the model parameters and estimators of the MSEs of the predictors. Let θ_ℓ be a component of $\boldsymbol{\theta}$ and $\alpha \in (0,1)$. The following procedure calculates a $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ and a parametric bootstrap estimator of $MSE(\widehat{\mu}^{in}_{yijk})$. It also provides bootstrap estimates for the quantiles of the distribution of the predictor $\widehat{\mu}^{in}_{yijk}$ so as to define risk measures in Section 2.4.4.

1. Fit the model and calculate the ML parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \widehat{\phi}_1, \widehat{\phi}_2)'$.
2. Let $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$.
   Repeat B times ($b = 1, \ldots, B$):
   (a) Generate $u^{*(b)}_{1,jk} \sim N(0,1)$, $u^{*(b)}_{2,jk} \sim N(0,1)$ and calculate
   $$\begin{aligned} p^{*(b)}_{ijk} &= \exp\left\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u^{*(b)}_{1,jk}\right\}\left(1 + \exp\left\{\boldsymbol{x}_{1,ijk}\widehat{\boldsymbol{\beta}}_1 + \widehat{\phi}_1 u^{*(b)}_{1,jk}\right\}\right)^{-1}, \\ \mu^{*(b)}_{ijk} &= \exp\left\{\boldsymbol{x}_{2,ijk}\widehat{\boldsymbol{\beta}}_2 + \widehat{\phi}_2 u^{*(b)}_{2,jk}\right\}. \end{aligned}$$
   (b) Generate $z^{*(b)}_{ijk} \sim \text{BE}(p^{*(b)}_{ijk})$. If $z^{*(b)}_{ijk} = 1$, do $y^{*(b)}_{ijk} = 0$. If $z^{*(b)}_{ijk} = 0$, generate $y^{*(b)}_{ijk} \sim GA(\mu^{*(b)}_{ijk}, \nu)$.
   (c) Calculate $\mu^{*(b)}_{yijk} = (1 - p^{*(b)}_{ijk})\mu^{*(b)}_{ijk}$.
   (d) Based on the sample $(y^{*(b)}_{ijk}, \boldsymbol{x}_{ijk})$, calculate the ML bootstrap estimate $\widehat{\theta}^{*(b)}_\ell$. In addition, calculate the predictor $\widehat{\mu}^{*(b)}_{yijk}$.
3. Sort the values $\widehat{\theta}^{*(b)}_\ell$, $b = 1, \ldots, B$, from smallest to largest. They are $\widehat{\theta}^{*}_{\ell(1)} \leqslant \ldots \leqslant \widehat{\theta}^{*}_{\ell(B)}$. A $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ is $\left(\widehat{\theta}^{*}_{\ell(\lfloor(\alpha/2)B\rfloor)}, \widehat{\theta}^{*}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}\right)$.
4. To estimate error measures, define $mse^*(\widehat{\mu}_{yijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mu}^{*(b)}_{yijk} - \mu^{*(b)}_{yijk}\right)^2$,
   $$rmse^*(\widehat{\mu}_{yijk}) = (mse^*(\widehat{\mu}_{yijk}))^{\frac{1}{2}},\ rrmse^*(\widehat{\mu}_{yijk}) = \frac{rmse^*(\widehat{\mu}_{yijk})}{\widehat{\mu}_{yijk}}.$$
5. Sort the values $\widehat{\mu}^{*(b)}_{yijk}$, $b = 1, \ldots, B$, from smallest to largest. They are $\widehat{\mu}_{yijk(1)} \leqslant \ldots \leqslant \widehat{\mu}_{yijk(B)}$. The bootstrap quantile of the distribution of the predictor $\widehat{\mu}_{yijk}$ that leaves its left-hand probability α is $\widehat{q}_{ijk,\alpha} := \widehat{\mu}_{yijk(\lfloor\alpha B\rfloor)}$.

2.4.3 Description of the 2007-2015 GFFS weekly data

The case study assesses the applicability of the aZIG22 mixed model (2.12)-(2.13) to explain the occurrence of large fires in Spain between 2007 and 2014, by province and week, and provide forecasts for 2015. Data are from the General Forest Fire Statistics (GFFS). The dependent variable y_{ijk} can be either the total burned area (in Ha) of a region during a certain period of time, or its value averaged over the number of reported forest fires, denoted by n_{ijk}. It is said that $\overline{y}_{ijk} = y_{ijk}/n_{ijk}$ denotes an average forest fire. The indexes i, j and k stand for year, week and province, respectively. The application to real data is limited to $K = 41$ Spanish provinces for reasons of data availability (see Figure 1.1). In fact, the light shaded provinces in Figure 1.1 have been excluded from this application to real data because we do not have data for all the explanatory variables we will consider and which are listed in Table 2.16. Furthermore, due to the seasonal nature of the megafires, the study is limited to the months of July, August, September and October, with data collected between the 27th and 44th weeks of $I = 9$ years, so there are $J = 18$ weeks.

Tables 2.14–2.15 are included to illustrate the suitability of using a zero-inflated mixed model. Table 2.14 shows the provincial deciles of the total and average burned area for weeks 27 to 44 and years 2007 to 2015. At least 20% of the data are equal to 0, which is an extremely high percentage for continuous distributions, such as the GA distribution. This motivates the modelling of burned forest areas by also including latent variables to account for excess zeros. Moreover, the total burned area has much more dispersion than its average by the number of fires, as expected. Table 2.15 depicts the proportion of zeros per year. It follows that the zero-inflated structure is stable over the years.

	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
Total	0	0	0	0.200	0.900	2.200	5.080	11.307	27.048	80.092	15256.210
Average	0	0	0	0.190	0.500	0.932	1.600	2.800	5.073	12.305	4674.110

Table 2.14: Deciles of total (top) and average (bottom) monthly and provincial burned areas for weeks 27 to 44 of the years 2007 to 2015.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Proportion	0.257	0.295	0.203	0.276	0.168	0.324	0.251	0.339	0.314

Table 2.15: Proportion of zeros per year, i.e. domains with no reported forest fires. Domains are defined as crossings between years, weeks and provinces.

The area-level auxiliary variables are described in Table 2.16. The climatological variables (first eight rows on the left of Table 2.16) were obtained in Section 2.3.3 for monthly data, and the process is simply extended here to weekly data. In terms of pre-processing, it was decided to standardize the variables of fire extinguishing means and the number of forest fires to avoid problems of location and scale. Thus, although these variables were initially of count type, they are interpreted as measures scaled to the mean and are unitless. The area-level auxiliary variables in Table 2.16 have been grouped according to their description into six categories: climatological (8), distance (2), firefighting staff (4), land-based machinery (4),

aerial machinery (4) and fire count (1).

Variable	Description	Units
Climatological variables (8)		
dir	direction of max. wind speed	°∠
prec	average precipitation	mm
sol	duration of insolation	h
tmax	average max. temperature	°C
tmed	average mean temperature	°
tmin	average min. temperature	°C
wmed	average speed elaborated from 07, 13, 18 UTC	m/s
hr	average relative humidity	tenths of mm
Distance variables (2)		
bui1	distance between the fire and the nearest building	km
bui10	average distance between the fire and the 10 nearest buildings	km
Firefighting staff variables (4)		
tch.st	technicians and/or forestry agents	
brg.st	brigade personnel	
sec.st	state security force	
oth.st	others, e.g. volunteers	
Land-based machinery variables (4)		
bll.eq	no. of bulldozers	
trc.eq	no. of tractors	
oth.eq	no. of other machines	
atb.eq	no. of fire engines	
Aerial machinery variables (4)		
ext.air	no. of firefighting helicopters	
car.air	no. of aircrafts	
tra.air	no. of transport helicopters	
amp.air	no. of amphibious aircraft	
Fire count variable (1)		
n.fir	no. of fires by domain	

Table 2.16: Two-column description and units of the domain-level auxiliary variables used in the application to the 2007-2015 GFFS weekly data.

As a preliminary analysis to the statistical modelling in Section 2.4.4, we applied a clustering method, the K-means algorithm, described in Appendix B. This clustering algorithm allows the identification of domains with a markedly anomalous profile. Table 2.17 shows the center, size and distribution of the average megafires in the clusters. To find a trade-off between interpretability, complexity and variability (measured by comparing within-cluster and between-cluster sums of squares), three clusters were considered (Forgy, 1965). As recommended, random sets of different observations were repeatedly selected as initial centres.

There is one cluster with only 5 observations (0.07%) and a slightly larger cluster with 34 observations (0.50%). The remaining one forms a large cluster with 6603 observations (99.41%). The average burned area in cluster 1 exceeds the threshold of 500 Ha. The same is true for 48% of the observations in cluster 2 and for none in cluster 3. Results show that $0.07 + 0.50 = 0.57\%$ of the studied events concentrate the problem and it is on these that modelling should focus. Thus, the K-means algorithm successfully detects average megafires and supports the need for models capable of quantifying and forecasting the spatio-temporal risk of extreme events. From Table 2.17, average megafires are described as follows:

Cluster	*dir*	*prec*	*sol*	*tmax*	*tmed*	*tmin*	*wmed*	*hr*	*bui1*	*bui10*	*tch.st*	*brg.st*	*sec.st*
1	16.360	1.702	9.660	30.976	24.886	18.808	3.188	53.000	1.292	2.530	12.456	8.700	19.366
2	20.795	1.561	11.311	32.066	24.295	16.530	2.776	44.206	1.884	3.673	4.219	3.716	4.910
3	9.546	8.705	8.912	26.756	20.337	13.918	2.554	57.896	1.086	2.266	-0.031	-0.026	-0.040
Cluster	*oth.st*	*atb.eq*	*bll.eq*	*trc.eq*	*oth.eq*	*amp.air*	*car.air*	*ext.air*	*tra.air*	*n.fir*	$\overline{y}_{ijk}$	*size*	*count*
1	2.272	6.321	7.118	4.139	8.015	8.873	4.711	2.134	10.235	-0.209	3710.843	5	5
2	2.521	3.759	6.064	1.803	1.877	4.560	2.014	2.445	3.681	0.026	656.026	34	16
3	-0.015	-0.024	-0.037	-0.012	-0.016	-0.030	-0.014	-0.014	-0.027	0.000	5.919	6603	0

Table 2.17: Results of the K-means algorithm described in Appendix B. Cluster centres and sizes (*size*) and count (*count*) of average megafires for the 3-group case.

Weather conditions: Low rainfall and humidity, plenty of sunshine, higher than expected temperatures and strong winds. Wind direction is not discriminating.

Distances: Proximity to urban settlements more influential than isolated human buildings.

Firefighting resources (standardized variables): Need for much more personnel, ground equipment and air support. The differences with the average of the extinguishing systems are higher for megafires. Indeed, they skew the mean –not robust enough– and force it to be slightly negative for cluster 3, which captures fires with "more common" patterns.

Simultaneity of fires (standardized variable): The more forest fires, the more virulent they are. In cluster 1, large forest fires occur in a period with fewer events than the average, but the other fires in cluster 2 are generally associated with higher simultaneity.

2.4.4 Application to the 2007-2015 GFFS weekly data

As a follow-up to the study by Bugallo et al. (2023), the aim of this section is to model virulent fires with provincial spatial and weekly time scales, and to define risk measures. Based on the aZIG22 mixed model (2.12)-(2.13), this section analyses the variables *total burned area*, y_{ijk}, and *average burned area* , $\overline{y}_{ijk}$, in year i, week j and province k, accounting for excess zeros. For ease of exposition, we will denote the models aZIG for totals and averages as aZIGT and aZIGA, respectively. Each model aZIG has two submodels, with BE and GA distributions, called the BE-submodel and the GA-submodel, respectively.

Tables 2.18 and 2.19 show the ML parameter estimators of the model parameters $\boldsymbol{\beta}_1$, ϕ_1 (BE submodel) and $\boldsymbol{\beta}_2$, ϕ_2 (GA submodels), the p-values to test $H_0 : \beta_{t\ell} = 0$, $t = 1, 2$, $\ell = 1, ..., q_t$, and $H_0 : \phi_t = 0$, $t = 1, 2$, and the normal-asymptotic and bootstrap CIs at the 95% confidence level. For convenience, their lower (LB) and upper (UB) bounds are provided. Normal-asymptotic CIs are discussed in Appendix A and bootstrap CIs in Section 2.4.2. The models are fitted to data from 2007–2014, keeping 2015 to test near future retroconditions.

The final model incorporates only those variables that are significant at 1%. As a result,

the BE submodel, fitted with the climatological variables of Table 2.16, contains $q_1 = 5$ covariables: $x_{1,1} = prec$, $x_{1,2} = tmax$, $x_{1,3} = tmed$, $x_{1,4} = wmed$, $x_{1,5} = hr$. The GA-submodel of model aZIGT contains $q_2 = 11$ covariates: $x_{2,1} =$, $x_{2,2} = bll.eq$, $x_{2,3} = trc.eq$, $x_{2,4} = amp.air$, $x_{2,5} = car.air$, $x_{2,6} = tra.air$, $x_{2,7} = prec$, $x_{2,8} = tmax$, $x_{2,9} = tmed$, $x_{2,10} = wmed$, $x_{2,11} = hr$. The GA-submodel of model aZIGA contains $q_2 = 7$ covariates: $x_{2,1} = n.fir$, $x_{2,2} = bui1$, $x_{2,3} = prec$, $x_{2,4} = tmax$, $x_{2,5} = tmed$, $x_{2,6} = wmed$, $x_{2,7} = hr$. The estimates of the scale parameters are $\nu_T = 0.808$ and $\nu_A = 0.805$, respectively. There is no fixed intercept in either of the three submodels and the climatological variables of the two GA submodels are the same.

		BE submodel					
		β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	ϕ_1
	Estimate	0.015	-0.285	0.235	-0.447	0.041	1.141
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
Asymp	LB 95%	0.010	-0.339	0.167	-0.559	0.035	1.020
	UB 95%	0.020	-0.232	0.302	-0.335	0.046	1.275
Boot	LB 95%	0.011	-0.351	0.183	-0.555	0.036	1.015
	UB 95%	0.020	-0.248	0.317	-0.330	0.047	1.267

Table 2.18: ML parameter estimators of the model parameters for the BE submodel of models aZIG. Model fitted with 2007-2014 data aggregated by province and week.

It can be observed in Table 2.18 that for both y_{ijk} and $\bar{y}_{ijk}$ the ML-Laplace algorithm returns the same estimates for the BE-submodel. The reason is that the objective function to be maximised is additively separable and the optimization of the BE-submodel summand does not depend on y_{ijk} or $\bar{y}_{ijk}$. Furthermore, the BE-submodel is an area-level BE mixed model (aBE) with random effects dependent on week and province crosses, and stable over years. If we fit the aBE model directly to the binary target variables ξ_{ijk}'s, which indicate the events of zero or more than one forest fire, and apply the ML-Laplace algorithm, we obtain the same estimates as for the completed models aZIGT and aZIGA.

At the same time, we have modelled the target variables taking into account that they are semi-continuous, i.e. with many exact zeros and continuous positive outcomes. Linear models for normally distributed variables are the simplest and most commonly used statistical models. However, linear models are not appropriate for positive variables with asymmetric distributions. Here are some reasons for choosing a GA mixed model for the conditional target data, $y_{ijk}|n_{ijk} > 0$ and $\overline{y}_{ijk}|n_{ijk} > 0$. If the response variable is positively skewed, a model based on the normal distribution does not take place. The GA distribution is appropriate when the response variables take values in $(0, \infty)$, where small values are expected to have small variability and large values are expected to have large variability. The link function of the GA GLMM is logarithmic. One reason is that these models assume multiplicative effects of the predictors on the original outcome and are easier to interpret. This is not the case with the canonical link of the GA GLMM (Lee et al., 2010). In addition, it is suitable to reduce the variability of positive variables with some tiny and some unusually high values. Finally, the usual GA GLMM assumes that the shape parameter is constant, just as the normal linear

		Total burned area - model aZIGT											
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}	β_{28}	β_{29}	β_{210}	β_{211}	$\phi_{2,T}$
	Estimate	0.316	0.384	0.274	0.267	0.171	0.336	0.008	0.356	-0.347	0.301	-0.016	1.347
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Asymp	LB 95%	0.236	0.305	0.221	0.192	0.104	0.266	0.003	0.316	-0.400	0.226	-0.020	1.266
	UB 95%	0.397	0.462	0.326	0.342	0.239	0.407	0.012	0.396	-0.294	0.377	-0.012	1.433
Boot	LB 95%	0.290	0.334	0.217	0.188	0.107	0.295	0.002	0.308	-0.395	0.218	-0.019	1.261
	UB 95%	0.407	0.437	0.305	0.311	0.225	0.414	0.011	0.388	-0.285	0.369	-0.012	1.416

		Average burned area - model aZIGA							
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}	$\phi_{2,A}$
	Estimate	0.243	0.237	0.006	0.115	-0.080	0.331	-0.021	1.023
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Asymp	LB 95%	0.191	0.188	0.002	0.079	-0.128	0.257	-0.025	0.959
	UB 95%	0.296	0.287	0.010	0.152	-0.032	0.406	-0.017	1.091
Boot	LB 95%	0.207	0.173	0.002	0.083	-0.121	0.253	-0.026	0.926
	UB 95%	0.310	0.252	0.011	0.152	-0.027	0.393	-0.018	1.077

Table 2.19: ML parameter estimators of the model parameters for the GA-submodels of models aZIG. Models fitted with 2007-2014 data aggregated by province and week.

model assumes a constant variance. Likewise, we have maintained this assumption.

To validate models aZIGT and aZIGA, we first define the raw residuals (RR) as

$$\widehat{e}_{ijk} = y_{ijk} - \widehat{\mu}^{in}_{yijk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

Standardized residuals (SR) are defined by dividing the RRs by its standard deviation, i.e.

$$\widehat{e}_{ijk}\nu^{-1}, \quad \text{where} \quad \nu = \left(\frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\widehat{e}_{ijk} - \widehat{\bar{e}}_{...} \right)^2 \right)^{\frac{1}{2}}, \quad \widehat{\bar{e}}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \widehat{e}_{ijk}.$$

Up to this point, it is important to make the following comments.

Model aZIGT: Due to the range of the response variable, calculations are restricted to the log-scale. Consequently, the RRs are the differences between the log-observed values and the log-fitted values. For SRs, the sample mean is subtracted and divided by the sample standard deviation. At this point, there is a problem with null counts because the corresponding residuals are not defined. They must be omitted or an artificial value must be assigned to log(0). Since we are interested in zero inflation, both options are problematic. This drawback is a strong argument against the model aZIGT (Bugallo et al., 2024c).

Model aZIGA: Calculations are performed at the original scale and both the RRs and SRs are well defined for all domains. Values are consistent and interpretable.

Because of the advantages and disadvantages mentioned above, we will pay more attention to the aZIGA model, whose estimated model parameters are given in Tables 2.18 and 2.19 (bottom). The aim is to investigate whether there are unfavourable cases for the tolerance limits set and to relate them to the megafires. Of the 5904 observations, only 18 are average megafires, but they are particularly interesting and distort the residual plots strongly. It is safe to say that they must be analysed separately.

Figure 2.10 plots the SRs of the model aZIGA, stressing their magnitude and colouring average megafires in dark blue. As expected, it is challenging to accurately fit such events, which adds great value to our research. Indeed, average megafires are more volatile and their residuals skyrocket. Consequently, the model underpredicts them. In contrast, residuals belonging to null observations are conveniently close to zero. In addition, the majority of the SRs are in the interval $[-3, 3]$. Finally, no patterns are observed over the years.

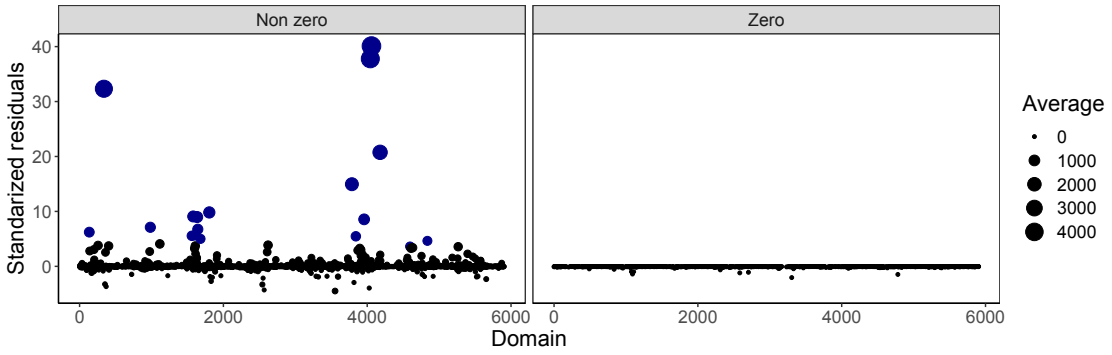


Figure 2.10: SRs of model aZIGA by domain, for non-zero (left) and zero (right) outcomes. Megafires are dark blue and the size of the dots is proportional to the average burned area.

Once the week and the province have been determined and the auxiliary variables are known, the proposed models make it possible to predict the total and average burned area in different provinces. There are two main approaches. One is to assess the goodness-of-fit of the model using the fit period of the data. The other is to work with hypothetical scenarios, i.e. artificial values of the area-level auxiliary variables and simulated target values. However, it is quite difficult to reproduce the variability of the real process, so we would have no guarantee that the results would be close to a realistic background. For this reason, we have decided to set aside 2015 ($I = 9$) and use the available information to make retroconditions for the near future. We have taken the actual scenario of 2015 as the forecast scenario.

Figure 2.11 shows line charts of the observed area-level values and retroconditions of the 2015 forest fires, based on the aZIGT and aZIGA models. The results have been averaged on a provincial basis according to the territorial division of Spain shown in Figure 1.1. Thus, what is shown is the average behaviour of the provinces over time, as a summary measure of the three specified regions (Northwest Spain, Peninsular Center and Mediterranean Coast). We conclude that the averaged observations and retroconditions follow similar patterns over time. Figure 2.11 also indicates that in 2015, the month of July was, on average, the period with the highest number of reported medium to large fires, with a decreasing trend until the last month considered, October. We do not include a non-aggregated plot because the

cardinality of the set of crosses between weeks and provinces clouds the interpretation.

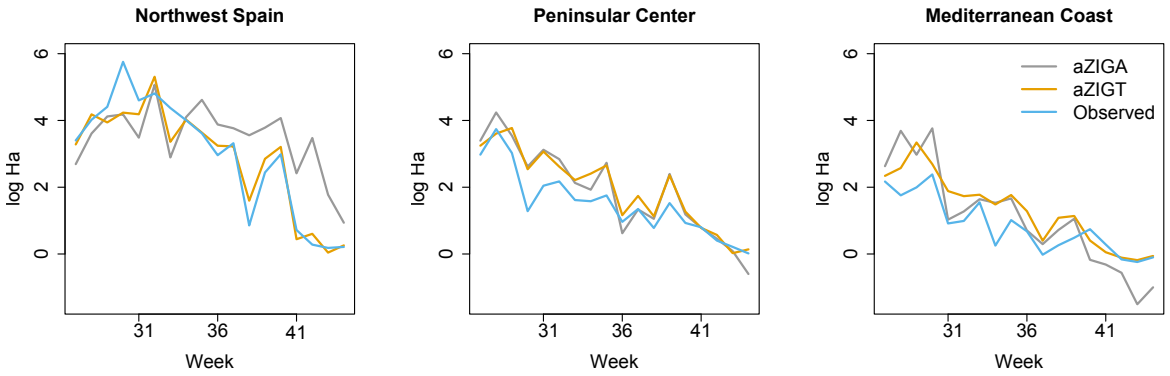
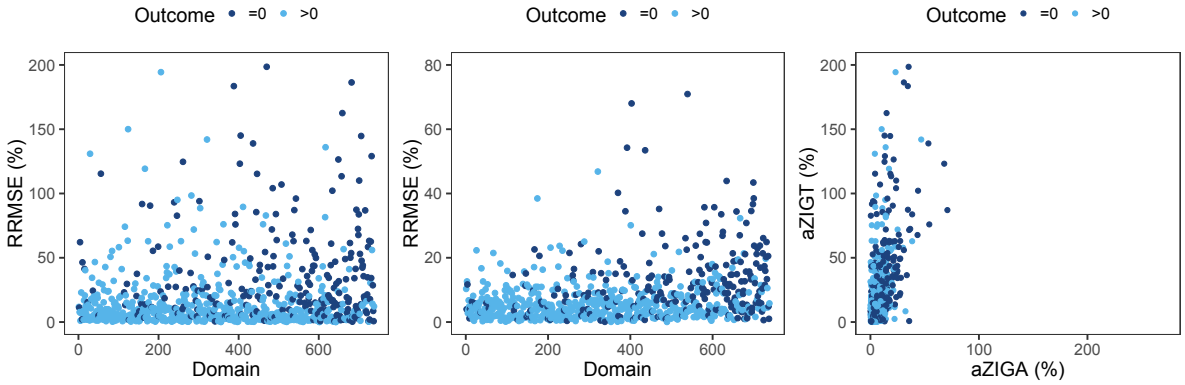


Figure 2.11: Spanish forest fire plug-in retroconditions for 2015 over the study weeks (from July to October) and aggregated by the three pre-defined regions of the Iberian Peninsula.

We also provide estimates of the MSE of the plug-in predictor for the 2015 retroconditions. For this purpose, we apply the resampling method described in Section 2.4.2 and consider both models aZIGT and aZIGA. Figure 2.12 shows the RRMSE estimates for domains of model aZIGT (left), model aZIGA (center) and one model against the other (right), all of them in %. To analyse what happens, a distinction is made between zero and non-zero observations. The model aZIGT performs worse, which is in line with the problems of fitting and handling large outcomes already mentioned. For totals, the main quartiles are $q_0 = 0.036$, $q_{0.25} = 3.752$, $q_{0.5} = 10.085$, $q_{0.75} = 25.342$ and $q_1 = 198.558$. In contrast, the model aZIGA is much better, with considerably lower RRMSE estimates. For averages, the main quartiles are $q_0 = 0.039$, $q_{0.25} = 2.564$, $q_{0.5} = 5.386$, $q_{0.75} = 10.187$ and $q_1 = 70.940$. As can be inferred from Figure 2.12 (right), domains with higher RRMSE estimates match for both models, although they are higher for totals. These are either domains with no records or very small weekly and provincial values, both in number and magnitude. This is mainly explained by the relative nature of the RRMSE. In addition, it is difficult to fit a model when the distribution of the data across provinces and weeks is so dissimilar. However, our results are reassuring: it is convenient to predict better in those domains with larger average forest fires, as the severity of the environmental problem is of much more concern.

Last but not least, we have proposed a measure of risk. The aim is to have solid evidence for firefighting, bearing in mind that the important issue is to know whether a fire is particularly dangerous or not, without seeking to predict exactly how many hectares it will burn. As a starting point, bootstrap estimates of the quantiles of the predicted average burned areas, $\widehat{q}_{ijk,\alpha}$, are calculated for $\alpha \in \{0.05, 0.15, 0.2, 0.3\}$, according to Section 2.4.2. We set 50 Ha as a threshold to introduce the risk scale: *medium-low* if $\widehat{q}_{ijk,0.30} \leqslant 50$ ($< 70\%$), *moderate* if $\widehat{q}_{ijk,0.20} \leqslant 50 < \widehat{q}_{ijk,0.30}$ (70-80%), *high* if $\widehat{q}_{ijk,0.15} \leqslant 50 < \widehat{q}_{ijk,0.20}$ (80-85%), *very high* if $\widehat{q}_{ijk,0.05} \leqslant 50 < \widehat{q}_{ijk,0.15}$ (80-85%), and *extreme* if $\widehat{q}_{ijk,0.05} > 50$ ($> 95\%$). The risk conditions establish a hierarchy over the Spanish provinces, providing weekly horizon prospects.

Figure 2.13 shows the risk maps for weeks 28, 29 and 30 of 2015. For the study period, the presence of constant provincial patterns over time is quite evident, highlighting the high risk associated with the provinces of Galicia and Extremadura, as well as some provinces of



(a) Model aZIGT (b) Model aZIGA (c) Compared to each other

Figure 2.12: RRMSE estimates of the 2015 plug-in retroconditions by domains. Domains are defined as crossings between years, weeks and provinces.

Castilla and León. In these domains, the detection of a forest fire is an alarm signal because, with high probability, it can grow considerably in size and, in particular, burn more than 50 Ha. On the opposite side are the provinces of the Mediterranean Coast and the whole of Andalucía. Thus, our results are consistent with the facts: north-western Spain is the region most affected by the incidence and intensity of forest fires because of its large forest mass.

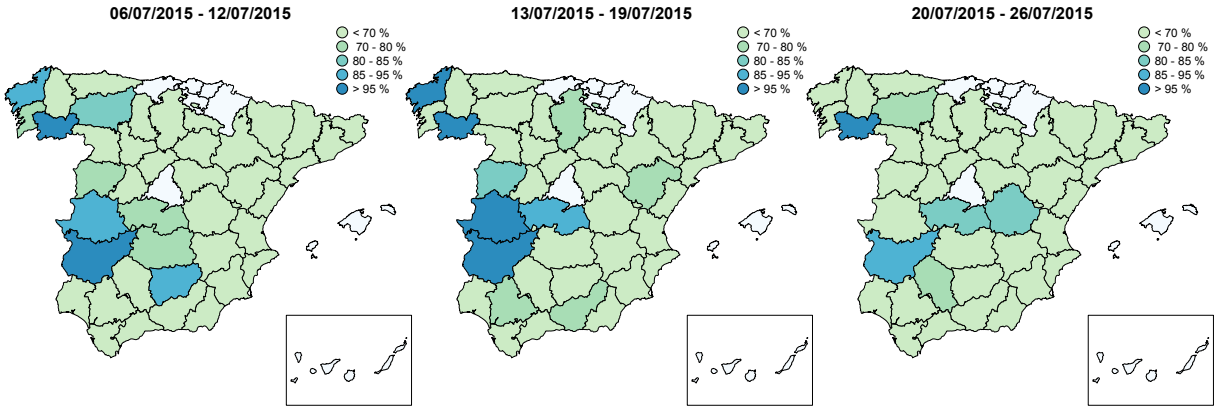


Figure 2.13: Risk maps for weeks 28, 29 and 30 of year 2015. The risk scale is calculated according to the bootstrap quantile estimation method of Section 2.4.2.

Several conclusions are drawn. The modelling of burned forest area has demonstrated the suitability of the GA distribution to model continuous and asymmetric outcomes, and the BE distribution to tackle the excess of zeros. Added to that, the application to real data provides solid and convincing results for the following reasons: (1) forest fire data are strongly affected by uncertainty and opportunism; (2) short-term (weekly) fire modelling is key to managing resources in potentially vulnerable locations and planning future firefighting practices; (3) *easy-to-predict* meteorological data have been used, which simplifies forecasting tasks; (4) the

error inherent in the modelling of large events is conditioned by simplified representations of the underlying physical processes.

2.4.5 R codes

As for the `R` codes, the GitHub repository https://github.com/mbugallo/aZIG11Fires (accessed on: November 4, 2024) contains our dataset and computer code, as well as a detailed description of its contents. It includes a README file that provides basic instructions for the correct execution of the available software.

Chapter 3

Three-fold Fay-Herriot model and prediction of segregation indexes

This chapter is self-contained and follows Bugallo et al. (2024d) as a reference point. It describes a new statistical methodology for the small area prediction of dissimilarity indexes of occupational segregation by sex in administrative areas and time periods. In this respect, there is a large amount of literature on the measurement of segregation, with several indexes in use, all of which have different properties. However, one of the most popular measures of segregation is the Duncan Segregation Index (DSI) (Duncan and Duncan, 1955), which allows for the assessment of differences between categories in the calculation of various socio-economic indicators. The DSI is a measure of segregation that is applied to individuals differentiated by a dichotomous classification variable in groups defined by sex, race, origin, religion or culture, to name just but a few. Locations should be interpreted in a broad sense. Examples of locations are residential areas, educational levels or occupational sectors.

The current research examines the use of the DSI to measure occupational segregation by sex, where the group variable is sex and the location variable is the occupational sector. This is done by comparing the percentage of men and women employed in each occupational sector and providing a numerical value which is lower the closer the occupational distribution is to equality. If all sectors have the same proportion of employed men and women, the DSI is zero. Otherwise, the DSI is one and segregation reaches its maximum.

The estimation of segregation indexes has been widely studied but, to our knowledge, little attention has been paid to the sample sizes used in the inferential processes. Recent contributions include Salardi (2016), who examines the evolution of racial and sex segregation in Brazilian labour markets, and Das and Kotikula (2019), who analyse the causes of gender-based occupational segregation. As a general feature, the studies cited above assume that the available information is fully reliable. In practice, data may come from surveys and are therefore subject to sampling errors. On condition the data come from administrative registers or surveys with large sample sizes, the calculation of the DSI is straightforward. When sample sizes are small, direct estimators may be unreliable and the problem deserves further methodological research. The main advantage of the three-fold Fay-Herriot (FH3) model over the existing literature is that it is an area-level model with hierarchical nesting, which suits

the nature of our data. Nested error regression (NER) models may also be appropriate, but the lack of census data and administrative registers would limit their predictive power to that of ANOVA-type models. It is highly doubtful that these models would have performed better in predicting DSIs over small areas.

This chapter is structured as follows. Section 3.1 introduces the data, the dissimilarity indexes and the small area problem. Section 3.2 describes the FH3 model, fitted to the proportion of employed men by province, occupation and time period, and then derives EBPs and plug-in predictors of the DSI. A parametric bootstrap algorithm is implemented to estimate the MSE by following Marcis et al. (2023). Section 3.3 includes some simulation experiments to investigate the performance of the DSI predictors and MSE estimators. The simulation scenario is based on the case study. Section 3.4 deals with the application to real data. Data from the 2020.4-2021.4 Spanish Labour Force Survey (SLFS) are used to illustrate the performance of the new statistical methodology and to shed some light on the current state of sex occupational segregation by province in Spain.

3.1 Dissimilarity indexes and 2020.4-2021.4 SLFS data

The application to real data aims to estimate sex occupational segregation by Spanish provinces ($D = 52$) and time period. Data are from $T = 5$ Spanish Labour Force Surveys (SLFS), starting in the last quarter of 2020 (SLFS2020.4) and up to the last quarter of 2021 (SLFS2021.4). The population of interest is made up of people aged 16 and over ($age \geqslant 16$), with permanent residence in Spain. Respondents under 16 years of age are not considered because they are not above the minimum age for working in Spain. The occupational sector (OC) variable has been derived from the 2011 Spanish National Classification of Occupations (CNO2011), statistical classification published online by the Spanish National Statistical Office (INE) (https://www.ine.es/en/daco/daco42/clasificaciones/nota_epa_cno11_en.pdf; accessed on: November 4, 2024). Three categories have, however, been aggregated due to the smallness of the sample sizes and because they are roughly similar in description. The final encoding of the occupational sector variable, with $R = 7$ mutually exclusive categories, is described in Table 3.1. The set of categories covers a wide range of occupations and provides an accurate picture about the respondents in terms of their main occupation.

In order to calculte the DSI by province and time period, some mathematical definitions are given below. Let U_{drt} be a subset (estimation domain) of the population, relative to time period t and conformed by N_{drt} employed people aged 16 or over, resident in province d and working in sector r. Let y_{drt1j} be a dichotomic variable such that $y_{drt1j} = 1$ if the individual j of U_{drt} is male, and $y_{drt1j} = 0$, otherwise. Let $y_{drt2j} = 1 - y_{drt1j}$ be the analogous variable for females. The population means of these variables are

$$\overline{Y}_{drt1} = \frac{1}{N_{drt}} \sum_{j=1}^{N_{drt}} y_{drt1j}, \quad \overline{Y}_{drt2} = \frac{1}{N_{drt}} \sum_{j=1}^{N_{drt}} y_{drt2j}, \quad d = 1, \ldots, D, \ r = 1, \ldots, R, \ t = 1, \ldots, T. \tag{3.1}$$

Code	Description
OC1	Directors and managers. Senior public and private figures
OC2	Scientists and intellectual technicians and professionals
OC3*	(i) Military occupations. (ii) Technicians and support staff
OC4	Accounting, administrative and other office employees
OC5	Catering, protection and commercial staff
OC6*	(i) Unskilled workers. (ii) Primary sector workers
OC7*	(i) Plant and machinery operators. (ii) Craftsmen and skilled workers in the manufacturing and construction industries.

Table 3.1: Encoding of the occupational sector (OC) variable. The * means that the categories have been aggregated because they are roughly similar in description.

The Duncan Segregation Index (DSI) of province d at time period t is

$$S_{d.t} = \frac{1}{2}\sum_{r=1}^{R} S_{drt} \text{ where } S_{drt} = \left| \frac{N_{drt}\overline{Y}_{drt1}}{\sum_{i=1}^{R} N_{dit}\overline{Y}_{dit1}} - \frac{N_{drt}\overline{Y}_{drt2}}{\sum_{i=1}^{R} N_{dit}\overline{Y}_{dit2}} \right|. \tag{3.2}$$

In our research, the DSI measures how evenly (or unevenly) the population of both sexes is distributed in each occupational sector (Bugallo et al., 2024d). Segregation is measured as the degree to which the spatial distribution of the female group deviates from that of the male one. As long as men and women are distributed in equal proportions in the different occupational sectors, there is no segregation. Therefore, the DSI has a straightforward interpretation: it corresponds to the proportion of women (or men) who would have to move to another occupational sector to balance the distribution. Movements would have to occur from occupations in which the group is overrepresented to occupations in which it is underrepresented.

In practice, the theoretical DSI values defined in (3.2) should be estimated by using SLFS data. However, our estimation domains are not planned in the SLFS so we first investigate whether the sample sizes are large enough to provide accurate direct estimates of the dissimilarities S_{drt}'s. For this purpose, we introduce some additional notation below. Let n_{drt} and $\widehat{N}_{drt}^{dir}$ be the sample size and the estimated population size (sum of the sampling weights) of U_{drt}. Let $n = \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{t=1}^{T} n_{drt}$ be the global sample size. The estimated sampling fractions (in %), are defined as relative sample sizes as

$$f_{drt} = 100\frac{n_{drt}}{\widehat{N}_{drt}^{dir}}, \tag{3.3}$$

and are not uniformly distributed in U_{drt}. The latter is shown in Table 3.2, which presents the deciles of the sample sizes (SS) and the estimated sampling fractions (SF).

It can be observed in Table 3.2 that 20% (50%) of the U_{drt}'s have samples sizes smaller than 56 (121) and that the average sample size, equal to 149, is between $q_{0.6} = 143$ and $q_{0.7} = 175$, indicating that the sample size distribution is positively skewed. Furthermore, sampling fractions allow us to know the percentage of individuals of the subsets U_{drt} who

	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
SS	6	34	56	82	104	121	143	175	218	297	1013
SF	0.091	0.202	0.249	0.290	0.350	0.407	0.469	0.538	0.652	0.811	1.779

Table 3.2: Deciles of the sample sizes (SS) and the estimated sampling fractions (SF), in %, for the 2020.4-2021.4 SLFS data.

actually belong to the sample. As they are all lower than 1.779, the representativeness of the samples in the crosses is quite small. Consequently, this is a SAE problem and direct estimators, such as the Hájek estimator, are not accurate enough. The inference problem requires the incorporation of more sophisticated prediction methods.

Table 3.3 shows the total and the proportion of men and women in the subset of employees by main occupation for the SLFS2021.4 data. We conclude that OC7 and, to a lesser extent, OC1 are of particular interest for the analysis of sex occupational segregation.

Occupation sector		OC1	OC2	OC3	OC4	OC5	OC6	OC7
Men	Total	11,023	31,816	29,101	12,720	31,845	27,798	63,984
	Proportion	0.698	0.395	0.631	0.317	0.380	0.508	0.906
Women	Total	5,191	44,518	17,947	28,116	48,398	29,021	6,694
	Proportion	0.302	0.605	0.369	0.683	0.620	0.492	0.094

Table 3.3: Employed men and women by occupation sector in the SLFS of 2021.4.

The Hájek estimators of $\overline{Y}_{drt1}$ and $\overline{Y}_{drt2}$ are direct estimators that are calculated by using only data of the SLFS sample s_{drt} of the subset U_{drt} and the sampling weights w_{drt}'s. They are therefore ratios between two quantities, $\widehat{Y}^{dir}_{drt\kappa}$ and $\widehat{N}^{dir}_{drt}$, given by

$$\widehat{\overline{Y}}^{dir}_{drt\kappa} = \frac{\widehat{Y}^{dir}_{drt\kappa}}{\widehat{N}^{dir}_{drt}} = \frac{\sum_{j\in s_{drt}} w_{drtj} y_{drt\kappa j}}{\sum_{j\in s_{drt}} w_{drtj}}, \quad \kappa = 1, 2. \tag{3.4}$$

To overcome the lack of precision of the Hájek estimator, we incorporate auxiliary and hierarchical information, and derive model-based predictors, which are the ones that drive our research. The selected auxiliary variables are the Hájek estimates of the proportion of individuals in U_{drt} that belong to the categories of the following factors:

Age group, with 3 categories: between 16 and 30 years (*age3-1*), between 30 and 50 years (*age3-2*) and over 50 years (*age3-3*).

Citizenship, with 2 categories: Spanish (*cit1*) and not Spanish (*cit2*).

Education, with 4 categories: primary or less (*edu1*), basic secondary education (*edu2*), advanced secondary education (*edu3*) and higher education, such as university (*edu4*).

Working hours, with 2 categories: full-time (*work1*) and part-time work (*work2*).

Professional status, with 5 categories: self-employed (*st1*), cooperative or family business (*st2*), public (*st3*) and private (*st4*) sector salaried employee and others (*st5*).

The set of categories of each factor is exhaustive, so the estimated proportions sum one. Based on their socio-economic meaning, we have limited to 11 auxiliary variables defined at the level of the U_{drt} subsets. We have removed *age3-2, cit2, edu2, work2* and *st5*. First of all, *cit1* and *work1* are complementary to *cit2* and *work2*, respectively, so the selection of the former or the latter is of little interest. As for *age3-2*, it represents the intermediate category, so we consider it more informative to include the age variables that account for the two edge groups, which to some extent also applies to *edu2*. Finally, we dropped *st5*, defined as "*others*", for being the most ambiguous variable to account for professional status.

For the sake of accuracy, we jointly use data from the last five SLFSs to estimate the covariates for each quarter and the population sizes N_{drt} used to calculate the DSI values in (3.2). Therefore, the effects of the variances of the covariate means and population sizes in the properties of the prediction procedure are considered negligible. This allows for an approximate 5-fold increase in available data and reduces temporal variability. As an example, the vector of covariates for $t = 1$ (SLFS2020.4) is estimated using the SLFS data from 2019.4 to 2020.4, both surveys included. In simulation experiments in Section 4.5, we empirically verify that this does not lead to underestimating the final variability. In addition, the elevation factors are the inverses of the inclusion probabilities, which are deterministic, after a calibration process whose randomness is minimal. As a result, the population sizes estimated as sums of elevation factors have negligible variability.

Table 3.4 compares the quartiles of the variances of the Hájek estimates of the selected auxiliary variables with those of the response variable. All area-level variables being proportions, it is safe to say that the variability of the covariates is significantly lower than that of $\widehat{\overline{Y}}^{dir}_{drt1}$ and close to zero. To provide further evidence to the previous point, the loss of precision of model-based predictors when using area-level auxiliary variables obtained with sampling errors will be studied. The results are reported in Section 3.3 of the simulation experiments, where the framework is based on the application to real data performed in Section 3.4.

	age3-1	*age3-3*	*cit1*	*edu1*	*edu3*	*edu4*	*work1*	*st1*	*st2*	*st3*	*st4*	$\widehat{\overline{Y}}^{dir}_{drt1}$
$q_{0.25}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
$q_{0.5}$	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.002
$q_{0.75}$	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.001	0.000	0.001	0.002	0.004

Table 3.4: Quartiles of the variances of the Hájek estimates of the selected auxiliary variables and the response variable. Data from the 2020.4-2021.4 SLFSs.

To take advantage of the area-level auxiliary data to refine the estimation of the proportions of men and women employed in each occupational sector, and to obtain DSI predictions by province and time period, Section 3.2 details the FH3 model-based statistical methodology. No specific model has been used for the proportions because priority has been given to obtaining DSI predictors from a three-fold nested model that allows the population to be hierarchised in provinces, occupational sectors and time periods.

3.2 Three-fold Fay-Herriot statistical methodology

The three-fold Fay-Herriot (FH3) model (Marcis et al., 2023) is defined in two steps, with the simplified notation $y_{drt} = \widehat{\overline{Y}}^{dir}_{drt1}$ and $\mu_{drt} = \overline{Y}_{drt1}$. The first step starts from the sampling model, indicating that y_{drt} is an unbiased estimator of μ_{drt}, i.e.

$$y_{drt} = \mu_{drt} + e_{drt},\ e_{drt} \sim N(0, \sigma^2_{drt}),\ \sigma^2_{drt} > 0,\ \ d = 1, \ldots, D,\ r = 1, \ldots, R,\ t = 1, \ldots, T, \tag{3.5}$$

where the error variances σ^2_{drt}'s are assumed to be known.

The selection of σ^2_{drt} is worthy of comment. In practice, we use the generalized variance function (GVF) method (see Chapter 5 of Wolter (1985)) to calculate σ^2_{drt}. For this purpose, a regression model is fitted to the direct estimates of the design-based variance of y_{drt}, $\widehat{\sigma}^{dir,2}_{drt}$, obtained in advance from the unit-level survey data. See e.g. Remark 2.3 in Morales et al. (2021). Following Section 16.4 in Morales et al. (2021), we define the log-linear model

$$\log(\widehat{\sigma}^{dir,2}_{drt}) = b_0 + b_1 y_{drt} + b_2 n_{drt} + \varepsilon_{drt}, \tag{3.6}$$

where the ε_{drt}'s are i.i.d. $N(0, \sigma^2_A)$ and $\sigma^2_A > 0$. Intuitively, b_1 is expected to be positive and b_2 negative. The final σ^2_{drt} equals the variance values predicted by the GVF model (3.6), i.e.

$$\sigma^2_{drt} = \exp(\widehat{\sigma}^2_A/2) \cdot \exp\left(\widehat{b}_0 + \widehat{b}_1 y_{drt} + \widehat{b}_2 n_{drt}\right), \tag{3.7}$$

where the factor $\exp(\widehat{\sigma}^2_A/2)$ is the usual bias correction term in a log-linear regression analysis to prevent underestimation. This allows for the smoothing of the direct estimates $\widehat{\sigma}^{dir,2}_{drt}$.

In a second step, a linking model is constructed assuming a hierarchical linear relationship between μ_{drt} and a row vector $\boldsymbol{x}_{drt}$ of p auxiliary variables, i.e.

$$\mu_{drt} = \boldsymbol{x}_{drt}\boldsymbol{\beta} + u_{1,d} + u_{2,dr} + u_{3,drt},\ \ d = 1, \ldots, D,\ r = 1, \ldots, R,\ t = 1, \ldots, T, \tag{3.8}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ row vector of model parameters, $u_{1,d} \sim N(0, \sigma^2_1)$, $u_{2,dr} \sim N(0, \sigma^2_2)$, $u_{3,drt} \sim N(0, \sigma^2_3)$ and $\sigma^2_1,\ \sigma^2_2,\ \sigma^2_3 > 0$ are the variance parameters. We further assume independence between errors and random effects.

The FH3 model is a linear mixed model that can be expressed in the single form

$$y_{drt} = \boldsymbol{x}_{drt}\boldsymbol{\beta} + u_{1,d} + u_{2,dr} + u_{3,drt} + e_{drt}, \quad d = 1, \ldots, D,\ r = 1, \ldots, R,\ t = 1, \ldots, T. \tag{3.9}$$

For $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\sigma^2_1, \sigma^2_2, \sigma^2_3)$, the REML log-likelihood function is

$$l_{reml}(\boldsymbol{\theta}) = -\frac{DRT - p}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{X}| - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}, \tag{3.10}$$

where the column and diagonal operators define the vectors and matrices

$$\begin{aligned}
\boldsymbol{X} &= \operatorname*{col}_{1\leqslant d\leqslant D}\Big(\operatorname*{col}_{1\leqslant r\leqslant R}\Big(\operatorname*{col}_{1\leqslant t\leqslant T}(\boldsymbol{x}_{drt})\Big)\Big), \quad \boldsymbol{V}_e = \operatorname*{diag}_{1\leqslant d\leqslant D}\Big(\operatorname*{diag}_{1\leqslant r\leqslant R}\Big(\operatorname*{diag}_{1\leqslant r\leqslant R}(\sigma^2_{drt})\Big)\Big),\\
\boldsymbol{V} &= \sigma^2_1 \operatorname*{diag}_{1\leqslant d\leqslant D}(\mathbf{1}_{RT}\mathbf{1}'_{RT}) + \sigma^2_2 \operatorname*{diag}_{1\leqslant d\leqslant D}\Big(\operatorname*{diag}_{1\leqslant r\leqslant R}(\mathbf{1}_T\mathbf{1}'_T)\Big) + \sigma^2_3\boldsymbol{I}_{DRT} + \boldsymbol{V}_e,\\
\boldsymbol{y} &= \operatorname*{col}_{1\leqslant d\leqslant D}\Big(\operatorname*{col}_{1\leqslant r\leqslant R}\Big(\operatorname*{col}_{1\leqslant t\leqslant T}(y_{drt})\Big)\Big), \quad \boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1},
\end{aligned}$$

and $\mathbf{1}_m$ and $\boldsymbol{I}_m$ denote the $m \times 1$ vector of ones and the $m \times m$ identity matrix, respectively. The REML estimators of the variance components, $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$, are obtained by maximizing $l_{reml}(\boldsymbol{\theta})$ in (3.10). We apply the Fisher-Scoring algorithm with updating equation

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{F}^{-1}(\boldsymbol{\theta}^{(k)})\boldsymbol{S}(\boldsymbol{\theta}^{(k)}), \tag{3.11}$$

where $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{\theta}) = (S_1, S_2, S_3)'$ is the score vector and $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{\theta}) = (F_{ab})_{a,b=1,2,3}$ is the Fisher information matrix. For $a, b = 1, 2, 3$, the components of $\boldsymbol{S}$ and $\boldsymbol{F}$ are

$$S_a = \frac{\partial l_{reml}}{\partial \theta_a} = -\frac{1}{2}\text{tr}(\boldsymbol{P}\boldsymbol{V}_a) + \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{V}_a\boldsymbol{P}\boldsymbol{y}, \quad F_{ab} = \frac{1}{2}\text{tr}(\boldsymbol{P}\boldsymbol{V}_a\boldsymbol{P}\boldsymbol{V}_b),$$

where

$$\boldsymbol{V}_1 = \frac{\partial \boldsymbol{V}}{\partial \theta_1} = \underset{1\leqslant d\leqslant D}{\text{diag}}\,(\mathbf{1}_{RT}\mathbf{1}'_{RT}), \;\; \boldsymbol{V}_2 = \frac{\partial \boldsymbol{V}}{\partial \theta_2} = \underset{1\leqslant d\leqslant D}{\text{diag}}\Big(\underset{1\leqslant r\leqslant R}{\text{diag}}\,(\mathbf{1}_T\mathbf{1}'_T)\Big), \;\; \boldsymbol{V}_3 = \frac{\partial \boldsymbol{V}}{\partial \theta_3} = \boldsymbol{I}_{DRT}.$$

To estimate $\boldsymbol{\beta}$ and to predict $\boldsymbol{u} = (\boldsymbol{u}'_1, \boldsymbol{u}'_2, \boldsymbol{u}'_3)'$, where

$$\boldsymbol{u}_1 = \underset{1\leqslant d\leqslant D}{\text{col}}(u_{1,d}), \quad \boldsymbol{u}_2 = \underset{1\leqslant d\leqslant D}{\text{col}}\Big(\underset{1\leqslant r\leqslant R}{\text{col}}(u_{2,dr})\Big), \quad \boldsymbol{u}_3 = \underset{1\leqslant d\leqslant D}{\text{col}}\Big(\underset{1\leqslant r\leqslant R}{\text{col}}\Big(\underset{1\leqslant t\leqslant T}{\text{col}}(u_{3,drt})\Big)\Big),$$

we use the REML estimator of $\boldsymbol{\beta}$ and the REML-EBLUP of $\boldsymbol{u}$, i.e.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{V}}_u\boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right), \tag{3.12}$$

where $\widehat{\boldsymbol{V}}$ is obtained by plugging, $\widehat{\boldsymbol{V}}_u = \text{diag}(\widehat{\sigma}_1^2\boldsymbol{I}_D, \widehat{\sigma}_2^2\boldsymbol{I}_{DR}, \widehat{\sigma}_3^2\boldsymbol{I}_{DRT})$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$, and

$$\boldsymbol{Z}_1 = \underset{1\leqslant d\leqslant D}{\text{diag}}\left(\mathbf{1}_{RT}\right), \quad \boldsymbol{Z}_2 = \underset{1\leqslant d\leqslant D}{\text{diag}}\Big(\underset{1\leqslant r\leqslant R}{\text{diag}}\,(\mathbf{1}_T)\Big), \quad \boldsymbol{Z}_3 = \boldsymbol{I}_{DRT}.$$

The EBLUP of μ_{drt} is $\widehat{\mu}_{drt} = \boldsymbol{x}_{drt}\widehat{\boldsymbol{\beta}} + \widehat{u}_{1,d} + \widehat{u}_{2,dr} + \widehat{u}_{3,drt}$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$ are given in (3.12). Consequently, each $\widehat{\mu}_{drt}$ contains area-level auxiliary information that will reduce the variance of the Hájek estimates $\widehat{\overline{Y}}^{dir}_{drt1}$ in (3.4) without needing to increase the sample sizes.

3.2.1 Small area prediction of Duncan Segregation Indexes

Below we derive several model-based predictors for the DSI indicators defined in (3.2), assuming that y_{drt} follows the FH3 model (3.5)-(3.8). Let $d = 1, \ldots, D, r = 1, \ldots, R$, $t = 1, \ldots, T$. First, let us define $\boldsymbol{u}_{drt} = (u_{1,d}, u_{2,dr}, u_{3,drt})'$, so that

$$\boldsymbol{u}_{drt} \sim N_K(\mathbf{0}, \boldsymbol{V}_{u,drt}), \quad \boldsymbol{V}_{u,drt} = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2), \quad \boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2,)$$

and $K = 1 + R + RT$. We consider the domain target parameters

$$S_{drt} = \left| \frac{N_{drt}\mu_{drt}}{\sum_{i=1}^R N_{dit}\mu_{dit}} - \frac{N_{drt}(1-\mu_{drt})}{\sum_{i=1}^R N_{dit}(1-\mu_{dit})} \right|. \tag{3.13}$$

The plug-in predictors of S_{drt} and $S_{d.t}$ are

$$\widehat{S}^{in}_{d.t} = \frac{1}{2}\sum_{r=1}^{R} \widehat{S}^{in}_{drt}, \quad \widehat{S}^{in}_{drt} = \left| \frac{N_{drt}\widehat{\mu}_{drt}}{\sum_{i=1}^{R} N_{dit}\widehat{\mu}_{dit}} - \frac{N_{drt}(1-\widehat{\mu}_{drt})}{\sum_{i=1}^{R} N_{dit}(1-\widehat{\mu}_{dit})} \right|; \tag{3.14}$$

and the marginal predictor (MP) of S_{drt} is

$$\widehat{S}^{mp}_{drt} = E[S_{drt}|y_{drt}] = \frac{\int_{\mathbb{R}^3} S_{drt}(\boldsymbol{u}_{drt},\boldsymbol{\beta}) f(y_{drt}|\boldsymbol{u}_{drt}) f(\boldsymbol{u}_{drt})\, d\boldsymbol{u}_{drt}}{\int_{\mathbb{R}^3} f(y_{drt}|\boldsymbol{u}_{drt}) f(\boldsymbol{u}_{drt})\, d\boldsymbol{u}_{drt}} = \frac{A_{drt}(y_{drt},\boldsymbol{\beta},\boldsymbol{\theta})}{B_d(y_{drt},\boldsymbol{\beta},\boldsymbol{\theta})},$$

where

$$\begin{aligned} A_{drt}(y_{drt},\boldsymbol{\beta},\boldsymbol{\theta}) &= \int_{\mathbb{R}^3} S_{drt}(\boldsymbol{u}_{drt},\boldsymbol{\beta}) \exp\Big\{ -\frac{1}{2\sigma^2_{drt}} e^2_{drt} \Big\} f(\boldsymbol{u}_{drt}) d\boldsymbol{u}_{drt}, \\ B_d(y_{drt},\boldsymbol{\beta},\boldsymbol{\theta}) &= \int_{\mathbb{R}^3} \exp\Big\{ -\frac{1}{2\sigma^2_{drt}} e^2_{drt} \Big\} f(\boldsymbol{u}_{drt}) d\boldsymbol{u}_{drt} \end{aligned}$$

and $e_{drt} = y_{drt} - \boldsymbol{\mu}_{drt} = y_{drt} - \boldsymbol{x}_{drt}\boldsymbol{\beta} - u_{1,d} - u_{2,dr} - u_{3,drt}$.

The empirical marginal predictor (EMP) of S_{drt} is $\widehat{S}^{emp}_{drt} = A_{drt}(y_{drt},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})/B_d(y_{drt},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})$. Therefore, the following algorithm gives a Monte Carlo approximation of $\widehat{S}^{emp}_{drt}$.

1. Fit the model and obtain the REML estimates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$.
2. For $\ell = 1, \ldots, L$, do
   (a) Draw $u^{(\ell)}_{1,d} \sim N(0,\widehat{\sigma}^2_1)$, $u^{(\ell)}_{2,dr} \sim N(0,\widehat{\sigma}^2_2)$, $u^{(\ell)}_{3,drt} \sim N(0,\widehat{\sigma}^2_3)$, $\boldsymbol{u}^{(\ell)}_{drt} = (u^{(\ell)}_{1,d}, u^{(\ell)\prime}_{2,dr}, u^{(\ell)\prime}_{3,drt})'$ and set $\boldsymbol{u}^{(L+\ell)}_{drt} = -\boldsymbol{u}^{(\ell)}_{drt}$.
   (b) Calculate $\widehat{S}^{emp}_{drt} = \widehat{A}_{drt}/\widehat{B}_d$, where

$$\widehat{A}_{drt} = \frac{1}{2L}\sum_{\ell=1}^{2L} S_{drt}(\boldsymbol{u}^{(\ell)}_{drt},\widehat{\boldsymbol{\beta}}) \exp\Big\{ -\frac{e^2_{drt}}{2\sigma^2_{drt}} \Big\}, \quad \widehat{B}_d = \frac{1}{2L}\sum_{\ell=1}^{2L} \exp\Big\{ -\frac{e^2_{drt}}{2\sigma^2_{drt}} \Big\}$$

   and $e^{(\ell)}_{drt} = y_{drt} - \boldsymbol{x}_{drt}\widehat{\boldsymbol{\beta}} - u^{(\ell)}_{1,d} - u^{(\ell)}_{2,dr} - u^{(\ell)}_{3,drt}$.

All in all, the EMP of $S_{d.t}$ is

$$\widehat{S}^{emp}_{d.t} = \frac{1}{2}\sum_{r=1}^{R} \widehat{S}^{emp}_{drt}.$$

The BP of S_{drt}, $\widehat{S}^{bp}_{drt} = E[S_{drt}|\boldsymbol{y}]$, is also a potentially attractive alternative: theoretically it has minimum MSE within the class of unbiased predictors. However, its computation requires to approximate an integral in $\mathbb{R}^K$, with $K = 43$ in the application to real data. This is computationally intensive and is the main reason why we do not consider the EBP approach under the proposed FH3 model to be a useful alternative for predicting the Duncan Segregation Index in academia or in the production of public statistics.

3.2.2 Bootstrap inference

This section presents bootstrap-based CIs for the model parameters and estimators of the MSE of $\widehat{\mu}_{drt}$ and $\widehat{S}_{d.t} \in \{\widehat{S}^{in}_{d.t}, \widehat{S}^{emp}_{d.t}\}$, $d = 1, \ldots, D, r = 1, \ldots, R$, $t = 1, \ldots, T$. Let β_{ℓ_1} be a component of $\boldsymbol{\beta}$, $\ell_1 = 1, \ldots, p$, θ_{ℓ_2} a component of $\boldsymbol{\theta}$, $\ell_2 = 1, 2, 3$, and $\alpha \in (0, 1)$. Under the FH3 model, we adapt the parametric bootstrap procedure proposed by Marcis et al. (2023) to calculate a $(1-\alpha)\%$ percentile bootstrap CI for β_{ℓ_1}, $\ell_1 = 1, \ldots, p$, or θ_{ℓ_2}, $\ell_2 = 1, 2, 3$, and estimate the MSE of $\widehat{\mu}_{drt}$ and $\widehat{S}_{d.t} \in \{\widehat{S}^{in}_{d.t}, \widehat{S}^{emp}_{d.t}\}$.

The steps of our algorithm are described below.

1. Fit the FH3 model to the data $(y_{drt}, \boldsymbol{x}_{drt})$ and obtain the REML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

2. Repeat B times $(b = 1, \ldots, B)$:

   (a) For $d = 1, \ldots, D$, generate $u^{*(b)}_{1,d} \sim N(0, \widehat{\sigma}^2_1)$. Construct the vector $\boldsymbol{u}^{*(b)}_1 = \underset{1 \leqslant d \leqslant D}{\text{col}}(u^{*(b)}_{1,d})$.

   (b) For $d = 1, \ldots, D$, $r = 1, \ldots, R$, generate $u^{*(b)}_{2,dr} \sim N(0, \widehat{\sigma}^2_2)$. Construct the vector $\boldsymbol{u}^{*(b)}_2 = \underset{1 \leqslant d \leqslant D}{\text{col}}(\underset{1 \leqslant r \leqslant R}{\text{col}}(u^{*(b)}_{2,dr}))$.

   (c) For $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$, generate $u^{*(b)}_{3,drt} \sim N(0, \widehat{\sigma}^2_3)$. Construct the vector $\boldsymbol{u}^{*(b)}_3 = \underset{1 \leqslant d \leqslant D}{\text{col}}(\underset{1 \leqslant r \leqslant R}{\text{col}}(\underset{1 \leqslant t \leqslant T}{\text{col}}(u^{*(b)}_{3,drt})))$.

   (d) For $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$, generate $e^{*(b)}_{drt} \sim N(0, \sigma^2_{drt})$. Construct the vector $\boldsymbol{e}^{*(b)} = \underset{1 \leqslant d \leqslant D}{\text{col}}(\underset{1 \leqslant r \leqslant R}{\text{col}}(\underset{1 \leqslant t \leqslant T}{\text{col}}(e^{*(b)}_{drt})))$.

   (e) Calculate the bootstrap vectors

   $$\boldsymbol{y}^{*(b)} = \boldsymbol{\mu}^{*(b)} + \boldsymbol{e}^{*(b)}, \quad \boldsymbol{\mu}^{*(b)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_1\boldsymbol{u}^{*(b)}_1 + \boldsymbol{Z}_2\boldsymbol{u}^{*(b)}_2 + \boldsymbol{Z}_3\boldsymbol{u}^{*(b)}_3.$$

   (f) For $d = 1, \ldots, D$, calculate the bootstrap quantities

   $$S^{*(b)}_{d.t} = \frac{1}{2}\sum_{r=1}^{R} S^{*(b)}_{drt}, \quad S^{*(b)}_{drt} = \left| \frac{N_{drt}\mu^{*(b)}_{drt}}{\sum_{i=1}^{R} N_{dit}\mu^{*(b)}_{dit}} - \frac{N_{drt}(1-\mu^{*(b)}_{drt})}{\sum_{i=1}^{R} N_{dit}(1-\mu^{*(b)}_{dit})} \right|.$$

   (g) Fit the FH3 model to the bootstrap vector $\boldsymbol{y}^{*(b)}$. Calculate the ML parameter estimators $\widehat{\boldsymbol{\theta}}^{*(b)}$, $\widehat{\boldsymbol{\beta}}^{*(b)}$, the EBLUP $\widehat{\boldsymbol{\mu}}^{*(b)}$, with components $\widehat{\mu}^{*(b)}_{drt}$, and the predictors $\widehat{S}^{*(b)}_{d.t}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$.

3. Sort the values $\widehat{\beta}^{*(b)}_{\ell_1}$ or $\widehat{\theta}^{*(b)}_{\ell_2}$, $\ell_1 = 1 \ldots, p$, $\ell_2 = 1, 2, 3$, $b = 1, \ldots, B$, from smallest to largest. They are $\widehat{\beta}^{*}_{\ell_1(1)} \leqslant \ldots \leqslant \widehat{\beta}^{*}_{\ell_1(B)}$ and $\widehat{\theta}^{*}_{\ell_2(1)} \leqslant \ldots \leqslant \widehat{\theta}^{*}_{\ell_2(B)}$. A $(1-\alpha)\%$ percentile bootstrap CI for β_{ℓ_1} is $\left(\widehat{\beta}^{*}_{\ell_1(\lfloor(\alpha/2)B\rfloor)}, \widehat{\beta}^{*}_{\ell_1(\lfloor(1-\alpha/2)B\rfloor)}\right)$. A $(1-\alpha)\%$ percentile bootstrap CI for θ_{ℓ_2} is $\left(\widehat{\theta}^{*}_{\ell_2(\lfloor(\alpha/2)B\rfloor)}, \widehat{\theta}^{*}_{\ell_2(\lfloor(1-\alpha/2)B\rfloor)}\right)$.

4. For $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$, we calculate the MSE estimates as

$$mse^*(\widehat{\mu}_{drt}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mu}_{drt}^{*(b)} - \mu_{drt}^{*(b)}\right)^2, \quad mse^*(\widehat{S}_{d.t}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{S}_{d.t}^{*(b)} - S_{d.t}^{*(b)}\right)^2. \quad (3.15)$$

Remark 3.2.1. The auxiliary variables of the FH3 model must be known at domain level, from censuses or administrative records, as they must be free of sampling errors to reduce the variability of the small area predictions. In practice, however, this is not the norm, leading researchers to resort to strategies that allow estimating such area-level variables with low variability. A common technique is to use data from many consecutive surveys to increase sample sizes in the direct estimation of the auxiliary information.

If the auxiliary variables have non-negligible sampling errors, the algorithm described above could lead to underestimates of the actual MSE of $\widehat{\mu}_{drt}$ and $\widehat{S}_{d.t} \in \{\widehat{S}_{d.t}^{in}, \widehat{S}_{d.t}^{emp}\}$. As a solution to this potential problem, we propose to modify Step 2 (e) so as to include the potencial non-negligible variability of $\boldsymbol{x}_{drt}$. The proposed modification assumes uncorrelation between the columns of $\boldsymbol{x}_{drt}$ and between $\boldsymbol{x}_{drt}$ and y_{drt}. Nonetheless, correlation relationships are expected to be even lower.

Let us rewrite Step 2 (e) as follows:

2. (e) For $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$, $k = 1, \ldots, p$, generate $v_{drtk}^{*(b)} \backsim N(0, \sigma_{drtk}^2)$, where σ_{drtk}^2 is the design-based variance of the k-th component of $\boldsymbol{x}_{drt} = (x_{drt1}, \ldots, x_{drtp})$ and p is the dimension of $\boldsymbol{x}_{drt}$. This must be skipped for the intercept. If we use Hájek estimates, σ_{drtk}^2 can be replaced by the direct estimate of the design-based variance of x_{drtk}. We calculate the modified bootstrap vectors

$$\boldsymbol{y}^{*(b)} = \boldsymbol{\mu}^{*(b)} + \boldsymbol{e}^{*(b)}, \quad \boldsymbol{\mu}^{*(b)} = (\boldsymbol{X} + \boldsymbol{v}^{*(b)})\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_1\boldsymbol{u}_1^{*(b)} + \boldsymbol{Z}_2\boldsymbol{u}_2^{*(b)} + \boldsymbol{Z}_3\boldsymbol{u}_3^{*(b)},$$

where $\boldsymbol{v}^{*(b)} = \underset{1\leqslant d\leqslant D}{\text{col}}\left(\underset{1\leqslant r\leqslant R}{\text{col}}\left(\underset{1\leqslant t\leqslant T}{\text{col}}(\boldsymbol{v}_{drt}^{*(b)})\right)\right)$ and $\boldsymbol{v}_{drt}^{*(b)} = (v_{drt1}^{*(b)}, \ldots, v_{drtp}^{*(b)}) \in \mathbb{R}^p$.

3.3 Simulations based on the 2020.4-2021.4 SLFS data

Based on the case study, two model-based simulation experiments were performed. The real set of area-level auxiliary variables, the variance of the direct estimator and the fitted model, described around Table 3.8, were used to simulate the target variables. At this regard, y_{drt} represents the direct estimator of the proportion of employed men in province d, occupational sector r and time period t, i.e. $y_{drt} = \widehat{\overline{Y}}_{drt1}^{dir}$. Simulation 1 investigates the performance of the Fisher-Scoring algorithm (3.11) and studies the behaviour of the DSI predictors derived in Section 3.2.1. The loss of precision of model-based predictors when using area-level auxiliary variables obtained with sampling errors will also be studied. Simulation 2 deals with the MSE estimation and provides a recommendation on the number of bootstrap replicates to be used. The behaviour of the estimators and predictors is studied under the assumption that the fitted model is the true one. For the final FH3 model, we use x_{drt1} = intercept, $x_{drt2} = cit1$, $x_{drt3} = edu1$, $x_{drt4} = edu4$, $x_{drt5} = work1$, $x_{drt6} = st1$, $x_{drt7} = st2$ and $x_{drt8} = st4$. As mentioned above, the estimates of the model parameters are shown in Table 3.8 of Section 3.4, devoted to the application to real data.

3.3.1 Simulation 1

The goal of Simulation 1 is to investigate the behaviour of the fitting algorithm and the performance of the predictors of S_{drt} and $S_{d.t}$, $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$. We run Simulation 1 with $I = 10^3$ iterations. For a model parameter $\widehat{\tau} = \widehat{\beta}_k$, $k = 1, \ldots, 8$ or $\widehat{\tau} = \sigma_l^2$, $l = 1, 2, 3$, we calculate

$$BIAS(\widehat{\tau}) = \frac{1}{I}\sum_{i=1}^{I}(\widehat{\tau}^{(i)} - \tau), \quad RMSE(\widehat{\tau}) = \left(\frac{1}{I}\sum_{i=1}^{I}(\widehat{\tau}^{(i)} - \tau)^2\right)^{1/2},$$

and for a predictor $\widehat{S}_{d.t} \in \{\widehat{S}_{d.t}^{in}, \widehat{S}_{d.t}^{emp}\}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$, we calculate

$$ABIAS = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\frac{1}{I}\sum_{i=1}^{I}(\widehat{S}_{d.t}^{(i)} - S_{d.t}^{(i)})\right|, \quad RMSE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\left(\frac{1}{I}\sum_{i=1}^{I}(\widehat{S}_{d.t}^{(i)} - S_{d.t}^{(i)})^2\right)^{1/2}.$$

The corresponding relative performance measures (in %) are

$$RBIAS(\widehat{\tau}) = 100\frac{BIAS(\widehat{\tau})}{\tau}, \quad RRMSE(\widehat{\tau}) = 100\frac{RMSE(\widehat{\tau})}{\tau},$$

$$RBIAS_{dt} = 100\frac{BIAS_{dt}}{S_{d.t}}, \; RRMSE_{dt} = 100\frac{RMSE_{dt}}{S_{d.t}}, \; S_{d.t} = \frac{1}{I}\sum_{i=1}^{I} S_{d.t}^{(i)},$$

$$ARBIAS = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|RBIAS_{dt}|, \quad RRMSE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T} RRMSE_{dt}.$$

An analysis of Table 3.5 (top) illustrates that, for the $\boldsymbol{\beta}$ coefficients, the biases are small but the root-MSEs (RMSE) are not so small, implying that the variance is the main component of the MSE. Such variability is probably attributable to the relationship between the number of estimation domains and the number of model parameters, $DRT/(8+3) = 165.45$, which is not large enough to activate the asymptotic properties of the ML estimators. For the estimators of the variances, the RBIAS is small and the RRMSE does not present notably large values either, with the worst result being the one corresponding to $\widehat{\sigma}_1^2$.

Table 3.6 (left) provides the absolute and relative performance measures for the EMPs and the plug-in predictors of the DSI values. We use $L = 500$ iterations in the integral approximation performed when calculating the EMPs. For the plug-in predictor, the average across DSI-domains of the absolute relative bias (ARBIAS) is close to 11% and the average RRMSE does not exceed 28%, which is quite satisfactory. Incidentally, we use the plug-in predictor in the application to real data in Section 3.4. In the case of the EMP, the ARBIAS is greater than 56% and the RRMSE is close to 80%. The EMP is not obtained exactly, only approximately, because the integrals that appear in its expression cannot be calculated analytically. It should be pointed out that approximations are generated by the antithetic Monte Carlo method and calculations are subject to the number of iterations, partly justifying its poor results. Moreover, good theoretical properties are attributed to the BP, not to marginal or empirical versions.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	σ_1^2	σ_2^2	σ_3^2
Estimate	-0.327	0.142	0.089	-0.304	0.889	0.205	0.620	0.135	0.012	0.002	0.001
BIAS	-0.001	0.000	0.001	-0.001	0.001	0.000	0.014	0.000	0.000	0.000	0.000
RMSE	0.044	0.018	0.038	0.019	0.034	0.029	0.186	0.021	0.002	0.000	0.000
RBIAS	-0.387	-0.305	1.500	-0.209	0.157	-0.129	2.230	-0.166	-0.008	0.165	-0.024
RRMSE	13.346	12.681	42.924	6.216	3.836	14.228	29.992	15.423	20.792	10.701	11.360

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	σ_1^2	σ_2^2	σ_3^2
Estimate	-0.327	0.142	0.089	-0.304	0.889	0.205	0.620	0.135	0.012	0.002	0.001
BIAS	-0.002	0.001	0.004	0.002	0.000	0.001	0.001	0.001	0.000	0.000	0.000
RMSE	0.042	0.019	0.039	0.018	0.035	0.028	0.194	0.020	0.002	0.000	0.000
RBIAS	-0.450	0.594	4.609	0.537	-0.019	0.371	0.147	0.430	-0.066	-0.296	0.020
RRMSE	12.933	13.378	43.129	5.956	3.883	13.583	31.248	15.139	20.358	10.661	11.594

Table 3.5: Performance of REML estimators of $\boldsymbol{\beta}$ and θ under the assumption that the auxiliary variables are deterministic (top) and taking into account the sampling errors (bottom).

	plug-in	EMP	plug-in	EMP
ABIAS	0.051	0.324	0.050	0.342
RMSE	0.100	0.349	0.100	0.350
ARBIAS	11.186	56.484	11.730	57.284
RRMSE	27.659	79.371	27.955	78.634

Table 3.6: Performance of predictors of $S_{d.t}$ under the assumption that the auxiliary variables are deterministic (left) and taking into account the sampling errors (right).

Up to this point, we have assumed that the area-level auxiliary variables are deterministic. This assumption leads to the results in Table 3.5 (top) and Table 3.6 (left). As mentioned in Remark 3.2.1, if the auxiliary data does not come from censuses or administrative registers, but from estimates, it is potentially likely to add more variability to the small area predictions. For this reason, we have also considered in the real data simulations the scenario in which the area-level auxiliary variables have non-negligible sampling errors. For each iteration $i = 1, \ldots, I$, the new area-level auxiliary variables are generated as follows:

$$\boldsymbol{X} + \boldsymbol{v}^{*(i)}, \text{ where } \boldsymbol{v}^{*(i)} = \underset{1 \leqslant d \leqslant D}{\text{col}}\Big(\underset{1 \leqslant r \leqslant R}{\text{col}}\Big(\underset{1 \leqslant t \leqslant T}{\text{col}}(\boldsymbol{v}_{drt}^{*(i)})\Big)\Big), \ \boldsymbol{v}_{drt}^{*(i)} = (v_{drt1}^{*(i)}, \ldots, v_{drtp}^{*(b)}) \in \mathbb{R}^p, \quad (3.16)$$

$v_{drtk}^{*(i)} \backsim N(0, \sigma_{drtk}^2)$, $d = 1, \ldots, D$, $r = 1, \ldots, R$, $t = 1, \ldots, T$, $k = 1, \ldots, p$, and σ_{drtk}^2 is the design-based variance of the k-th component of $\boldsymbol{x}_{drt} = (x_{drt1}, \ldots, x_{drtp})$.

Table 3.5 (bottom) shows the results for the model parameter estimators under scenario (3.16). So as to compare the differences between the two scenarios (the deterministic scenario and scenario (3.16)), the error measures must be interpreted in absolute terms to avoid small variations caused by changes in the denominators when relativizing. Having said that, it is

concluded that there is virtually no change in the performance of the Fisher-Scoring algorithm when the random terms $v_{drtk}^{*(i)}$'s are added. This is another argument in favour of our methodology. In addition, and as can be seen in Table 3.6 (right), generating the area-level auxiliary variables according to scenario (3.16) leads to virtually no changes in the performance measures of the predictors $S_{d.t}$'s. This justifies that the variability added by estimating them with five consecutive periods of the SLFS is minimal. In light of the above, we conclude that it is not necessary to propose a measurement error model for the problem at hand, i.e. the small area prediction of DSIs by province and time period.

3.3.2 Simulation 2

Simulation 2 studies the behaviour of the parametric bootstrap estimator of the MSE of the plug-in predictor of $\widehat{S}_{d.t}^{in}$, denoted by mse_{dt}, $d = 1, \ldots, D$, $t = 1, \ldots, T$. The real MSE of $\widehat{S}_{d.t}^{in}$ is taken from Simulation 1 and denoted by MSE_{dt}. It is assumed that the area-level auxiliary variables are deterministic. As this simulation is more computationally-demanding, we run Simulation 2 with $I = 500$ iterations. Moreover, as absolute measures are more difficult to interpret, we focus our study on relative measures.

For $d = 1, \ldots, D$, $t = 1, \ldots, T$, we calculate

$$B_{dt} = \frac{1}{I}\sum_{i=1}^{I}\left(mse_{dt}^{*(i)} - MSE_{dt}\right),\ RE_{dt} = \left(\frac{1}{I}\sum_{i=1}^{I}\left(mse_{dt}^{*(i)} - MSE_{dt}\right)^2\right)^{1/2},$$

Then we define the relative performance measures (in %)

$$RB_{dt} = 100\frac{B_{dt}}{MSE_{dt}},\ RRE_{dt} = 100\frac{RE_d}{MSE_{dt}},\ d = 1, \ldots, D,\ t = 1, \ldots, T;$$

$$ARB = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|RB_{dt}|,\ RRE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}RRE_{dt}.$$

Figure 3.1 plots five boxplots of the relative biases (left), RB_{dt}, and the relative root-MSEs (right), RRE_{dt}, $d = 1, \ldots, D$, $t = 1, \ldots, T$, for $B = 50, 100, 150, 200, 300, 400$.

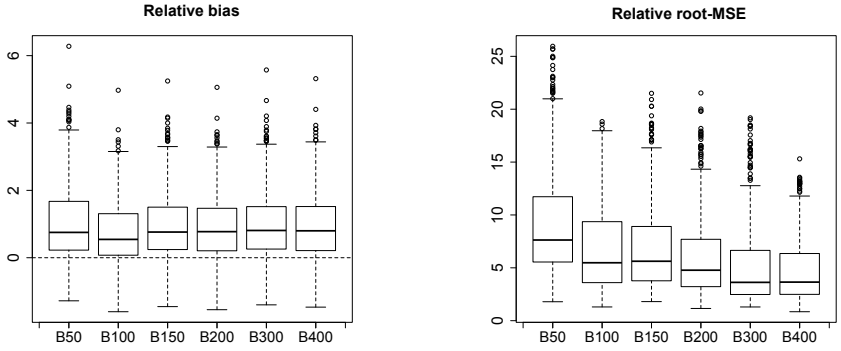


Figure 3.1: Study of the parametric bootstrap estimator of the MSE of $\widehat{S}_{d.t}^{in}$. Boxplots of RB_{dt}'s (left) and RRE_{dt}'s (right) for $B = 50, 100, 150, 200, 300, 400$.

The left boxplots show that the relative biases do not decrease as the size of B increases, showing a slight positive bias around 1.2%. The right boxplots show that the relative root-MSEs are lower than 20% and decrease as B increases, achieving good results for values greater than or equal to 300 resamples. Table 3.7 confirms it, with the average of the absolute relative biases (ARB) stabilized around 1.2% and the average of the relative root-MSEs (RRE) decreasing as B increases, but suggesting some stabilization around $B = 300$ iterations.

B	50	100	150	200	300	400
ARB	1.229	0.963	1.163	1.107	1.180	1.121
RRE	9.506	6.933	7.185	6.507	5.624	4.830

Table 3.7: Study of the parametric bootstrap estimator of the MSE of $\widehat{S}^{in}_{d.t}$. Average relative performance measures for $B = 50, 100, 150, 200, 300, 400$.

3.4 Application to the 2020.4-2021.4 SLFS data

3.4.1 Model fitting and validation

In this section we apply the FH3 model-based statistical methodology described in Section 3.2 to the SLFS 2020.4-2021.4 data. Although we fit the FH3 model to all data, we focus mainly on the results of the last quarter (SLFS2021.4) to draw conclusions. The main reason is the temporal proximity, which allows us to analyse results closer to the present day, but also to assess brevity. To fit the FH3 model to each y_{drt}, we recursively removed those auxiliary variables that are not significant at 5%. Specifically, *age3-1*, *age3-3*, *edu3* and *st3* were eliminated. For the final FH3 model, we use $x_{drt1} =$ intercept, $x_{drt2} = cit1$, $x_{drt3} = edu1$, $x_{drt4} = edu4$, $x_{drt5} = work1$, $x_{drt6} = st1$, $x_{drt7} = st2$ and $x_{drt8} = st4$. The failure to consider age groups suggests that sex segregation is persistent over time, despite the age of the worker. At this regard, the conclusions are subject to the available information and, therefore, with other territorial divisions, occupational sectors or time periods, the final set of auxiliary variables may vary. Table 3.8 presents the REML estimates of $\boldsymbol{\beta}$ and the p-values to test $H_0 : \beta_k = 0$, $k = 1, \ldots, 8$. It also includes the lower (LB) and upper (UB) bounds of the normal-asymptotic and bootstrap-percentile CIs at the 95% confidence level, being the latter discussed in Section 3.2.2.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
	Estimate	-0.327	0.142	0.089	-0.304	0.889	0.205	0.620	0.135
	p-value	0.000	0.000	0.020	0.000	0.000	0.000	0.001	0.000
Asymp	LB 95%	-0.414	0.105	0.014	-0.340	0.820	0.156	0.243	0.095
	UB 95%	-0.241	0.180	0.165	-0.269	0.957	0.260	0.998	0.175
Boot	LB 95%	-0.550	0.126	0.246	-0.406	0.704	0.067	0.818	0.052
	UB 95%	-0.117	0.413	0.426	-0.206	1.069	0.343	2.095	0.315

Table 3.8: Model parameters of the final FH3 model for the SLFS 2020.4-2021.4 data.

The effect of the auxiliary variables derived from Table 3.8 is consistent with a socio-economic interpretation. Once the rest of the variables are fixed, their sign indicates their contribution (positive or negative) to estimate the proportion of employed men by estimation domain. Regarding the model variances, we obtain $\widehat{\sigma}_1^2 = 0.012$, $\widehat{\sigma}_2^2 = 0.002$ and $\widehat{\sigma}_3^2 = 0.001$. At the 95% confidence level, the asymptotic CIs for the variances are

$$\mathrm{CI}_{\sigma_1^2}^{asymp} = (0.007, 0.016), \quad \mathrm{CI}_{\sigma_2^2}^{asymp} = (0.002, 0.003), \quad \mathrm{CI}_{\sigma_3^2}^{asymp} = (0.001, 0.001);$$

and the respetive bootstrap-percentile CIs are

$$\mathrm{CI}_{\sigma_1^2}^{boot} = (0.002, 0.004), \quad \mathrm{CI}_{\sigma_2^2}^{boot} = (0.001, 0.003), \quad \mathrm{CI}_{\sigma_3^2}^{boot} = (0.001, 0.002).$$

As they do not contain zero, it is justified to make further inferences based on the FH3 model. Moreover, the similarity between asymptotic-normal and bootstrap-percentile CIs indicates that both distributions are close, which is a reassuring finding.

For the diagnosis of the FH3 model, we consider the raw residuals (RR), defined by

$$\widehat{e}_{drt} = y_{drt} - \widehat{\mu}_{drt}, \ d = 1, \ldots, D, \ r = 1, \ldots, R, \ t = 1, \ldots, T,$$

and the standardized residuals (SR), defined by dividing by the standard deviation, i.e.

$$\widehat{e}_{drt}\nu^{-1}, \ \text{ where } \ \nu = \left(\frac{1}{DRT} \sum_{d=1}^{D} \sum_{r=1}^{T} \sum_{t=1}^{T} \left(\widehat{e}_{drt} - \widehat{\overline{e}}_{\cdots} \right)^2 \right)^{\frac{1}{2}}, \quad \widehat{\overline{e}}_{\cdots} = \frac{1}{DRT} \sum_{d=1}^{D} \sum_{r=1}^{R} \sum_{t=1}^{T} \widehat{e}_{drt}.$$

To detect outliers, three boxplots are shown in Figure 3.2. From left to right, the SRs are grouped by time period, province and main occupation. The last two boxplots use only data from the 2021.4 SLFS. We observe that: (1) the SRs present a homogeneous pattern over time periods, (2) the provinces have a more pronounced influence, although none of them shows particularly anomalous behaviour, and (3) the Hájek estimates tend to overestimate occupational categories OC1, OC2, OC3 and OC7 because their boxes fall mostly in the positive half-plane. Another important result that can be inferred is the adequacy of the SRs in terms of rank: they take values from -3 to 2, with a single outlier, located in Melilla.

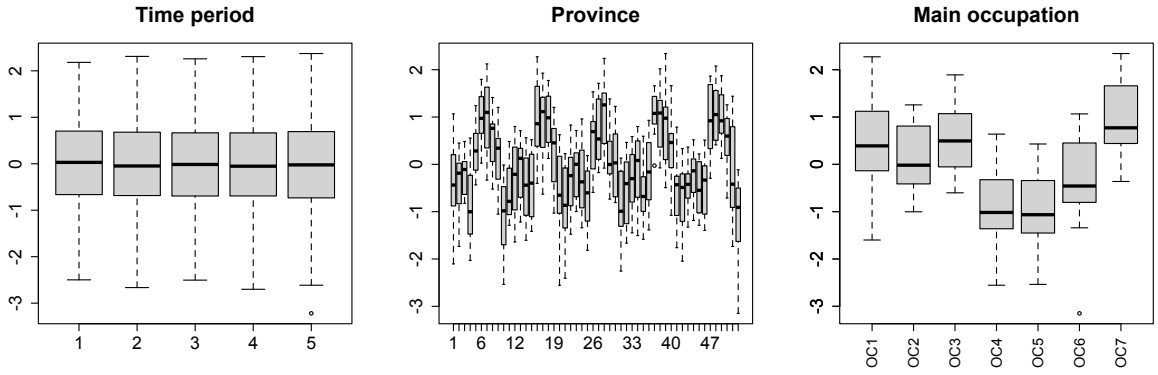


Figure 3.2: Boxplot of the SRs of the final FH3 model the SLFS 2020.4-2021.4 data by time period (left), province (center) and main occupation (right).

3.4.2 Prediction, error measures and maps

The aim now is to predict proportions of men who are employed in each occupational sector and, eventually, sex occupational segregation across provinces and time periods. Figure 3.3 (left) plots the EBLUPs and the Hájek direct estimates of the proportion of employed men in the last quarter of 2021. The dotted line $y = 0.5$ is included to compare the distance between both approaches and the balanced distribution of the population. As desired, it can be seen that model-based predictions smooth the behaviour of the Hájek estimates, with atypically high and low proportions, and show a better predictive performance. It is observed that the EBLUPs and the direct estimates follow the same trend, although the first ones are closer to $y = 0.5$. Figure 3.3 (right) includes some boxplots of the EBLUPs and the Hájek direct estimates of the proportion of employed men, for each occupational sector and the lastest time period SLFS2021.4. The boxes of the EBLUPs and the direct estimates follow the same pattern, although they are not completely identical.

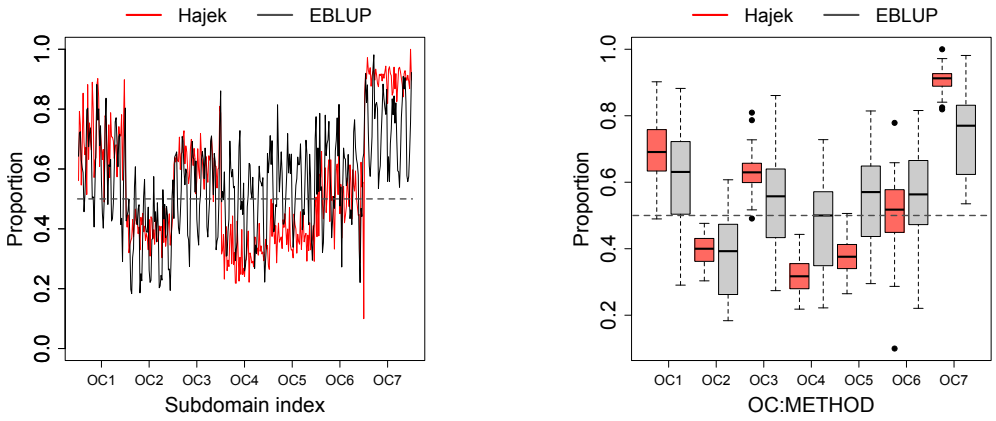


Figure 3.3: Line charts (left) and boxplots (right) of EBLUPs and Hájek estimates of the proportion of employed men. Data from the SLFS of 2021.4.

To make a fair comparison of the relative error measures, we estimate the RRMSE of $\widehat{\mu}_{drt}$ by dividing the squared root of the bootstrap estimate $mse^*(\widehat{\mu}_{drt})$, defined in (3.15), by the Hájek estimate y_{drt}. Next, we run the bootstrap algorithm with $B = 2000$ resamples, taking into account Remark 3.2.1, and estimate the RRMSE of the EBLUP as follows:

$$\text{RRMSE}(\widehat{\mu}_{drt}) = \frac{\sqrt{mse^*(\widehat{\mu}_{drt})}}{y_{drt}}, \quad d = 1, \ldots, D,\ r = 1, \ldots, R,\ t = 1, \ldots, T. \tag{3.17}$$

Table 3.9 contains the deciles of the model-based estimates of the RRMSEs of the EBLUP proportions of employed men and CVs of the Hájek estimator for the 2021.4 SLFS data. It is obtained that the deciles of the CVs prior to the median are lower, as they correspond to estimation domains with higher sample sizes, where direct estimates report reliable results. However, after the median, the CVs have higher deciles than those of the RRMSEs of the EBLUP. The reason, again, is the sample size.

Since the sample sizes are highly variable in our estimation domains for the SLFS data,

	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
n_{drt}	6	30	57	81	101	118	142	172	211	294	956
RRMSE	0.027	0.047	0.066	0.078	0.088	0.097	0.108	0.129	0.150	0.187	0.315
CV	0.000	0.029	0.063	0.077	0.088	0.099	0.113	0.127	0.153	0.187	0.373

Table 3.9: Deciles of sample sizes, RRMSEs of the EBLUP proportions of employed men and CVs of the Hájek estimator. Data from the SLFS of 2021.4.

it is advisable to use model-based predictors instead of direct estimators. Under the model-based approach, the EBLUP also has some theoretical good properties, such as asymptotic unbiasedness. Overall, the proposed model performs satisfactorily, both in terms of the significance level of the estimated parameters and in the reduction of the CVs of the Hájek estimator when the sample sizes are small. On balance, its use to calculate plug-in predictions of the DSI by province from 2020.4 to 2021.4 is justified.

Table 3.10 presents the provincial averages of the DSI plug-in predictions for $t = 5$, i.e.

$$\widehat{S}^{in}_{.r5} = \frac{1}{D}\sum_{d=1}^{D} \widehat{S}^{in}_{dr5}, \quad r = 1, \ldots, R. \tag{3.18}$$

Among the main occupations with highest DSI plug-in predictions, OC2 and OC7 stand out. Therefore, sex occupational segregation is concentrated in two main groups: high-skilled scientific and intellectual jobs and traditionally manual or low-skilled jobs (Figure 3.4).

	OC1	OC2	OC3	OC4
$\widehat{S}^{in}_{.r5}$	0.011	0.131	0.019	0.036
	OC5	OC6	OC7	AC
$\widehat{S}^{in}_{.r5}$	0.040	0.040	0.160	0.060

Table 3.10: For 2021.4, DSI plug-in predictions and mean value of the so-called average contributions (AC), given by

$$s_{dt} = \frac{1}{R}\sum_{r=1}^{R} S_{drt},\ d = 1, ..., D,\ t = 1, ..., T.$$

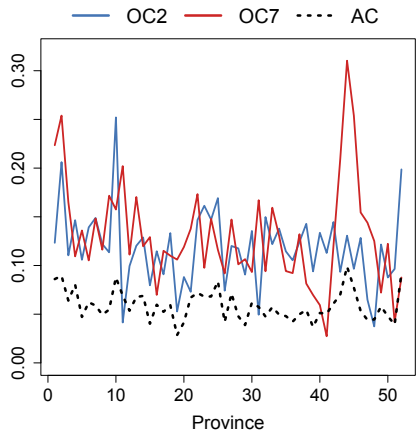


Figure 3.4: Unbalanced occupations and average contributions (AC) by province.

The problem of inclusion and equal opportunities particularly affects the lowest and most precarious occupational categories, and the highest positions of professor, manager, director or equivalent. Moreover, this gap not only has an immediate effect on women's labour conditions, but also on their career progression, as it is a process of continuous training and promotion. In contrast, directors and managers of public and private institutions and, in

general, accountants, administrative and other office employees work in less sex-segregated jobs. Intuitively, sex segregation is expected to be less evident in the public sector, where placement is theoretically based on objective merit criteria. However, multiple studies have shown that public institutions, such as universities, are not exempt from these problems either (Massó et al., 2021). Nevertheless, the average measures presented here mask the provincial variability of sex occupational segregation.

Figure 3.5 (left) colours Spain according to the DSI predictions for the fourth quarter of 2021. It therefore allows us to analyse how sex segregation differs across provinces. We observe that the largest discrepancies are found in Teruel, Albacete and Álava, among others. Indeed, between 30 and 35% of the employed population of Teruel would have to change their occupational sector to achieve a uniform distribution by province. The cause of the high sex segregation in Álava is owing to the mining industry. Historically, male labour has always been more predominant in this sector, including plant and machinery operators and assemblers, as well as the construction and mining industries. In the other highlighted provinces, sex segregation mainly occurs in the category OC2, which covers highly skilled scientific and intellectual jobs. Research claims that there is a sex gap that persists despite advances in the inclusion of women in the labour market in recent years and that is related to the unequal sharing of family responsabilities and the stigmas still present in modern societies.

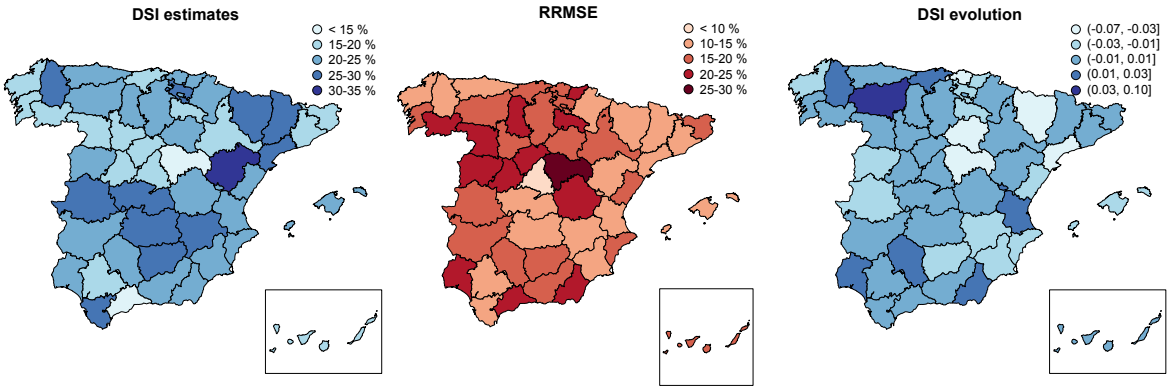


Figure 3.5: DSI predictions (left) and RRMSEs (center) for the SLFS of 2021.4, and evolution of the DSI predictions over the horizon 2020.4-2021.4 (right).

There is no clear spatial pattern in the sense that it is not possible to say that certain larger regions of the Iberian Peninsula are more prone to sex segregation than others. However, the distribution among provinces with similar demographic and socio-economic conditions is, in general, homogeneous. In terms of labour equality, the high predicted values for many provinces reveal the magnitude of the problem: the labour market disadvantages women and the occupational distribution is clearly non-homogeneous. According to our research, public and private institutions should implement measures of work equality and promote the inclusion of men and women in those sectors in which their presence is minority.

As for the error measures, we calculate the parametric bootstrap estimator of $\widehat{S}^{in}_{d.t}$, $mse^*(\widehat{S}^{in}_{d.t})$, given by (3.15). We generate $B = 2000$ bootstrap resamples, taking into account Remark

3.2.1. The RRMSE of $\widehat{S}^{in}_{d.t}$ is obtained by dividing the RMSE by the DSI estimates, i.e.

$$\text{RRMSE}(\widehat{S}^{in}_{d.t}) = \frac{\sqrt{mse^*(\widehat{S}^{in}_{d.t})}}{\widehat{S}^{in}_{d.t}}, \quad d = 1, \ldots, D,\ t = 1, \ldots, T. \tag{3.19}$$

Figure 3.5 (center) shows the bootstrap estimates of the RRMSE for the DSI predictions, which enables us to visually quantify the precision of our results. It can be concluded that most provinces are accompanied by RRMSEs below 25%, which is quite acceptable in the SAE setup. Most RRMSEs are lower than 20% and even 10% in several domains.

Taking advantage of the available temporal information, Figure 3.5 (right) maps the DSI differences between the last quarter of 2020 and the last quarter of 2021, i.e. $\widehat{S}^{in}_{d.5} - \widehat{S}^{in}_{d.1}$, $d = 1, \ldots, D$. We have observed that segregation shows appreciable changes over the observation period, with a maximum decrease close to 7 percentage units and a maximum increase bordering on 10 percentage. However, several provinces in the center of Spain do not seem to be affected by any change. In absolute terms, the situation has worsened in 17 provinces, improved in 7 and remained stable in 26 (between -0.01 and 0.01). In Madrid and Barcelona, which are the most populated regions, no significant changes have been predicted.

Even so, the changes observed over 2021 do not refer to a sufficiently long period of time to capture the results of potentially applicable policy decisions, and therefore they are not statistically significant (Bugallo et al., 2024d). But in spite of this, our model-based methodology provides relevant advances in the study of the temporal evolution of sex segregation in SAE situations. Consequently, the proposed approach could be applied in other studies with data from longer time periods, such as years or decades.

3.5 R codes

As for the `R` codes, the GitHub repository https://github.com/small-area-estimation/FH3DUNCAN (accessed on: November 4, 2024) contains our dataset and computer code, as well as a detailed description of its contents. It includes a README file that provides basic instructions for the correct execution of the available software.

Chapter 4

Unit-level multinomial mixed models and prediction of labour indicators

This chapter is self-contained and follows Bugallo et al. (2024a) as a reference point. It describes a new statistical methodology for the small area prediction of labour indicators under unit-level multinomial logit mixed models. Namely, the proportion of employed, unemployed and inactive people, and of unemployment rates. The novel empirical best and plug-in predictors are based on a multinomial mixed model for a trivariate response vector, and fitted to unit-level data. Model parameters are estimated by ML and MSEs by parametric bootstrap, following Hall and Maiti (2006) and González-Manteiga et al. (2008, 2010).

The first point is to motivate the applicability that guides this research. It stands to reason that unemployment is a cause of social instability that affects a country's economy and social welfare. The effects of unemployment can be economic, such as a decline in real productivity, a fall in demand or an increase in the public deficit. But it can also have social consequences, as psychological or discriminatory. The same applies to inactivity proportions, covering all citizens over the age of 16 who are neither employed nor unemployed. On this basis, accurate information is sought to monitor these problems so as to be able to take decisions aimed at reducing them. As a matter of fact, governments are interested in mapping labour indicators at a sufficient level of detail. This could be of great help in assessing the level of development and progress of a country and can be applied sequentially over time.

Our research deals with vector data that has a one-in-one component and zeros in the rest. That is, a one in the k-th component of the vector specifies that the respondent belongs to the k-th category of the employment status variable. Given that the employment status has three categories in the population residing in Spain, aged 16 and over, the response vector has the components $k = 1$ for employed and $k = 2$ for unemployed. Inactive category ($k = 3$) is complementary. The compositional models by Esteban et al. (2023) are not applicable to this type of data due to the amount of zeros. An alternative is the unit-level multinomial logit mixed model with a size parameter equal to one, in what follows called multi-BE model. It should be stressed that the estimation, inference and prediction in these models presents specific difficulties that are mitigated in the case of multinomial logit models with large size parameters. In multi-BE models, the approximation to the normal distribution cannot be

used. PQL estimators (Breslow and Clayton, 1993) might not be consistent, so it is high time to calculate ML estimators. Accordingly, we develop two specific algorithms to maximise the model log-likelihood, calculate ML estimators of the model parameters and predict the random effects.

The chapter is structured as follows. Section 4.1 presents the data and the SAE problem. Section 4.2 introduces the unit-level multinomial mixed model and the fitting algorithms. Section 4.2.1 describes the H-cubature algorithm and Section 4.2.2 describes the Laplace algorithm. Section 4.3 presents several model-based and population-based predictors. Section 4.4 develops bootstrap estimators of their MSEs. Section 4.5 contains simulation experiments to investigate the behavior of the two fitting algorithms, the predictors and the MSE estimators. Section 4.6 deals with the application to real data and the mapping of labour indicators. Data are from the first Spanish Labour Force Survey (SLFS) of 2021 to map labour indicators by province, sex and age group.

Supplementary Material, available online at *Journal of the Royal Statistical Society: Series A*[1], contains additional content, structured into 5 sections. Section A develops the theory of the more general unit-level multinomial logit mixed model for a target vector with $q \geqslant 3$ components. Section B sets out the mathematical background of the proposed predictors in a more general context, without imposing restrictions on the unit-level auxiliary information. Section C presents algorithms to calculate the bootstrap MSE estimators. Section D describes the steps of the simulations and contains further results. Section E plots additional maps for the application to real data.

4.1 Labour indicators and 2021.1 SLFS data

The application to real data aims at estimating labour indicators by Spanish province, sex (*sex 1*: men, *sex2*: women) and age group (*age1*: between 16 and 45 years; *age2*: between 46 and 55 years; *age3*: between 56 and 64 years; *age4*: 65 years or older). Data are from the Spanish Labour Force Survey (SLFS) of the first quarter of 2021 (SLFS2021.1), which covers about 58,000 dwellings, corresponding to 140,000 people. The population of interest, U, is made up of people aged 16 and over (*age*$\geqslant$16), with permanent residence in Spain. Respondents under the age of 16 are not taken into account, as they are below the minimum working-age in Spain. This reduces the size of the survey file to approximately 122,000 working-age respondents.

There are $D = 52 \cdot 2 = 104$ domains, $U_d \subset U$, defined by the crosses of province and sex, and $S = 52 \cdot 2 \cdot 4 = 416$ subdomains, $U_{d,t} \subset U_d$, defined by the crosses of province, sex and age group, respectively. The population U of size N is hierarchically partitioned in domains $U = \bigcup_{d=1}^{D} U_d$ and subdomains $U_d = \bigcup_{t=1}^{4} U_{d,t}$, $d = 1, \ldots, D$. The total number of people in a domain U_d and a subdomain $U_{d,t}$ are N_d and $N_{d,t}$ respectively, which only contain individuals aged 16 and over. The sizes N_d and $N_{d,t}$ are taken from the population projections published by the Spanish National Statistical Office (INE), and are the official sizes of the domains and subdomains. For $n = 122{,}000$ working-age respondents, it is expected an average of

[1] http://academic.oup.com/jrsssa/article-lookup/doi/10.1093/jrsssa/qnae033; accessed on: November 4, 2024.

$n/D = 1173$ and $n/S = 293$ respondents per domain and subdomain, respectively.

Going deeper into the issue of the sizes n_d and $n_{d,t}$ of the samples s_d and $s_{d,t}$, the estimated sampling fractions are

$$f_d = 100\frac{n_d}{\widehat{N}_d}, \quad f_{d,t} = 100\frac{n_{d,t}}{\widehat{N}_{d,t}}, \quad \widehat{N}_d = \sum_{j\in s_d} w_{dj}, \quad \widehat{N}_{d,t} = \sum_{j\in s_{d,t}} w_{dj}, \tag{4.1}$$

where $\widehat{N}_d$ and $\widehat{N}_{d,t}$ are the estimated domain and subdomain sizes and w_{dj} is the elevation factor of the j-th individual of s_d or $s_{d,t}$, $d = 1, \ldots, D$, $t = 1, \ldots, 4$.

Table 4.1 presents the deciles of the sample sizes (SS) and the estimated sampling fractions (SF) for subdomains (top) and domains (bottom). At the domain level, the sample sizes are all greater than 179, which appears to be large enough to obtain sufficiently accurate direct estimates. However, we observe that 20% (40%) of the subdomain sample sizes are lower than 106 (178) and the average sample size 293 is between $q_{0.6} = 266$ and $q_{0.7} = 314$, which suggests that the sample size distribution is positively skewed. Added to that, sampling fractions allow us to know the percentage of individuals from domains or subdomains that actually belong to the sample. As they are all lower than 1.804 (in %), the representativeness of the samples is low. Since sample sizes and sampling fractions are small in many subdomains, mapping labour indicators using direct estimators is not sufficiently accurate. This motivates the incorporation of model-based predictors using SAE methods.

		q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
Subdomain	SS	11	74	106	138	178	222	266	314	454	576	1821
	SF	0.112	0.220	0.273	0.331	0.398	0.449	0.512	0.625	0.725	0.923	1.804
Domain	SS	179	555	611	660	959	1046	1169	1248	1451	2227	3916
	SF	0.126	0.219	0.277	0.311	0.370	0.434	0.494	0.577	0.679	0.860	1.501

Table 4.1: Deciles of the sample sizes (SS) and the estimated sampling fractions (SF), in %, for subdomains (top) and domains (bottom) in the SLFS2021.4.

The binary target variables, y_{dj1}, y_{dj2} and y_{dj3}, are equal to 1 if individual j of domain d is employed, unemployed, and inactive, respectively, and are equal to 0 otherwise. Table 4.2 provides an overview of the distribution of the respondents according to their employment status, sex and age group. We observe that active population is mostly concentrated in the first two age groups, with inactivity proportions increasing from the age of 56 onwards, with a very notable jump at the age of 65 (retirement age). The latter justifies why the total number of employed and unemployed respondents decreases with age group. In addition, unemployment and inactivity are more common among female respondents, regardless of the age group they belong to.

The domain and subdomain labour indicators of interest are the proportions of employed, unemployed and inactive people and the corresponding unemployment rates, i.e.

$$\overline{Y}_{dk} = \frac{1}{N_d}\sum_{j\in U_d} y_{djk}, \quad \overline{Y}_{dk,t} = \frac{1}{N_{d,t}}\sum_{j\in U_{d,t}} y_{djk}, \quad k = 1,2,3;\ d = 1, \ldots, D,\ t = 1, \ldots, 4, \tag{4.2}$$

age4	men	women	men	women	men	women	men	women
1	27038	27555	17399	15699	3058	3690	6581	8166
2	11352	12232	8724	7540	1003	1285	1625	3407
3	5095	5597	2447	2155	339	276	2309	3166
4	14655	18560	495	360	16	23	14144	18177
	respondents		employed		unemployed		inactive	

Table 4.2: From left to right, number of respondents by employment status, sex (columns) and age group (rows) in the SLFS2021.1.

and

$$R_d = \frac{\overline{Y}_{d2}}{\overline{Y}_{d1} + \overline{Y}_{d2}}, \quad R_{d,t} = \frac{\overline{Y}_{d2,t}}{\overline{Y}_{d1,t} + \overline{Y}_{d2,t}}, \quad d = 1, \ldots, D,\ t = 1, \ldots, 4. \tag{4.3}$$

The quantities (4.2) and (4.3) can be estimated using direct estimators, that is, relying only on data from sample units in the domain and subdomain of interest, respectively. The Hájek estimators of $\overline{Y}_{dk}$ and $\overline{Y}_{dk,t}$ are

$$\widehat{\overline{Y}}_{dk}^{dir} = \frac{\sum_{j \in s_d} w_{dj} y_{djk}}{\sum_{j \in s_d} w_{dj}}, \ \widehat{\overline{Y}}_{dk,t}^{dir} = \frac{\sum_{j \in s_{d,t}} w_{dj} y_{djk}}{\sum_{j \in s_{d,t}} w_{dj}}, \quad k = 1, 2, 3;\ d = 1, \ldots, D,\ t = 1, \ldots, 4, \tag{4.4}$$

and the Hájek estimators of R_d and $R_{d,t}$ are

$$\widehat{R}_d^{dir} = \frac{\widehat{\overline{Y}}_{d2}^{dir}}{\widehat{\overline{Y}}_{d1}^{dir} + \widehat{\overline{Y}}_{d2}^{dir}}, \quad \widehat{R}_{d,t}^{dir} = \frac{\widehat{\overline{Y}}_{d2,t}^{dir}}{\widehat{\overline{Y}}_{d1,t}^{in} + \widehat{\overline{Y}}_{d2,t}^{dir}}, \quad d = 1, \ldots, D,\ t = 1, \ldots, 4. \tag{4.5}$$

4.2 Unit-level multinomial logit mixed model

Let y_{djk} be the indicator variable of the labour status k for individual j of domain d. That is, $y_{djk} = 1$ if individual j of domain d is in labour status k, $y_{djk} = 0$ otherwise, and $y_{dj1} + y_{dj2} + y_{dj3} = 1$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$. Let $n = \sum_{d=1}^{D} n_d$ be the global sample size. For $k = 1, 2$, let $x_{djk} = (x_{djk1}, \ldots, x_{djkp_k})$ be a row vector containing p_k auxiliary variables and let $\beta_k = (\beta_{k1}, \ldots, \beta_{kp_k})'$ be a column vector of size p_k containing the model parameters, with $p = p_1 + p_2$. For $d = 1, \ldots, D$, $k = 1, 2$, let us consider independent random effects $u_{dk} \sim N(0, 1)$. The domain random effects are $u_d = (u_{d1}, u_{d2})' \sim N(0, I_2)$, $d = 1, \ldots, D$, where I_m denotes the $m \times m$ identity matrix, and $u = \underset{1 \leqslant d \leqslant D}{\text{col}}(u_d) \sim N(0, I_{2D})$.
The probability density function (p.d.f.) of u is

$$f(u) = \prod_{d=1}^{D} f(u_d) = \prod_{d=1}^{D} \prod_{k=1}^{2} f(u_{dk}) = (2\pi)^{-D} \exp\left\{-\frac{1}{2}\, u'u\right\}.$$

The unit-level multinomial logit mixed model, with independent domain-category random effects, assumes that the distribution of the target vector $y_{dj} = (y_{dj1}, y_{dj2})'$, conditioned to

the random vector u_d, is multinomial with size parameter equal to one, i.e.,

$$y_{dj}|_{u_d} \sim M(1; p_{dj1}, p_{dj2}), \quad d = 1, \ldots, D,\ j = 1, \ldots, n_d, \tag{4.6}$$

with the logit link for the natural parameter, i.e.

$$\eta_{djk} = \log \frac{p_{djk}}{p_{dj3}} = x_{djk}\beta_k + \phi_k u_{dk}, \quad d = 1, \ldots, D,\ j = 1, \ldots, n_d,\ k = 1, 2, \tag{4.7}$$

where $p_{dj1} + p_{dj2} + p_{dj3} = 1$, $p_{djk} > 0$, $\phi_k > 0$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, $k = 1, 2, 3$. Here the multinomial distribution is denoted by the letter M. Finally, the model assumes that the vectors y_{dj}'s are independent conditioned to u. The vector of model parameters is $\theta = (\beta', \phi')'$, where $\beta = (\beta_1', \beta_2')'$ and $\phi = (\phi_1, \phi_2)'$. The model formula (4.7) has random intercepts $v_{dk} = \phi_k u_{dk} \sim N(0, \phi_k^2)$, $d = 1, \ldots, D$, $k = 1, 2$.

For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, the conditioned probability of y_{dj}, given u, is

$$P_\theta(y_{dj}|u) = P_\theta(y_{dj}|u_d) = p_{dj1}^{y_{dj1}} p_{dj2}^{y_{dj2}} p_{dj3}^{y_{dj3}},$$

where

$$p_{dj3} = \frac{1}{1 + \exp\{\eta_{dj1}\} + \exp\{\eta_{dj2}\}}, \quad p_{djk} = \frac{\exp\{\eta_{djk}\}}{1 + \exp\{\eta_{dj1}\} + \exp\{\eta_{dj2}\}}, \quad k = 1, 2.$$

The vectors of dimensions $2n_d \times 1$ and $2n \times 1$ that contain the values of the target variables are $y_d = \underset{1 \leqslant j \leqslant n_d}{\text{col}}(y_{dj})$ and $y = \underset{1 \leqslant d \leqslant D}{\text{col}}(y_d)$, respectively. The model likelihood is

$$P_\theta(y) = \int_{\mathbb{R}^{2D}} P_\theta(y|u) f(u)\, du, \quad P_\theta(y|u) = \prod_{d=1}^{D} \prod_{j=1}^{n_d} P(y_{dj}|u). \tag{4.8}$$

The ML parameter estimator $\widehat{\theta}$ maximizes the log-likelihood function $\ell(\theta; y) = \log P_\theta(y)$. Since the objective function is a multidimensional integral, we present two maximization approaches that combine approximation and optimization algorithms. In doing so, the aim is to provide computationally efficient estimates with high accuracy (Bugallo et al., 2024a). The first algorithm (algorithm HC) contains a sub-algorithm to approximate multiple integrals and a derivative-free optimization algorithm (algorithm NB) to maximize the approximated log-likelihood. Similarly, the Laplace algorithm contains a sub-algorithm to approximate multiple integrals (algorithm NR) and algorithm NB to maximize the approximated log-likelihood. The following two subsections describe the H-cubature and Laplace algorithms.

4.2.1 H-cubature algorithm

Algorithm HC uses the H-cubature approach to calculate a quadrature approximation of the integral in (4.8) at each step of a given optimization algorithm. For this sake, we apply the `hcubature` function of the `R cubature` package. The algorithm and its properties are described by Genz and Malik (1980) and Berntsen et al. (1991). The `hcubature` function performs adaptive multidimensional integration of vector-valued integrands over hypercubes,

that recursively subdivide the integration domain into smaller subdomains, using the same rule (weighted sum of integrand values in evaluation points), until convergence is achieved. In each subdomain, the use of this rule gives a vector of integration results, a vector of error estimates and a coordinate for a subsequent subdivision of that subdomain.

Concerning the optimization of the log-likelihood function, we use the `nloptr` function of the `R` package `nloptr`, with the global optimization method NEWUOA (Powell, 2004) adapted to bound constraints. Hence, we implement the derivative-free algorithm NB (NEWUOA-BOUND), which is a variant of the method NEWUOA that supports constraint problems and iteratively constructs quadratic approximations for the objective function.

Nesting the `hcubature` function to approximate the log-likelihood of the multinomial mixed model in the optimization function `nloptr` allows any type of integral function to be maximized. It is a simple solution by chaining two `R` functions, but it is a global algorithm. Besides, it is not an algorithm adapted to the log-likelihood that must be maximized.

4.2.2 Laplace algorithm

Let $h : \mathbb{R}^m \mapsto \mathbb{R}$ be a continuously twice differentiable function with a global maximum at x_0. This is to say, let us assume that the vector of first partial derivatives is $\dot{h}(x_0) = \frac{\partial h}{\partial x}\big|_{x=x_0} = 0$ and the matrix of second partial derivatives, $\ddot{h}(x_0) = \frac{\partial^2 h}{\partial x^2}\big|_{x=x_0}$, is negative definite. A Taylor series expansion of $h(x)$ around x_0 yields to

$$\begin{aligned} h(x) &= h(x_0) + \dot{h}'(x_0)(x - x_0) + \frac{1}{2}(x - x_0)'\ddot{h}(x_0)(x - x_0) + o\big(\|x - x_0\|^2\big) \\ &\approx h(x_0) + \frac{1}{2}(x - x_0)'\ddot{h}(x_0)(x - x_0). \end{aligned}$$

The multivariate Laplace approximation is

$$\begin{aligned} \int_{\mathbb{R}^m} e^{h(x)}\, dx &\approx \int_{\mathbb{R}^m} e^{h(x_0)} \exp\Big\{ -\frac{1}{2}(x - x_0)'\big(-\ddot{h}(x_0)\big)(x - x_0)\Big\}\, dx \\ &= (2\pi)^{m/2}\big|-\ddot{h}(x_0)\big|^{-1/2} e^{h(x_0)}. \end{aligned} \tag{4.9}$$

Let us now approximate the likelihood (4.8) of the unit-level multinomial logit mixed model. As the target vectors $y_1, \ldots, y_D$ are unconditionally independent and the random vectors $u_d \sim N(0, I_2)$ are independent, the marginal distribution of y_d is

$$\begin{aligned} P_\theta(y_d) &= \int_{\mathbb{R}^2} P_\theta(y_d|u_d) f(u_d)\, du_d = \int_{\mathbb{R}^2} \left(\prod_{j=1}^{n_d} p_{dj1}^{y_{dj1}} p_{dj2}^{y_{dj2}} p_{dj3}^{y_{dj3}} \right) f(u_d)\, du_d \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \exp\Big\{ -\sum_{j=1}^{n_d} \log\Big(1 + \sum_{k=1}^{2} \exp\{x_{djk}\beta_k + \phi_k u_{dk}\}\Big) \\ &+ \sum_{k=1}^{2}\sum_{j=1}^{n_d} y_{djk}(x_{djk}\beta_k + \phi_k u_{dk}) - \frac{1}{2}\, u_d' u_d \Big\}\, du_d = \frac{1}{2\pi} \int_{\mathbb{R}^2} \exp\{h_d(u_d)\}\, du_d, \end{aligned} \tag{4.10}$$

where

$$\begin{aligned} h_d(u_d) &= h_d(u_d; y_d, \theta) = -\sum_{j=1}^{n_d} \log\Big(1 + \sum_{k=1}^{2} \exp\{x_{djk}\beta_k + \phi_k u_{dk}\}\Big) \\ &+ \sum_{k=1}^{2}\sum_{j=1}^{n_d} y_{djk}(x_{djk}\beta_k + \phi_k u_{dk}) - \frac{1}{2}\, u_d' u_d. \end{aligned} \tag{4.11}$$

So as to apply the Laplace algorithm to the integral in (4.10), we maximize $h_d(u_d; y_d, \theta)$ in u_d, given y_d and θ. We could carry out the maximization by applying an `R` function of optimization. Alternatively, we implement a Newton-Raphson (NR) algorithm after calculating the first and second partial derivatives of h with respect to u_d, $d = 1, \ldots, D$, given y and θ.

The first derivatives of h_d with respect to u_{dk}, $k = 1, 2$, are

$$\frac{\partial h_d(u_d)}{\partial u_{dk}} = \sum_{j=1}^{n_d} \big\{ -\phi_k p_{djk} + \phi_k y_{djk} \big\} - u_{dk}.$$

The second derivatives of h_d with respect to u_{dk}, $k = 1, 2$, are

$$\frac{\partial^2 h_d(u_d)}{\partial u_{dk}^2} = -1 - \phi_k^2 \sum_{j=1}^{n_d} p_{djk}(1 - p_{djk}), \quad \frac{\partial^2 h_d(u_d)}{\partial u_{dk_1}\partial u_{dk_2}} = \phi_{k_1}\phi_{k_2} \sum_{j=1}^{n_d} p_{djk_1} p_{djk_2}, \quad k_1 \neq k_2,$$

since

$$\frac{\partial p_{djk}}{\partial u_{dk}} = \phi_k p_{djk}(1 - p_{djk}) \quad \text{and} \quad \frac{\partial p_{djk_1}}{\partial u_{dk_2}} = -\phi_{k_2} p_{djk_1} p_{djk_2}, \quad k_1 \neq k_2.$$

The score vector and the Jacobian matrix are

$$S_d(u_d, \theta) = \Big(\frac{\partial h_d(u_d)}{\partial u_{d1}}, \frac{\partial h_d(u_d)}{\partial u_{d2}}\Big)', \quad H_d(u_d, \theta) = \begin{pmatrix} \frac{\partial^2 h_d(u_d)}{\partial u_{d1}^2} & \frac{\partial^2 h_d(u_d)}{\partial u_{d1}\partial u_{d2}} \\ \frac{\partial^2 h_d(u_d)}{\partial u_{d2}\partial u_{d1}} & \frac{\partial^2 h_d(u_d)}{\partial u_{d2}^2} \end{pmatrix}.$$

For $\theta = (\beta', \phi')'$ fixed, the function $h_d(u_d)$, defined in (4.11), is maximized by using the Newton-Raphson algorithm. The updating equation is

$$u_d^{(r+1)} = u_d^{(r)} - H_d^{-1}(u_d^{(r)}, \theta) S_d(u_d^{(r)}, \theta), \quad d = 1, \ldots, D. \tag{4.12}$$

Let us denote by u_{0d} the argument of maxima of the function $h_d(u_d)$. It holds that $\dot{h}(u_{0d}) = 0$ and $\ddot{h}(u_{0d}) = H_d(u_{0d}, \theta)$ is negative definite. The model log-likelihood is

$$\ell = \ell(\theta; y) = \sum_{d=1}^{D} \log P_\theta(y_d) = \sum_{d=1}^{D} \ell_d.$$

By applying (4.9) at $u_d = u_{0d}$ to (4.10), we obtain the Laplace approximation ℓ_{0d} of the term ℓ_d, where

$$\begin{aligned} \ell_{0d} &= \ell_{0d}(\theta; y, u_{0d}) = h_d(u_{0d}) - \frac{1}{2}\log|-H_d(u_{0d}, \theta)| \\ &= -\sum_{j=1}^{n_d} \log\Big(1 + \sum_{k=1}^{2} \exp\{x_{djk}\beta_k + \phi_k u_{0dk}\}\Big) + \sum_{k=1}^{2}\sum_{j=1}^{n_d} y_{djk}(x_{djk}\beta_k + \phi_k u_{0dk}) \\ &- \frac{1}{2}\, u_{0d}' u_{0d} - \frac{1}{2}\log|-H_d(u_{0d}, \theta)|. \end{aligned} \tag{4.13}$$

The following step is to maximize $\ell_0(\theta) \stackrel{\Delta}{=} \sum_{d=1}^{D} \ell_{0d}(\theta; y, u_{0d})$ in $\theta \in \Theta$ by applying the algorithm NB. The final Laplace algorithm combines the approximation NR algorithm (4.12) and the optimization algorithm NB. It is described by the following steps:

1. Set the initial values $r = 0$, $\theta^{(0)}$, $\theta^{(-1)} = \theta^{(0)} + 1_{p+2}$, $u_d^{(0)} = 0_2$, $u_d^{(-1)} = 1_2$, $d = 1, \ldots, D$.
2. Until $\|\theta^{(r)} - \theta^{(r-1)}\|_2 < \varepsilon_1$, $\|u_d^{(r)} - u_d^{(r-1)}\|_2 < \varepsilon_2$, $d = 1, \ldots, D$, do
   (a) Apply algorithm (4.12) with seeds $u_d^{(r)}$, $d = 1, \ldots, D$, convergence tolerance ε_2 and $\theta = \theta^{(r)}$ fixed. Output: $u_d^{(r+1)}$, $d = 1, \ldots, D$.
   (b) Apply algorithm NB with seed $\theta^{(r)}$, convergence tolerance ε_1 and $u_d = u_d^{(r+1)}$ fixed, $d = 1, \ldots, D$. Output: $\theta^{(r+1)}$.
   (c) $r \leftarrow r + 1$.
3. Output: $\widehat{\theta} = \theta^{(r)}$, $\widehat{u}_d = u_d^{(r)}$, $d = 1, \ldots, D$.

Let us note that the Laplace algorithm gives at convergence not only ML parameter estimators of the model parameters, but also modal predictors of the random effects.

4.3 Small area prediction of labour indicators

This section derives predictors of unit-level probabilities and domain-level and subdomain-level population-dependent labour indicators. Once the unit-level multinomial logit mixed model (4.6)-(4.7) has been fitted to the sample data, the construction of small area predictors of the quantities of interest is based on model elements, target vector values in sample units and, in some cases, auxiliary variable values in population units. First, we predict the probability p_{djk} that individual j of domain d belongs to labour status k. Second, we predict y_{djk} and the domain totals and means. In the second case, we further derive predictors of unemployment rates. To differentiate sample vectors of size n_d and non-sample vectors of size $N_d - n_d$, from population vectors of size N_d, we introduce the notation

$$y_s = \underset{1 \leqslant d \leqslant D}{\text{col}}(y_{ds}),\ y_r = \underset{1 \leqslant d \leqslant D}{\text{col}}(y_{dr});\ y_{ds} = \underset{j \in s_d}{\text{col}}(y_{dj}),\ y_{dr} = \underset{j \in r_d}{\text{col}}(y_{dj});\ y_{dj} = \underset{1 \leqslant k \leqslant 2}{\text{col}}(y_{djk}),$$

where $s_d \subset U_d$ and $r_d = U_d - s_d$ are the sample and non-sample subsets of U_d, respectively.

Section 4.3 and Section 4.4 are fully adapted to the application to real data in Section 4.6. In particular, the covariate *age4* defined in Section 4.1, which delimits the subdomains $U_{d,t}$, takes $T = 4$ values. For this reason, we assume that the covariates have the same size $p_1 = p_2 \stackrel{\Delta}{=} p_0$ and contain categorical variables, such that

$$x_{djk} \in \{z_t \in \mathbb{R}^{p_0} : t = 1, \ldots, T\},\ d = 1, \ldots, D,\ j = 1, \ldots, N_d,\ k = 1, 2. \tag{4.14}$$

We also suppose that all the components of the target variable y_{dj} are explained with the same set of auxiliary variables, i.e. $x_{djk} = x_{dj}$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, $k = 1, 2$. Under

this categorical setup, it holds that $p_{djk} = p_{dk,t}$, $j \in U_{d,t} = \{j \in U_d : x_{dj} = z_t\}$, where

$$\sum_{j=1}^{N_d} p_{djk} = \sum_{t=1}^{T} N_{d,t} p_{dk,t}, \quad p_{dk,t} = p_{dk,t}(\theta, u_d) = \frac{\exp\{z_t\beta_k + \phi_k u_{dk}\}}{1 + \sum_{\ell=1}^{2} \exp\{z_t\beta_\ell + \phi_\ell u_{d\ell}\}}. \tag{4.15}$$

The sample and non-sample subsets of $U_{d,t}$ are $s_{d,t} = \{j \in s_d : x_{dj} = z_t\}$ and $r_{d,t} = \{j \in r_d : x_{dj} = z_t\}$, and the corresponding target vectors are $y_{d,ts} = \underset{j \in s_{d,t}}{\text{col}}\,(y_{dj})$ and $y_{d,tr} = \underset{j \in r_{d,t}}{\text{col}}\,(y_{dj})$. The size of $r_{d,t}$ is denoted by $N_{d,tr}$.

Remark 4.3.1. *In the simulation study presented in Section 4.5, we assume a more general setup, where the auxiliary variables x_{dj1} and x_{dj2} do not have to be equal. The corresponding formulas for the predictors under this more general case, and for any arbitrary values of q and T, are presented in Supplementary Material, available online at Journal of the Royal Statistical Society: Series A. For the sake of completeness, the mentioned Section B develops the contents of Section 4.3 and Section 4.4 also without assuming the categorical setup (4.14), i.e. for a situation that may include continuous auxiliary variables.*

Predictors of $p_{dk,t}$

First we look for a predictor $\widehat{p}_{dk,t}$ of $p_{dk,t} = p_{dk,t}(\theta, u_d)$ with minimum MSE, $E_\theta\big[(\widehat{p}_{dk,t} - p_{dk,t})^2\big]$, in the class of unbiased predictors, $E_\theta\big[\widehat{p}_{dk,t} - p_{dk,t}\big] = 0$, for θ known. The predictor fulfilling these properties is the BP, given by $\widehat{p}^{\,bp}_{dk,t}(\theta) = E_\theta[p_{dk,t}|y_s]$. It follows that

$$E_\theta[p_{dk,t}|y_s] = E_\theta[p_{dk,t}|y_{ds}] = \frac{\int_{\mathbb{R}^2} \frac{\exp\{z_t\beta_k + \phi_k u_{dk}\}}{1+\sum_{\ell=1}^{2}\exp\{z_t\beta_\ell + \phi_\ell u_{d\ell}\}} P(y_{ds}|u_d) f(u_d)\, du_d}{\int_{\mathbb{R}^2} P(y_{ds}|u_d) f(u_d)\, du_d} = \frac{A_{dk,t}}{D_d},$$

where $A_{dk,t} = A_{dk,t}(y_{d.}, \theta)$ and $D_d = D_d(y_{d.}, \theta)$ are

$$\begin{aligned} A_{dk,t} &= \int_{\mathbb{R}^2} \frac{\exp\{z_t\beta_k + \phi_k u_{dk}\}}{1 + \sum_{\ell=1}^{2}\exp\{z_t\beta_\ell + \phi_\ell u_{d\ell}\}} \\ &\cdot \exp\left\{ \sum_{k=1}^{2} \phi_k y_{d.k} u_{dk} - \sum_{j=1}^{n_d} \log\Big[1 + \sum_{\ell=1}^{2} \exp\{x_{dj\ell}\beta_\ell + \phi_\ell u_{d\ell}\}\Big]\right\} f(u_d)\, du_d, \\ D_d &= \int_{\mathbb{R}^2} \exp\left\{ \sum_{k=1}^{2} \phi_k y_{d.k} u_{dk} - \sum_{j=1}^{n_d} \log\Big[1 + \sum_{\ell=1}^{2} \exp\{x_{dj\ell}\beta_\ell + \phi_\ell u_{d\ell}\}\Big]\right\} f(u_d)\, du_d, \end{aligned} \tag{4.16}$$

and we have denoted $y_{d.} = (y_{d.1}, y_{d.2})'$ and $y_{d.k} = \sum_{j=1}^{n_d} y_{djk}$, $k = 1, 2$.

The EBP of $p_{dk,t}$ is obtained by substituting θ by its estimate $\widehat{\theta}$, i.e. $\widehat{p}^{\,ebp}_{dk,t} = \widehat{p}^{\,bp}_{dk,t}(\widehat{\theta})$. In practice, it is approximated by a Monte Carlo method as follows.

1. Calculate the ML parameter estimator $\widehat{\theta} = (\widehat{\beta}', \widehat{\phi})'$.
2. For $s = 1, \ldots, S$, generate $u_d^{(s)}$ i.i.d. $N(0, I_2)$ and set $u_d^{(S+s)} = -u_d^{(s)}$.

3. Calculate $\widehat{p}^{ebp}_{dk,t} = \widehat{A}_{dk,t}/\widehat{D}_d$, where

$$\begin{aligned}\widehat{A}_{dk,t} &= \frac{1}{2S}\sum_{s=1}^{2S}\frac{\exp\{z_t\widehat{\beta}_k + \widehat{\phi}_k u^{(s)}_{dk}\}}{1+\sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell + \widehat{\phi}_\ell u^{(s)}_{d\ell}\}} \\ &\cdot \exp\left\{\sum_{k=1}^{2}\widehat{\phi}_k y_{d.k}u^{(s)}_{dk} - \sum_{j=1}^{n_d}\log\left[1+\sum_{\ell=1}^{2}\exp\{x_{dj\ell}\widehat{\beta}_\ell + \widehat{\phi}_\ell u^{(s)}_{d\ell}\}\right]\right\}, \\ \widehat{D}_d &= \frac{1}{2S}\sum_{s=1}^{2S}\exp\left\{\sum_{k=1}^{2}\widehat{\phi}_k y_{d.k}u^{(s)}_{dk} - \sum_{j=1}^{n_d}\log\left[1+\sum_{\ell=1}^{2}\exp\left\{x_{dj\ell}\widehat{\beta}_\ell + \widehat{\phi}_\ell u^{(s)}_{d\ell}\right\}\right]\right\}.\end{aligned} \tag{4.17}$$

The interesting feature here is that the BP has the minimum MSE in the class of unbiased predictors. Unfortunately, this is not the case for the EBPs, which are obtained by replacing the true model parameters with their estimates, and are therefore not unbiased. Under the assumption that the estimates of the model parameters are consistent, the EBPs are asymptotically unbiased, but the domain sample sizes in SAE problems are typically small. Even in the case of θ known, we have to approximate the integrals of the numerator and denominator, either by Monte Carlo or by another numerical methods. Integral approximations have a high computational cost and introduce a second source of error that increases the variance of the final predictors. As a matter of fact, $\widehat{p}^{ebp}_{dk,t}$ should be called empirical approximated BP. Due to these drawbacks, we consider less computationally-demanding plug-in predictors.

The plug-in predictor of $p_{dk,t}$ is

$$\widehat{p}^{in}_{dk,t} = \frac{\exp\{z_t\widehat{\beta}_k + \widehat{\phi}_k\widehat{u}_{dk}\}}{1+\sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell + \widehat{\phi}_\ell\widehat{u}_{d\ell}\}}, \tag{4.18}$$

where $\widehat{u}_{dk}$ is the modal predictor of u_{dk}, $d = 1,\ldots,D$, $k = 1,2$, which is obtained at the output of the Laplace algorithm, given in Section 4.2.2.

Predictors of y_{djk}

The BP of y_{djk} is $\widehat{y}^{bp}_{djk} = E_\theta[y_{djk}|y_s]$. If $j \in s_{d,t}$, then $E_\theta[y_{djk}|y_s] = y_{djk}$. If $j \in r_{d,t}$, then $E_\theta[y_{djk}|y_s] = E_\theta[y_{djk}|y_{ds}]$ and

$$\begin{aligned}\widehat{y}^{bp}_{djk}(\theta) &= E_\theta[y_{djk}|y_{ds}] = \frac{\int_{\mathbb{R}^2}\left\{\sum_{y_{djk}=0}^{1} y_{djk}P(y_{djk}|u_d)\right\}P(y_{ds}|u_d)f(u_d)\,du_d}{\int_{\mathbb{R}^2}P(y_{ds}|u_d)f(u_d)\,du_d} \\ &= \frac{\int_{\mathbb{R}^2} p_{dk,t}P(y_{ds}|u_d)f(u_d)\,du_d}{\int_{\mathbb{R}^2}P(y_{ds}|u_d)f(u_d)\,du_d} = \frac{A_{dk,t}(y_{d.},\theta)}{D_d(y_{d.},\theta)} = E_\theta[p_{dk,t}|y_{ds}] = \widehat{p}^{bp}_{dk,t}(\theta),\end{aligned}$$

where $A_{dk,t} = A_{dk,t}(y_{d.},\theta)$ and $D_d = D_d(y_{d.},\theta)$ are defined in (4.16).

The EBP of y_{djk} is $\widehat{y}^{ebp}_{djk} = \widehat{y}^{bp}_{djk}(\widehat{\theta})$. It holds that $\widehat{y}^{ebp}_{djk} = y_{djk}$ if $j \in s_{d,t}$ and $\widehat{y}^{ebp}_{djk} = \widehat{p}^{ebp}_{dk,t}$ if $j \in r_{d,t}$, where $\widehat{p}^{ebp}_{dk,t}$ is the EBP of $p_{dk,t}$, which is approximated in (4.17).

The plug-in predictor of y_{djk} is $\widehat{y}^{in}_{djk} = y_{djk}$ if $j \in s_{d,t}$ and $\widehat{y}^{in}_{djk} = \widehat{p}^{in}_{dk,t}$ if $j \in r_{d,t}$, where $\widehat{p}^{in}_{dk,t}$ is the plug-in predictor of $p_{dk,t}$, which is given in (4.18).

Predictors of proportions

Under unit-level non-linear models, the construction of small area predictors requires the availability of census files. As this is not the case of the application to real data, we have assumed the categorical setup (4.14). The EBP and the plug-in predictor of $\overline{Y}_{dk}$ are

$$\widehat{\overline{Y}}_{dk}^{ebp} = \widehat{\overline{Y}}_{dk}(\widehat{\theta}) = \frac{1}{N_d}\Big\{ \sum_{j\in s_d} y_{djk} + \sum_{t=1}^{T} N_{d,tr}\widehat{p}_{dk,t}^{ebp}\Big\}, \;\; \widehat{\overline{Y}}_{dk}^{in} = \frac{1}{N_d}\Big\{ \sum_{j\in s_d} y_{djk} + \sum_{t=1}^{T} N_{d,tr}\widehat{p}_{dk,t}^{in}\Big\}, \quad (4.19)$$

and the EBP and the plug-in predictor of $\overline{Y}_{dk,t}$ are

$$\widehat{\overline{Y}}_{dk,t}^{ebp} = \widehat{\overline{Y}}_{dk,t}(\widehat{\theta}) = \frac{1}{N_{d,t}}\Big\{ \sum_{j\in s_{d,t}} y_{djk} + N_{d,tr}\widehat{p}_{dk,t}^{ebp}\Big\}, \;\; \widehat{\overline{Y}}_{dk,t}^{in} = \frac{1}{N_{d,t}}\Big\{ \sum_{j\in s_{d,t}} y_{djk} + N_{d,tr}\widehat{p}_{dk,t}^{in}\Big\}. \quad (4.20)$$

Predictors of non-linear quantities

Finally, we derive predictors of non-linear quantities $G_d = g_d(y_d)$, where $g_d : \mathbb{R}^{2N_d} \mapsto \mathbb{R}$ is a continuous function. Similarly, we obtain the corresponding predictors for subdomains. For brevity, we focus on unemployment rates, defined by

$$g_d(y_d) = R_d(y_d) = \frac{\overline{Y}_{d2}}{\overline{Y}_{d1} + \overline{Y}_{d2}}.$$

First, the BP of $G_d = g_d(y_d)$ is $\widehat{G}_d^{bp}(\theta) = E_\theta\big[g_d(y_d)|y_{ds}\big]$.

The conditional probability of y_{dr}, given y_{ds}, is

$$P(y_{dr}|y_{ds}) = \frac{P(y_{ds}, y_{dr})}{P(y_{ds})} = \frac{\int_{\mathbb{R}^2} P(y_{dr}|u_d)P(y_{ds}|u_d)f(u_d)\,du_d}{\int_{\mathbb{R}^2} P(y_{ds}|u_d)f(u_d)\,du_d}, \quad y_{dr} \in \mathcal{Y}_{dr},$$

where $\mathcal{Y}_{dr} = \Delta^{N_d - n_d}$ and $\Delta = \{(1,0),(0,1),(0,0)\} \subset \mathbb{R}^2$. It holds that

$$\begin{aligned} \widehat{G}_d^{bp}(\theta) &= E_\theta[g_d(y_d)|y_{ds}] = \sum_{y_{dr}\in\mathcal{Y}_{dr}} g_d(y_{ds}, y_{dr})P(y_{dr}|y_{ds}) \\ &= \frac{\int_{\mathbb{R}^2} \sum_{y_{dr}\in\mathcal{Y}_{dr}} g_d(y_{ds}, y_{dr})P(y_{dr}|u_d)P(y_{ds}|u_d)f(u_d)\,du_d}{\int_{\mathbb{R}^2} P(y_{ds}|u_d)f(u_d)\,du_d}. \end{aligned}$$

The EBP of $G_d = g_d(y_d)$ is $\widehat{G}_d^{ebp} = \widehat{G}_d^{bp}(\widehat{\theta})$ and the EBP of $R_d = R_d(y_d)$ is $\widehat{R}_d^{ebp} = \widehat{R}_d^{bp}(\widehat{\theta})$.

They are approximated by the Monte Carlo method as follows.

1. Calculate the ML parameter estimator $\widehat{\theta} = (\widehat{\beta}', \widehat{\phi})'$.
2. For $s_1 = 1, \ldots, S_1$, generate $u_d^{(s_1)}$ i.i.d. $N(0, I_2)$ and set $u_d^{(S_1+s_1)} = -u_d^{(s_1)}$.

3. For $d = 1, \ldots, D$, $k = 1, 2$, $t = 1, \ldots, T$, $s_1 = 1, \ldots, 2S_1$, calculate

$$p_{dk,t}^{(s_1)} = \frac{\exp\{z_t\widehat{\beta}_k + \widehat{\phi}_k u_{dk}^{(s_1)}\}}{1 + \sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell + \widehat{\phi}_\ell u_{d\ell}^{(s_1)}\}}.$$

4. For $s_2 = 1, \ldots, S_2$, generate $y_{dj}^{(s_1s_2)} \sim M\big(1; p_{d1,t}^{(s_1)}, p_{d2,t}^{(s_1)}\big)$, $j \in r_{d,t}$, and construct $y_{d,tr}^{(s_1s_2)} = \underset{j\in r_{d,t}}{\text{col}}\big(y_{dj}^{(s_1s_2)}\big)$, $y_{d,t}^{(s_1s_2)} = (y_{d,ts}', y_{d,tr}^{(s_1s_2)\prime})'$, $y_{d}^{(s_1s_2)} = \underset{1\leqslant t\leqslant T}{\text{col}}\big(y_{d,t}^{(s_1s_2)}\big)$.

5. Calculate

$$\overline{Y}_{dk}^{(s_1s_2)} = \frac{1}{N_d}\sum_{t=1}^{T}\left\{\sum_{j\in s_{d,t}} y_{dkj} + \sum_{r_{d,t}} y_{dkj}^{(s_1s_2)}\right\}, \quad k = 1, 2.$$

6. Calculate

$$\widehat{G}_d^{ebp} = \frac{1}{2S_1S_2}\sum_{s_1=1}^{2S_1}\sum_{s_2=1}^{S_2} g_d\big(y_d^{(s_1s_2)}\big), \quad \widehat{R}_d^{ebp} = \frac{1}{2S_1S_2}\sum_{s_1=1}^{2S_1}\sum_{s_2=1}^{S_2}\frac{\overline{Y}_{d2}^{(s_1s_2)}}{\overline{Y}_{d1}^{(s_1s_2)} + \overline{Y}_{d2}^{(s_1s_2)}}.$$

The plug-in predictor of G_d is $\widehat{G}_d^{in} = g_d(\widehat{y}_d^{in})$. To calculate $\widehat{y}_d^{in}$, we recall that $\widehat{y}_{djk}^{in} = y_{djk}$ if $j \in s_{d,t}$ and $\widehat{y}_{djk}^{in} = \widehat{p}_{dk,t}^{in}$ if $j \in r_{d,t}$. Therefore, we take $\widehat{p}_{dk,t}^{in}$ from (4.18) and construct the vectors $\widehat{y}_d^{in} = (\widehat{y}_{ds}^{in\prime}, \widehat{y}_{dr}^{in\prime})'$, with $\widehat{y}_{ds}^{in} = y_{ds}$ and $\widehat{y}_{dr}^{in} = \underset{1\leqslant k\leqslant 2}{\text{col}}\Big(\underset{1\leqslant t\leqslant T}{\text{col}}\Big(\underset{j\in r_{d,t}}{\text{col}}\,(\widehat{p}_{dk,t}^{in})\Big)\Big)$.

The in.ebp predictor of G_d is $\widehat{G}_d^{in.ebp} = g_d(\widehat{y}_d^{ebp})$. The components of $\widehat{y}_d^{ebp}$ are $\widehat{y}_{djk}^{ebp} = y_{djk}$ if $j \in s_{d,t}$ and $\widehat{y}_{djk}^{ebp} = \widehat{p}_{dk,t}^{ebp}$ if $j \in r_{d,t}$. We take $\widehat{p}_{dk,t}^{ebp}$ from (4.17) and construct $\widehat{y}_d^{ebp} = (\widehat{y}_{ds}^{ebp\prime}, \widehat{y}_{dr}^{ebp\prime})'$, with $\widehat{y}_{ds}^{ebp} = y_{ds}$ and $\widehat{y}_{dr}^{ebp} = \underset{1\leqslant k\leqslant 2}{\text{col}}\Big(\underset{1\leqslant t\leqslant T}{\text{col}}\Big(\underset{j\in r_{d,t}}{\text{col}}\,(\widehat{p}_{dk,t}^{ebp})\Big)\Big)$.

For unemployment rates, the plug-in and in.ebp domain predictors are

$$\widehat{R}_d^{in} = \frac{\widehat{\overline{Y}}_{d2}^{in}}{\widehat{\overline{Y}}_{d1}^{in} + \widehat{\overline{Y}}_{d2}^{in}}, \quad \widehat{R}_d^{in.ebp} = \frac{\widehat{\overline{Y}}_{d2}^{ebp}}{\widehat{\overline{Y}}_{d1}^{ebp} + \widehat{\overline{Y}}_{d2}^{ebp}}, \tag{4.21}$$

where $\widehat{\overline{Y}}_{dk}^{in}$ and $\widehat{\overline{Y}}_{dk}^{ebp}$ are given in (4.19). The corresponding subdomain predictors are

$$\widehat{R}_{d,t}^{in} = \frac{\widehat{\overline{Y}}_{d2,t}^{in}}{\widehat{\overline{Y}}_{d1,t}^{in} + \widehat{\overline{Y}}_{d2,t}^{in}}, \quad \widehat{R}_{d,t}^{in.ebp} = \frac{\widehat{\overline{Y}}_{d2,t}^{ebp}}{\widehat{\overline{Y}}_{d1,t}^{ebp} + \widehat{\overline{Y}}_{d2,t}^{ebp}}, \tag{4.22}$$

where $\widehat{\overline{Y}}_{dk,t}^{in}$ and $\widehat{\overline{Y}}_{dk,t}^{ebp}$ are given in (4.20).

Due to the size of the population, the EBP and in.ebp predictors were problematic when applied to the SLFS2021.1 data in Section 4.6, providing misleading approximations of divisions of extremely large quantities, although they performed well in our simulation studies in Section 4.5. The cause of the problem is the size of the census, which corresponds to the Spanish population over the age of 16 in 2021. Consequently, the quantities defined in (4.17)

are excessively large, leading their division to problems of numerical instability. This undermines the prediction of the proportions in (4.19) and (4.20), and the problem is inherited to predictors (4.21) and (4.22). In fact, these problems prevent us from providing bootstrap estimates of the MSE for such predictors. For the plug-in predictor, the calculation of (4.18) does not cause any trouble. It can therefore be stated that it is not advisable to use the EBP (or in.ebp) to predict R_d in practice. Although they have proven to be promising predictors according to our simulation experiments in Section 4.5, only the plug-in predictor is finally used in the application to real data in Section 4.6.

4.4 Bootstrap inference

This section provides parametric bootstrap algorithms to estimate the MSE of the plug-in predictors under the categorical setup (4.14). Analogous algorithms can be obtained for the EBP and the in.ebp predictors in a straightforward manner.

Bootstrap estimation of the MSE of $\widehat{\overline{Y}}_{dk}^{in}$

The algorithm to estimate the MSE of the plug-in predictor $\widehat{\overline{Y}}_{dk}^{in}$ is as follows:

1. Fit the model and calculate the ML parameter estimator $\widehat{\theta} = (\widehat{\beta}', \widehat{\phi}')'$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Bootstrap sample: Generate $\{u_d^{*(b)} : d = 1, \ldots, D\}$ i.i.d. $N(0, I_2)$. The bootstrap sample has the same units as the real sample, i.e. $s_d^{*(b)} = s_d$. For $d = 1, \ldots, D$, $j \in s_d$, generate the elements of the bootstrap sample
   $$y_{dj}^{*(b)} \sim \mathrm{M}(1; p_{dj1}^{*(b)}, p_{dj2}^{*(b)}), \quad p_{djk}^{*(b)} = \frac{\exp\{x_{dj}\widehat{\beta}_k + \widehat{\phi}_k u_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{2} \exp\{x_{dj}\widehat{\beta}_\ell + \widehat{\phi}_k u_{d\ell}^{*(b)}\}}, \; k = 1, 2.$$
   (b) Bootstrap population quantities: For $d = 1, \ldots, D$, $k = 1, 2$, calculate
   $$\overline{Y}_{dk}^{*(b)} = \frac{1}{N_d}\Big(\sum_{j \in s_d} y_{djk}^{*(b)} + \sum_{t=1}^{T} Y_{dk,tr}^{*(b)} \Big),$$
   where $Y_{d,tr}^{*(b)} = \big(Y_{d1,tr}^{*(b)}, Y_{d2,tr}^{*(b)}\big) \sim \mathrm{M}(N_{d,tr}; p_{d1,t}^{*(b)}, p_{d2,t}^{*(b)})$ and $p_{dk,t}^{*(b)} = \frac{\exp\{z_t\widehat{\beta}_k + \widehat{\phi}_k u_{d_k}^{*(b)}\}}{1+\sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell + \widehat{\phi}_\ell u_{d\ell}^{*(b)}\}}$.
   (c) Bootstrap model: Fit a unit-level multinomial logit mixed model to the bootstrap sample $(y_{dj}^{*(b)}, x_{dj})$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$. Calculate the ML parameter estimator $\widehat{\theta}^{*(b)} = (\widehat{\beta}^{*(b)\prime}, \widehat{\phi}^{*(b)\prime})'$ and the random effects modal predictors $\widehat{u}_d^{*(b)}$, $d = 1, \ldots, D$. Calculate the plug-in predictor of $\overline{Y}_{dk}^{*(b)}$, i.e.
   $$\widehat{\overline{Y}}_{dk}^{in*(b)} = \frac{1}{N_d}\Big\{ \sum_{j \in s_d} y_{djk}^{*(b)} + \sum_{t=1}^{T} N_{d,tr}\widehat{p}_{dk,t}^{in*(b)} \Big\}, \; \widehat{p}_{dk,t}^{in*(b)} = \frac{\exp\{z_t\widehat{\beta}_k^{*(b)} + \widehat{\phi}_k^{*(b)}\widehat{u}_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell^{*(b)} + \widehat{\phi}_\ell^{*(b)}\widehat{u}_{d\ell}^{*(b)}\}}.$$

3. Output: $mse^*(\widehat{\overline{Y}}_{dk}^{in}) = \frac{1}{B}\sum_{b=1}^{B}\big(\widehat{\overline{Y}}_{dk}^{in*(b)} - \overline{Y}_{dk}^{*(b)}\big)^2$.

Bootstrap estimation of the MSE of $\widehat{G}_d^{in}$

The algorithm to estimate the MSE of the plug-in predictor $\widehat{G}_d^{in}$ is as follows:

1. Fit the model and calculate the ML parameter estimator $\widehat{\theta} = (\widehat{\beta}', \widehat{\phi}')'$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Bootstrap population: Generate $\{u_d^{*(b)} : d = 1, \ldots, D\}$ i.i.d. $N(0, I_2)$. The bootstrap sample has the same units as the real sample, i.e. $s_d^{*(b)} = s_d$.
      i. For $d = 1, \ldots, D$, $j \in s_d$, generate the elements of the bootstrap sample
$$y_{dj}^{*(b)} \sim \mathrm{M}(1; p_{dj1}^{*(b)}, p_{dj2}^{*(b)}), \quad p_{djk}^{*(b)} = \frac{\exp\{x_{dj}\widehat{\beta}_k + \widehat{\phi}_k u_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{2}\exp\{x_{dj}\widehat{\beta}_\ell + \widehat{\phi}_k u_{d\ell}^{*(b)}\}}, \ k = 1, 2.$$
      ii. For $d = 1, \ldots, D$, $t = 1, \ldots, T$, $j \in r_{d,t}$, $k = 1, 2$, generate the elements of the bootstrap non-sample subset, i.e.
$$y_{dj,t}^{*(b)} \sim \mathrm{M}(1; p_{d1,t}^{*(b)}, p_{d2,t}^{*(b)}), \quad p_{dk,t}^{*(b)} = \frac{\exp\{z_t\widehat{\beta}_k + \widehat{\phi}_k u_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{2}\exp\{z_t\widehat{\beta}_\ell + \widehat{\phi}_\ell u_{d\ell}^{*(b)}\}}.$$
      iii. For $d = 1, \ldots, D$, construct the bootstrap population vectors
$$y_d^{*(b)} = (y_{ds}^{*(b)\prime}, y_{dr}^{*(b)\prime})', \quad y_{ds}^{*(b)} = \operatorname*{col}_{j\in s_d}(y_{dj}^{*(b)}), \quad y_{dr}^{*(b)} = \operatorname*{col}_{1\leqslant t\leqslant T}\big(\operatorname*{col}_{j\in r_{d,t}}(y_{dj,t}^{*(b)})\big).$$
      iv. Calculate the bootstrap population quantities $G_d^{*(b)} = g_d(y_d^{*(b)})$, $d = 1, \ldots, D$.
   (b) Bootstrap model: Fit a unit-level multinomial logit mixed model to the bootstrap sample $(y_{dj}^{*(b)}, x_{dj})$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$. Calculate the ML parameter estimator $\widehat{\theta}^{*(b)} = (\widehat{\beta}^{*(b)\prime}, \widehat{\phi}^{*(b)\prime})'$ and the random effects modal predictors $\widehat{u}_d^{*(b)}$, $d = 1, \ldots, D$. Calculate the plug-in predictor of $G_d^{*(b)}$, i.e. $\widehat{G}_d^{in*(b)}$, $d = 1, \ldots, D$.
3. Output: $mse^*(\widehat{G}_d^{in}) = \frac{1}{B}\sum_{b=1}^{B}\big(\widehat{G}_d^{in*(b)} - G_d^{*(b)}\big)^2$, $d = 1, \ldots, D$.

4.5 Model-based simulations

This section presents the results of the model-based simulations. Simulation 1 compares the implemented fitting algorithms, i.e. the Laplace and the H-cubature algorithms from Section 4.2.1 and Section 4.2.2, respectively. Simulations 2.A and 2.B examine the behaviour of the EBP, in.ebp, and plug-in predictors from Section 4.3 and compare their performance with the predictors from Dawber et al. (2022), based on MQ and expectile regression for

multi-category outcomes. Simulation 3 tests the parametric bootstrap methods described in Section 4.4 and provides a recommendation on the number of bootstrap replicates to use in practice.

We generate unit-level data for model-based simulations to investigate the properties of the proposed statistical methodology. We may be interested in: (1) studying the effect of increasing the sample size or the number of domains (among other elements), or (2) analysing the behaviour of estimators and predictors in scenarios close to that of the application to real data. The first approach gives freedom in data generation and allows a larger number of questions to be investigated. The second approach is more restrictive, but allows to learn more about the application to real data. The size of the population (Spanish census) makes the second option practically unfeasible, due to the scale of some intermediate calculations (overflows), and the high computational time needed to construct the predictors and bootstrap resampling to estimate the MSE. For this reason, we chose the first approach, inspired by the model-based simulations performed by Hobza and Morales (2016) and Hobza et al. (2018) for the SAE methodology based on unit-level binomial mixed models.

Simulations 1, 2.A and 3 generate the vectors of auxiliary variables $x_{dj1} = (x_{dj11}, x_{dj12})$ and $x_{dj2} = (x_{dj21}, x_{dj22})$, $x_{dj11} = x_{dj21} = 1$, $x_{dj12} \sim \text{BI}(1, 1/2)$, $x_{dj22} \sim \text{BI}(1, 1/2)$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. In order to consider the predictors of Dawber et al. (2022), Simulation 2.B generates $x_{dj12} = x_{dj22}$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. These variables remain fixed during the simulations. For the multinomial logit mixed model, we take $q = 3$, $p_1 = p_2 = 2$, $p = 4$, $\beta_1 = (\beta_{11}, \beta_{12})' = (0.5, -1)'$, $\beta_2 = (\beta_{21}, \beta_{22})' = (-0.5, 0.5)'$, $\phi_1 = 0.4$, $\phi_2 = 0.5$. At each step i, we generate $u_{dk}^{(i)} \sim N(0, 1)$, $d = 1, \ldots, D$, $k = 1, 2$, and $y_{dj}^{(i)} \sim M(1; p_{dj1}^{(i)}, p_{dj2}^{(i)})$, where

$$p_{djk}^{(i)} = \frac{\exp\{\eta_{djk}^{(i)}\}}{1 + \exp\{\eta_{dj1}^{(i)}\} + \exp\{\eta_{dj2}^{(i)}\}}, \quad \eta_{djk}^{(i)} = x_{djk}\beta_k + \phi_k u_{dk}^{(i)}, \quad j = 1, \ldots, n_d.$$

4.5.1 Simulation 1

The target of Simulation 1 is to investigate the behaviour of the Laplace (LA) algorithm and the H-cubature (HC) algorithm from Section 4.2.1 and Section 4.2.2, respectively. To do so, we compare the empirical bias (BIAS) and root-MSE (RMSE) and consider two cases: (1) $n_d = 10$, $D = 25, 50, 75, 100$; (2) $D = 25$, $n_d = 10, 25, 50, 75, 100$.

An analysis of Tables 4.3 and 4.4 shows that, for both fitting algorithms, the error measures of the ML estimators decrease as D or n_d increases. In terms of execution times, the following average results are obtained. For $D = 25$ and $n_d = 10$, algorithms NB and LA need 2.16 and 0.07 minutes to converge, respectively. For $D = 50$ and $n_d = 10$, algorithms NB and LA need 4.27 and 0.06 minutes to converge, respectively. As a result, the computational cost of the Laplace algorithm is much lower. In fact, as the size of the data and the number of domains increases, the computation could become burdensome with unit-level data. For this reason, we use the Laplace algorithm in the remaining simulations and in the application to the SLFS2021.1 data.

Alg.	D	25	50	75	100	25	50	75	100
LA	β_{11}	-0.008	0.009	0.011	-0.006	0.228	0.164	0.127	0.112
	β_{12}	0.014	-0.009	-0.012	0.007	0.290	0.201	0.159	0.134
	β_{21}	-0.028	0.000	0.004	-0.019	0.284	0.186	0.147	0.125
	β_{22}	0.029	-0.003	-0.002	0.003	0.302	0.210	0.169	0.146
	ϕ_1	-0.092	-0.062	-0.034	-0.048	0.251	0.203	0.168	0.151
	ϕ_2	-0.122	-0.074	-0.067	-0.036	0.297	0.221	0.166	0.137
HC	β_{11}	-0.006	0.009	0.012	-0.007	0.226	0.164	0.126	0.111
	β_{12}	0.013	-0.010	-0.014	0.005	0.290	0.202	0.159	0.134
	β_{21}	-0.014	0.011	0.012	-0.009	0.273	0.182	0.143	0.120
	β_{22}	0.029	-0.002	-0.001	0.004	0.303	0.210	0.169	0.146
	ϕ_1	-0.080	-0.046	-0.016	-0.029	0.253	0.204	0.170	0.149
	ϕ_2	-0.115	-0.060	-0.052	-0.020	0.299	0.220	0.163	0.136

Table 4.3: Comparison of the Laplace (LA) and H-cubature (HC) fitting algorithms. BIAS (left) and RMSE (right) for Case (1): $n_d = 10$, $D = 25, 50, 75, 100$.

Alg.	n_d	10	25	50	75	100	10	25	50	75	100
LA	β_{11}	-0.008	-0.002	-0.007	0.005	0.003	0.228	0.147	0.117	0.109	0.105
	β_{12}	0.014	0.003	0.011	0.000	0.001	0.290	0.179	0.126	0.107	0.085
	β_{21}	-0.028	0.006	-0.033	-0.011	-0.017	0.284	0.181	0.150	0.142	0.135
	β_{22}	0.029	-0.010	0.009	-0.002	0.002	0.302	0.194	0.133	0.106	0.095
	ϕ_1	-0.092	-0.042	-0.026	-0.025	-0.014	0.251	0.167	0.116	0.098	0.092
	ϕ_2	-0.122	-0.045	-0.008	-0.013	-0.004	0.297	0.173	0.119	0.112	0.100
HC	β_{11}	-0.006	-0.001	-0.010	0.001	-0.002	0.226	0.145	0.115	0.106	0.102
	β_{12}	0.013	0.003	0.011	0.000	0.000	0.290	0.179	0.126	0.106	0.085
	β_{21}	-0.014	0.023	-0.013	0.007	0.001	0.273	0.176	0.142	0.136	0.128
	β_{22}	0.029	-0.010	0.009	-0.002	0.002	0.303	0.194	0.133	0.106	0.095
	ϕ_1	-0.079	-0.039	-0.028	-0.028	-0.018	0.253	0.165	0.114	0.096	0.088
	ϕ_2	-0.117	-0.044	-0.014	-0.022	-0.015	0.299	0.169	0.113	0.105	0.094

Table 4.4: Comparison of the Laplace (LA) and H-cubature (HC) fitting algorithms. BIAS (left) and RMSE (right) for Case (2): $D = 25$, $n_d = 10, 25, 50, 75, 100$.

4.5.2 Simulation 2

The target of Simulation 2.A is to test the behaviour of the predictors of $\overline{Y}_{dk}$, $k = 1, 2$, and R_d. It implements the EBP (ebp) and the plug-in predictor (in) of $\overline{Y}_{dk}$, and the EBP (ebp), plug-in EBP (in.ebp) and plug-in predictor (in) of R_d. Simulation 2.B compares the EBP and plug-in predictor with the robust MQ and expectile regression predictors for multi-category

data proposed by Dawber et al. (2022). Simulation 2.A generates the data in the same way as Simulation 1, allowing x_{dj12} to be different from x_{dj22}, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, and only computes the predictors based on the multinomial logit mixed model. Simulation 2.B imposes $x_{dj12} = x_{dj22}$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. We take $N_d = 200$, $d = 1, \ldots, D$. Each domain $U_d = \{u_{dj} : j = 1, \ldots, N_d\}$ is partitioned in two subsets, $U_{ds} = \{u_{dj} : j = 1, \ldots, n_d\}$ and $U_{dr} = \{u_{dj} : j = n_d + 1, \ldots, N_d\}$. Simulations 2.A and 2.B apply the Laplace algorithm to calculate the ML estimators of the multinomial logit mixed model.

Simulation 2.A

Table 4.5 shows the simulation results for $D = 25$ and $n_d = 10, 25, 50, 75, 100$. The performance measures are the average across domains of the absolute biases (AB) and root-MSEs (RE). Table 4.5 points out that both predictors $\widehat{\overline{Y}}_k^{ebp}$ and $\widehat{\overline{Y}}_k^{in}$ behave similarly, $k = 1, 2$. For unemployment rates, $\widehat{R}^{in.ebp}$ or $\widehat{R}^{in}$ perform slightly better than $\widehat{R}^{ebp}$ and have a lower computational cost. However, due to the computational intensity of the bootstrap resampling, we use the plug-in predictor in Simulation 3.

n_d	10	25	50	75	100	10	25	50	75	100
$\widehat{\overline{Y}}_1^{ebp}$	0.012	0.012	0.013	0.011	0.009	0.105	0.095	0.083	0.068	0.057
$\widehat{\overline{Y}}_1^{in}$	0.007	0.008	0.006	0.005	0.004	0.098	0.086	0.074	0.060	0.050
$\widehat{\overline{Y}}_2^{ebp}$	0.014	0.011	0.007	0.007	0.005	0.097	0.080	0.060	0.050	0.040
$\widehat{\overline{Y}}_2^{in}$	0.009	0.008	0.008	0.005	0.005	0.099	0.085	0.070	0.058	0.047
$\widehat{R}^{ebp}$	0.013	0.013	0.008	0.008	0.007	0.142	0.133	0.118	0.102	0.088
$\widehat{R}^{in.ebp}$	0.011	0.010	0.008	0.006	0.006	0.134	0.117	0.098	0.081	0.067
$\widehat{R}^{in}$	0.011	0.011	0.009	0.006	0.006	0.135	0.120	0.100	0.082	0.068

Table 4.5: Comparison of the predictors based on the multinomial logit mixed model. AB (left) and RE (right) for $D = 25$ and $n_d = 10, 25, 50, 75, 100$.

Simulation 2.B

First and foremost, MQ regression can be summarised as a quantile-type generalisation of regression based on influence functions. The most common influence function is the Huber function, which depends on a parameter c to be specified. By setting the value of c, it is possible to trade robustness for efficiency in MQ regression models. The modelling of binary outcomes in small areas using MQ regression has been proposed by Chambers et al. (2016). More recently, Dawber et al. (2022) extend this methodology to multi-category outcomes, but propose models with the same set of auxiliary variables in each category of the response variable. In contrast, we do not impose this restriction. Following Section 5 and Section S.2 of Appendix S1 of Dawber et al. (2022), Simulation 2.B fits a multi-category MQ model with $c = 1.345$, and a multi-category expectile (EXP) model with $c = 100$.

Table 4.6 shows the domain averages of the absolute biases (AB), root-MSEs (RE), relative absolute biases (RAB) and relative root-MSEs (RRE) for $D = 25$ and $n_d = 10$. As expected, the predictors based on the data-generating model, EBP and plug-in, show better performance. However, it should be noted that our model-based predictors are not robust. Therefore, the results of Simulation 2.B could be extrapolated to the real world if the selected multinomial logit mixed model fits the data properly and has a good diagnostic performance.

	$\widehat{\overline{Y}}_1^{ebp}$	$\widehat{\overline{Y}}_1^{in}$	$\widehat{\overline{Y}}_1^{mq}$	$\widehat{\overline{Y}}_1^{exp}$	$\widehat{\overline{Y}}_2^{ebp}$	$\widehat{\overline{Y}}_2^{in}$	$\widehat{\overline{Y}}_2^{mq}$	$\widehat{\overline{Y}}_2^{exp}$
AB	0.016	0.017	0.035	0.049	0.020	0.019	0.057	0.041
RE	0.051	0.054	0.099	1.000	0.053	0.055	0.085	0.086
RAB	4.812	5.081	10.601	14.818	6.102	5.818	11.292	12.492
RRE	15.334	16.242	30.065	30.175	16.092	16.781	17.050	25.943

(a) Performance of the predictors of employed and unemployed proportions.

	$\widehat{R}^{ebp}$	$\widehat{R}^{in.ebp}$	$\widehat{R}^{in}$	$\widehat{R}_2^{mq}$	$\widehat{R}_2^{exp}$
AB	0.023	0.029	0.029	0.058	0.052
RE	0.073	0.054	0.061	0.120	0.107
RAB	6.841	5.843	5.819	11.5128	10.305
RRE	22.201	10.742	12.163	24.021	21.435

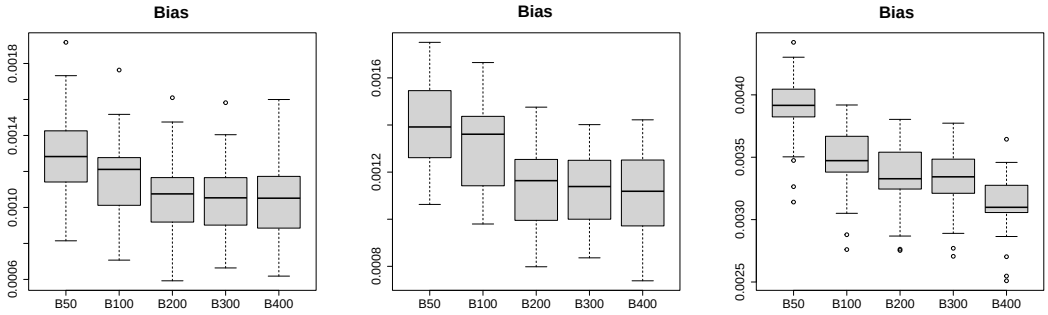
(b) Performance of the predictors of unemployment rates.

Table 4.6: Comparison of the predictors based on the multinomial logit mixed model and those proposed by Dawber et al. (2022). Absolute (top rows AB and RE) and relative (bottom rows RAB and RRE) performance measures for $n_d = 10$ and $D = 25$.

4.5.3 Simulation 3

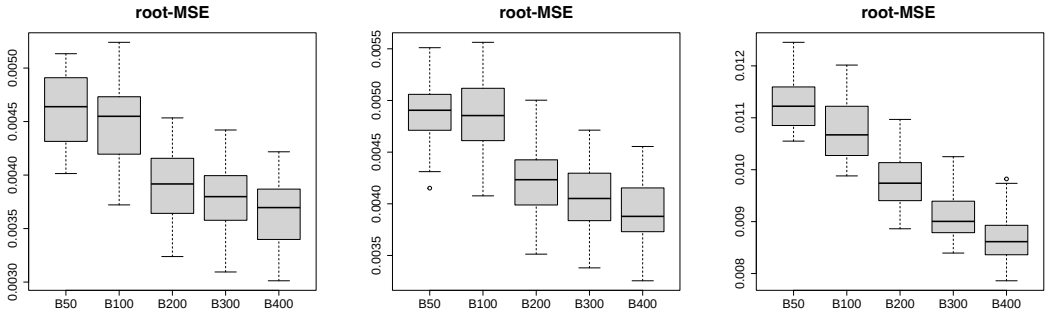
Simulation 3 calculates empirical biases (B_{dk}) and root-MSEs (RE_{dk}) for the parametric bootstrap estimator of the MSE of $\widehat{\overline{Y}}_{dk}^{in}$, $k = 1, 2$ and $\widehat{R}_d^{in}$ ($k = 0$). It takes the empirical MSEs, obtained from the output of Simulation 2.A, as *true* MSEs.

Figures 4.1 and 4.2 present boxplots of B_{dk}'s and RE_{dk}'s for $D = 25$, $n_d = 10$ and $B = 50, 100, 200, 400$. For $k = 1, 2$, biases take values between 0.0006 and 0.0018 and root-MSEs between 0.0030 and 0.0055. For $k = 0$, biases take values between 0.0025 and 0.0040 and root-MSEs between 0.008 and 0.012. In both cases, the main contribution to the root-MSE comes from the variance. Concerning the number of bootstrap replicates, both figures show that both biases and root-MSEs decreases as B increases. As a trade-off between computation time and precision, we recommend running the bootstrap algorithm with $B = 400$ replicates.



(a) Employed proportions. (b) Unemployed proportions. (c) Unemployment rates.

Figure 4.1: Study of the parametric bootstrap estimator of the MSE of $\widehat{\overline{Y}}_{d1}^{in}$ (left), $\widehat{\overline{Y}}_{d2}^{in}$ (center) and $\widehat{R}_d^{in}$ (right). Boxplots of B_{dk} for $k = 0, 1, 2$, $D = 25$, $n_d = 10$, $B = 50, 100, 200, 300, 400$.



(a) Employed proportions. (b) Unemployed proportions. (c) Unemployment rates.

Figure 4.2: Study of the parametric bootstrap estimator of the MSE of $\widehat{\overline{Y}}_{d1}^{in}$ (left), $\widehat{\overline{Y}}_{d2}^{in}$ (center) and $\widehat{R}_d^{in}$ (right). Boxplots of RE_{dk} for $k = 0, 1, 2$, $D = 25$, $n_d = 10$, $B = 50, 100, 200, 300, 400$.

4.6 Application to the 2021.1 SLFS data

4.6.1 Model fitting and validation

This section applies the developed methodology to the SLFS2021.1 data. We first fit the model (4.6)-(4.7) to the target data, with *age4* as the auxiliary variable and *age4-1* as the reference category. Table 4.7 shows the ML parameter estimators of the model parameters β_1, β_2, ϕ_1 and ϕ_2 of the multinomial mixed model (MMM), the p-values to test $H_0 : \beta_{kt} = 0$, $k = 1, 2$, $t = 1, 2, 3, 4$, and $H_0 : \phi_k = 0$, $k = 1, 2$, and the asymptotic and bootstrap CI at the 95% confidence level. It includes the lower (LB) and upper (UB) bounds. The table also includes the ML parameter estimators of the model parameters of the corresponding multinomial fixed effects model (MFM). In addition, a relative gap (Rgap) in % is included, and it is defined as the absolute difference between the MMM and the MFM estimates, divided by the MMM estimates and multiplied by 100. The purpose is to quantify the absolute

relative differences between the ML parameter estimators of the model parameters for the two multinomial models.

		β_{11}	β_{12}	β_{13}	β_{14}	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_1	ϕ_2
	MFM	0.808	0.365	-0.982	-4.441	-0.782	-0.006	-1.405	-5.938	–	–
	MMM	0.850	0.383	-0.995	-4.485	-0.824	0.002	-1.387	-5.910	0.347	0.224
	Rgap	4.913	4.715	1.288	0.993	5.138	442.583	1.298	0.472	–	–
	p-value	0.000	0.000	0.000	0.000	0.000	0.950	0.000	0.00	0.000	0.000
Asymp	LB 95%	0.830	0.345	-1.039	-4.556	-0.854	-0.056	-1.475	-6.225	0.332	0.196
	UB 95%	0.870	0.420	-0.951	-4.415	-0.795	0.059	-1.298	-5.595	0.367	0.252
Boot	LB 95%	0.822	0.348	-1.043	-4.558	-0.891	-0.058	-1.487	-6.231	0.298	0.179
	UB 95%	0.961	0.421	-0.952	-4.417	-0.786	0.059	-1.296	-5.626	0.402	0.263

Table 4.7: Model parameters of the unit-level multinomial logit mixed model for the SLFS2021.1 data.

An analysis of Table 4.7 shows that the Rgap is less than 6% except for β_{22}. If we exclude *age4-2* from the prediction of the disaggregated proportion of unemployed people, all p values for all coefficients are less than 0.05. Nonetheless, we treat *age4* as a factor, i.e. as a single categorical variable that can take a finite and fixed number of values. Therefore, the variable *age4* would be significant if any of its categories were, which supports its inclusion for both components of the target vector. The randomness of the two intercepts is also relevant. The standard deviation parameter estimates are $\widehat{\phi}_1 = 0.347$ and $\widehat{\phi}_2 = 0.224$. Neither the asymptotic nor the bootstrap CIs contain the zero, confirming the need to model the proportions with the random effects model. Consequently, we adopt the model presented in Table 4.7 and proceed with its validation.

As the fitted model is multi-BE, we perform the diagnosis at the subdomain level, i.e. at the intersections between domains and age groups. We are also interested in the reconciliation of the model-based and design-based approaches to SAE.

Under the categorical setup (4.14), we define the aggregated raw residuals (ARR) as

$$\widehat{e}_{dk,t} = \overline{y}_{dk,t} - \widehat{p}^{in}_{dk,t}, \quad \overline{y}_{dk,t} = \frac{1}{n_{d,t}} \sum_{j \in s_{d,t}} y_{djk}, \quad d = 1, \ldots, 104,\ t = 1, \ldots, 4,\ k = 1, 2.$$

The aggregated standardized residuals (ASR) are defined by dividing the ARRs by its standard deviation, given by $\nu_k = \left(\frac{1}{DT} \sum_{d=1}^{D} \sum_{t=1}^{T} \left(\widehat{e}_{dk,t} - \widehat{\overline{e}}_{.k,.} \right)^2 \right)^{\frac{1}{2}}$, $\widehat{\overline{e}}_{.k,.} = \frac{1}{DT} \sum_{d=1}^{D} \sum_{t=1}^{T} \widehat{e}_{dk,t}$, $k = 1, 2$. Accordingly, the ASRs are defined as $\widehat{e}_{dk,t} {\nu_k}^{-1}$, $d = 1, \ldots, 104$, $t = 1, \ldots, 4$, $k = 1, 2$.

Figure 4.3 plots the ASRs for employed (left) and unemployed (right) proportions, sorted by subdomain sample size. Unsurprisingly, their magnitude decreases progressively as the sample size increases, resulting in a conical structure in the line charts. All values oscillate symmetrically around $y = 0$ in a reasonable range for an outlier analysis. Out of a total of 416 subdomains, there are only 11 (2.64%) outside the interval $(-3, 3)$, both for the employed

and the unemployed categories. In conclusion, the proposed model performs satisfactorily in terms of the significance level of the model parameters and the validation via the ASRs.

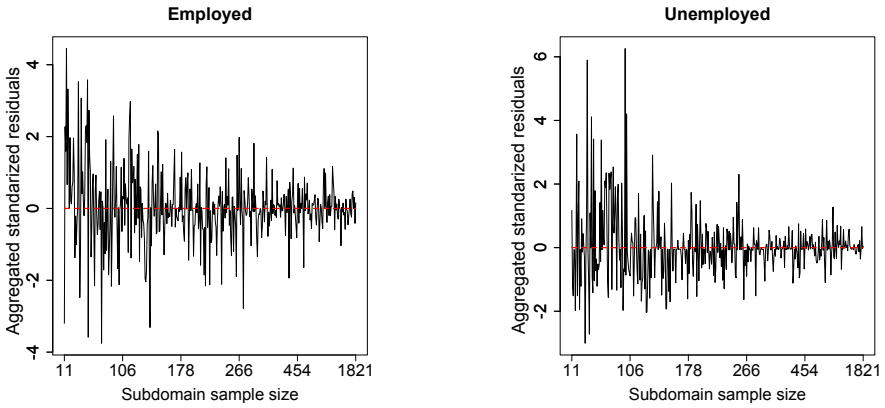


Figure 4.3: ASRs for employed (left) and unemployed (right) proportions by subdomain sample size for the SLFS2021.1 data.

4.6.2 Prediction, error measures and maps

This section presents Hájek estimates and plug-in predictions of unemployment rates by province, sex and age group. A completely parallel template could be followed to provide plug-in predictions and Hájek estimates of the proportions of employed, unemployed and inactive people. However, we focus on unemployment rates because they are more challenging to predict and more sought after by governments and private institutions.

In Figure 4.4 (left), the freedom of the non-parametric Nadaraya-Watson regression confirms the linear relationship between direct and model-based estimates. This highlights a key advantage of our approach: the theoretical properties of the Hájek estimator, such as asymptotic design-based unbiasedness, are, to some extent, inherited by the plug-in predictor. Figure 4.4 (right) plots Hájek estimates and plug-in predictions of unemployment rates against the subdomain index. The 416 subdomains are first sorted by age group, secondly by sex and thirdly by province. In this plot, we can see some smoothing effect of the plug-in predictor compared to the Hájek estimates. Especially in the fourth age group, which consists of people aged 65 and over. The latter is due to the imprecision of the Hájek estimates in these subdomains, where the number of respondents is quite small. It should be remembered that we are estimating a non-linear quantity ($R_{d,4}$, $d = 1, \ldots, D$) that depends directly on the total number of employed and unemployed people. It is therefore essential to provide measures of accuracy.

As error measures, we calculate the parametric bootstrap estimator of the MSE of $\widehat{R}^{in}_{d,t}$, $d = 1, \ldots, D$, $t = 1, \ldots, 4$. Section 4.4 describes the estimation procedure, applied with $B = 500$ bootstrap resamples. In terms of application, the MSE assesses the quality of a predictor, but it is scale-dependent, in the same way as the RMSE. However, the RRMSE is used to compare different predictors by expressing the error in relative or percentage terms. This is why we use RRMSEs and CVs to compare the performance of model-based predic-

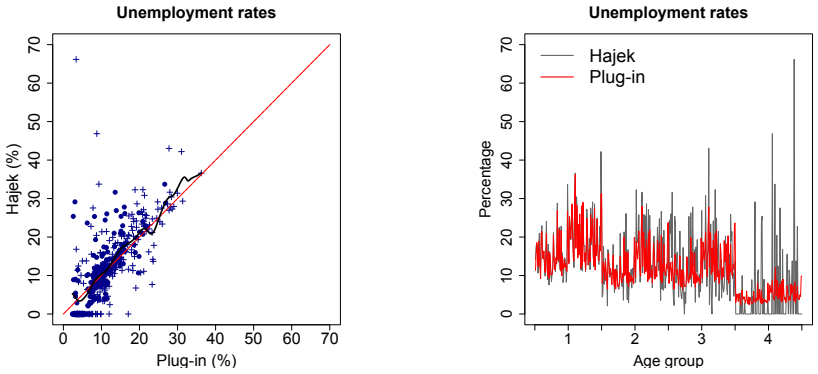


Figure 4.4: On the left, Hájek vs plug-in predicted subdomain unemployment rates. Men are dots and women crosses. On the right, line chart sorted by age group and domain.

tors and direct estimators. Table 4.8 shows the deciles of the model-based estimates of the RRMSEs (in %) for the plug-in predictor. They are calculated from the 416 subdomain-level estimates of the RRMSEs. This table also includes the design-based CVs (in %) of the Hájek estimator, assuming unbiasedness (see Morales et al. (2021), Chapter 3). For both the plug-in predictor and the Hájek estimator, the denominators of the estimated RRMSEs and CVs are the corresponding plug-in predictions and direct estimates, respectively.

Table 4.8 shows that the plug-in predictor is superior to the Hájek estimator. In fact, it follows that there is a significant reduction in all the estimated RRMSEs when using the proposed model and, in particular, the plug-in predictor. For the sake of completeness, the results are broken down by group in Table 4.9. This is done by calculating the quartiles for each age group by sorting the 104 crosses between province and sex.

	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
$\widehat{R}^{in}_{d,t}$	3.568	6.150	7.668	9.048	10.467	12.286	14.236	16.451	20.097	25.573	49.891
$\widehat{R}^{dir}_{d,t}$	6.222	11.529	14.579	18.101	21.828	26.631	32.699	39.152	52.865	71.979	113.012

Table 4.8: Deciles of subdomain-level RRMSEs and CVs (in %) of unemployment rates for the SLFS2021.1 data.

q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	q_1
3.568	7.138	9.098	13.256	22.474
3.753	7.641	10.018	14.662	25.201
4.360	8.112	10.805	15.313	25.232
8.250	17.558	23.136	30.047	49.891

(a) Plug-in predictor: $\widehat{R}^{in}_{d,k}$

q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	q_1
6.222	10.082	14.130	18.151	30.133
10.800	18.188	23.193	35.036	70.105
19.984	32.664	44.353	58.567	99.170
41.421	73.160	93.172	96.183	113.012

(b) Hájek estimator: $\widehat{R}^{dir}_{d,k}$

Table 4.9: Quartiles of subdomain-level RRMSEs and CVs (in %) of unemployment rates by age group for the SLFS2021.1 data.

Last but not least, RRMSE values below 30% are expected in SAE, as is the case of the

plug-in predictor derived from our model for *age4-1*, *age4-2* and *age4-3*. In the fourth age group, the over-65s, predicting unemployment rates is quite difficult because there is almost no labour force data. The values skyrocket for the direct estimator.

The model offers the opportunity to analytically read the appreciable differences by Spanish provinces. Figures 4.5-4.6 present maps for unemployment rates (in %), showing how they differ by province and age group and, in particular, the differences between men (Figure 4.5) and women (Figure 4.6). The fourth age group is not included for either sex because unemployment rates are below 10% in all provinces except in the domain of women living in Cádiz, with an unemployment rate close to 12%. In spite of the variety of measures that have been put in place to reduce gender inequality, the gap is still wide in terms of unemployment. Among the most notable results, there is a significant difference between sexes, with a clearly higher proportion of unemployed women for all age groups.

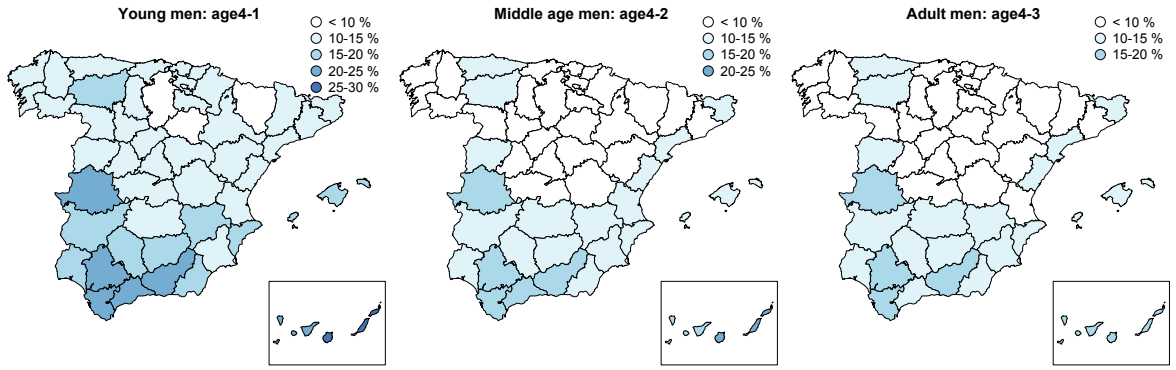


Figure 4.5: Unemployment rates for men in SLFS2021.1.

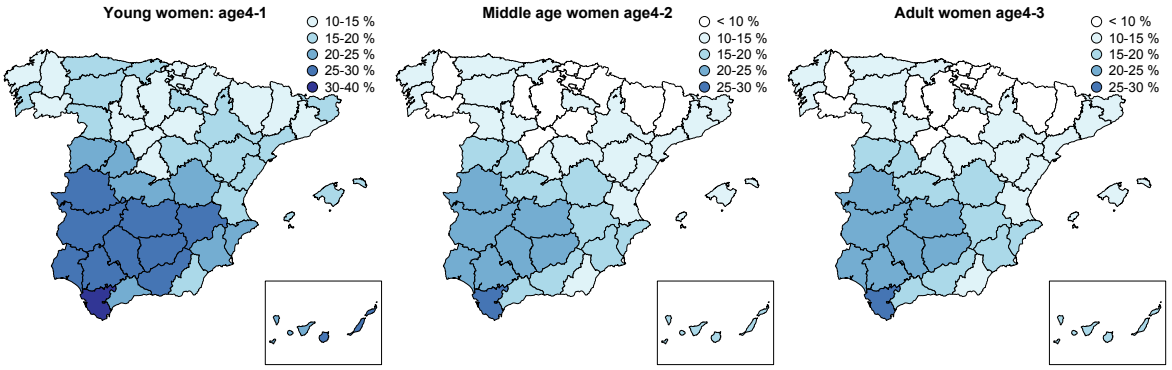


Figure 4.6: Unemployment rates for women in SLFS2021.1.

The highest unemployment rates are found in the center and southwest of the country, with lower percentages in the north and in the eastern Mediterranean Coast. What is more,

the south-west part of Spain suffers the worst unemployment situation, and it is even worse for women. Moreover, the distribution between neighbouring provinces, and among those with similar demographic and socio-economic conditions, is generally homogeneous. Santa Cruz de Tenerife and Cádiz are the provinces most affected by unemployment, followed by the other provinces of Andalucía and Extremadura.

Figures 4.7-4.8 map the RRMSE estimates of the plug-in predictions of the unemployment rates for men (Figure 4.7) and women (Figure 4.8), respectively, by age group, from left to right. According to Section 4.6.2, $B = 500$ bootstrap resamples are used. It can be observed that RRMSEs are lower in the center and south of the Iberian Peninsula, and more so for women. The same applies to the Canary and Balearic Islands. The highest estimated relative errors are reached in the north-east, although not exceeding 30% for men and 22% for women. The results are more than acceptable considering that we are predicting a non-linear indicator, such as the unemployment rate, in small and unplanned areas in the survey.

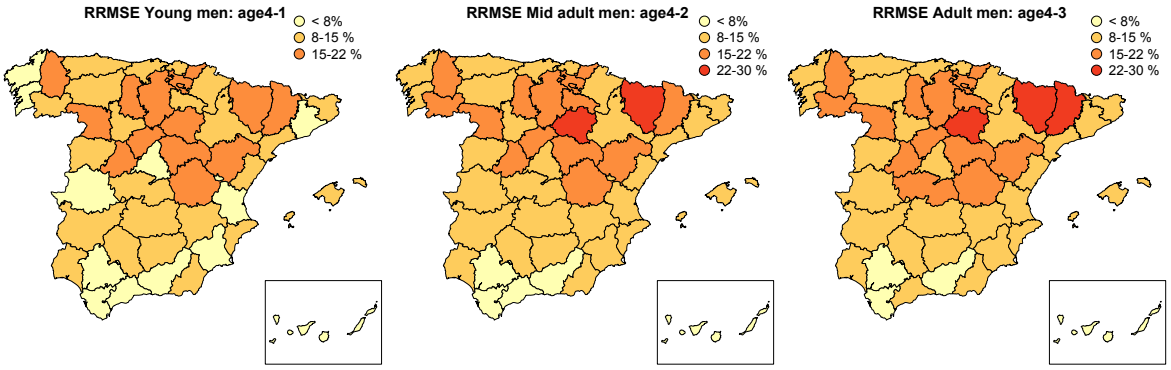


Figure 4.7: RRMSEs of unemployment rates for men in SLFS2021.1.

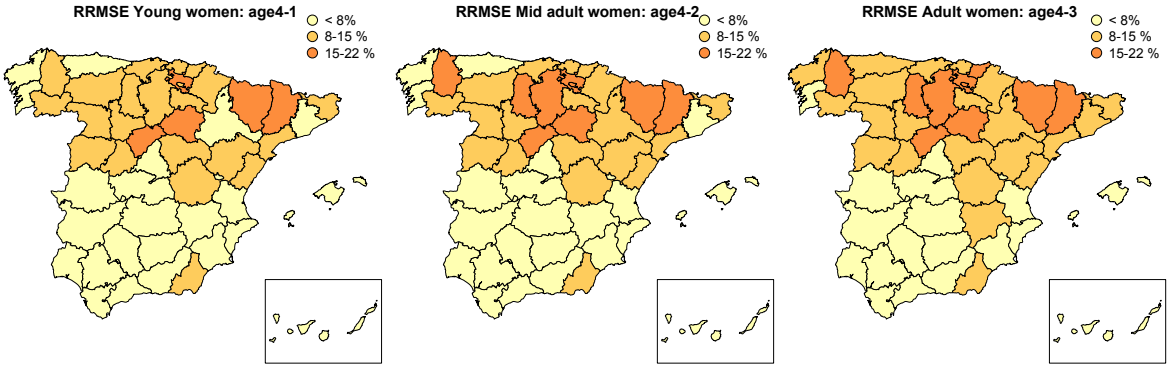


Figure 4.8: RRMSEs of unemployment rates for women in SLFS2021.1.

4.7 R codes

As for the `R` codes, the GitHub repository https://github.com/small-area-estimation/Small-area-estimation-of-labour-force-indicators-under-unit-level-multinomial-mixed-models (accessed on: November 4, 2024) contains our dataset and computer code, as well as a detailed description of its contents. It includes a README file that provides basic instructions for the correct execution of the available software.

Chapter 5

M-quantile regression

> Adapted from Koenker (2005)[page 294], "*Much of the early history of social statistics can be viewed as a search for the average man, that improbable man without qualities who could be confortable with his feet in the ice chest and his hands in the oven (...). Yet for all the mathematical elegance of the Gaussian law of errors, it should be tempered by a skeptical empiricism: a willingness to peer occasionally outside the cathedral of mathematics and see the world in all its diversity*".

Models for the conditional mean, with i.i.d. normal errors, are welcome approximations in many applications to real data. However, they can also be risky strategies. The strengths and weaknesses of these models depend on the fulfillment of their strong parametric assumptions. This is compounded by the oversimplification of area-level models to explain population patterns (Rao and Molina, 2015). Fortunately, unit-level models pose many advantages relative to area-level models, especially after the breakthrough of the MQ modelling approach to SAE (Chambers and Tzavidis, 2006). Indeed, MQ regression is considered a valuable alternative in SAE to relax some of the conventional assumptions of LMMs and obtain estimators that are robust against outliers. Another advantage of the MQ models is the computational speed of their fitting, thanks to the IRLS algorithm (Bianchi and Salvati, 2015).

This chapter contains two novel contributions: the extension of MQ models to the semi-parametric modelling of temporal dependencies and the pioneering proposal of data-driven criteria for the selection of robustness parameters. First and foremost, it is a smart strategy to rely on data measured over time. In this respect, LMMs are unable to properly capture time dependencies when the number of lags is somewhat large. Since there are no published studies dealing with robust prediction in small areas based on time-dependent data, it is sought to extend the MQ regression to this field of research, adding flexibility to the widely imposed assumption of unit-level independence. In light of the above, the final objective of this thesis is to propose temporal MQ models and then, derive robust bias-corrected predictors of small area linear indicators. As for the estimation of the MSE, we have obtained, under general conditions, a first-order approximation and proposed several analytical estimators.

Apart from all the above, we have defined an optimal criterion for an accurate selection of the robustness parameters for bias correction in MQ models (see Section 5.4.4). The idea is to

reduce the bias of the robust, model-based predictors but not unbalance the MSE. It can be applied to MQ models in general, not only to the predictors derived from the TWMQ linear models. Additionally, its potential role in outlier detection has been extensively studied, both in simulation experiments and in the application to real data.

This chapter is structured as follows. Section 5.1 introduces the MQ functions. Sections 5.2 and 5.3 review the theory of two-fold MQ (MQ2) linear models (Chambers and Tzavidis, 2006) and three-fold MQ (MQ3) linear models (Marchetti et al., 2018) for SAE, focusing on the most relevant aspects to our research. Section 5.4 describes the TWMQ statistical methodology. Sections 5.5 presents the results of the model-based simulations. In Sections 5.6 and 5.7, the new methods are illustrated with an application to socio-economic data, modelling the average level of income in 23 provinces of Empty Spain. A section for `R` codes is not included because they are not yet available in any online repository. This is a project in progress, involving both debugging and documentation of the code itself. This chapter is accompanied by two appendices. Appendix C describes an adaptation of the IRLS algorithm used to estimate the model parameters of the TWMQ linear models. Appendix D provides technical specifications and step-by-step proofs of Theorems 1 and 2 in Section 5.4.

5.1 M-quantile functions

An excellent account given by Koenker (2005) is a must in the literature on robust statistics. We present here only a brief review of the basic concepts necessary to introduce the contributions derived from our research on temporal MQ models for SAE.

Let Y be a random variable with cumulative distribution function (c.d.f.) $F_Y(y) = P(Y \leqslant y)$, $y \in \mathbb{R}$, and standard deviation $\sigma_Y > 0$.
For $0 < q < 1$, let us define the (q, σ_q, ψ)-check function

$$\rho_q(u, \sigma_q) = 2\sigma_q \left| q - I_{(-\infty,0)}(u) \right| \rho(\sigma_q^{-1} u), \quad u \in \mathbb{R},\ \sigma_q > 0, \tag{5.1}$$

where $\rho(u)$ is a continuously differentiable loss function and $\psi(u) = \frac{\partial \rho(u)}{\partial u}$.

For $0 < q < 1$, the partial derivative of $\rho_q(u, \sigma_q)$, with respect to u, is

$$\psi_q(u, \sigma_q) = \frac{\partial \rho_q(u, \sigma_q)}{\partial u} = 2\Big\{ q I_{(0,\infty)}(u) + (1-q) I_{(-\infty,0]}(u) \Big\}\, \psi(\sigma_q^{-1} u). \tag{5.2}$$

The quantile function of order q, scale parameter σ_q and influence function ψ, of Y is

$$Q_q(Y; \sigma_q, \psi) = \underset{\xi_q \in \mathbb{R}}{\operatorname{argmin}}\, E\big[\rho_q(Y - \xi_q, \sigma_q)\big] = \underset{\xi_q \in \mathbb{R}}{\operatorname{argmin}} \int_{\mathbb{R}} \rho_q(y - \xi_q, \sigma_q)\, dF_Y(y), \quad 0 < q < 1,$$

and it is called MQ function of order q of Y.

For $0 < q < 1$, $u = y - \xi_q \in \mathbb{R}$, it holds that

$$\rho_q(y - \xi_q, \sigma_q) \;=\; \sigma_q q\, \rho\big((\sigma_q^{-1}(y - \xi_q)\big) I_{(0,\infty)}(y - \xi_q) + \sigma_q (1-q)\, \rho\big((\sigma_q^{-1}(y - \xi_q)\big) I_{(-\infty,0)}(y - \xi_q),$$

and

$$\begin{aligned}\frac{\partial \rho_q(y-\xi_q,\sigma_q)}{\partial \xi_q} &= -q\psi\big((\sigma_q^{-1}(y-\xi_q)\big)I_{(0,\infty)}(y-\xi_q) - (1-q)\,\psi\big((\sigma_q^{-1}(y-\xi_q)\big)I_{(-\infty,0)}(y-\xi_q)\\ &= -\big\{qI_{(0,\infty)}(y-\xi_q) + (1-q)I_{(-\infty,0)}(y-\xi_q)\big\}\,\psi\big((\sigma_q^{-1}(y-\xi_q)\big)\\ &= -\psi_q(y-\xi_q,\sigma_q).\end{aligned}$$

Therefore, the MQ function of order q of Y is calculated as

$$Q_q(Y;\sigma_q,\psi) = \underset{\xi_q\in\mathbb{R}}{\text{solution}}\left\{\int_{\mathbb{R}} \psi_q(y-\xi_q,\sigma_q)\,dF_Y(y) = 0\right\}, \quad 0<q<1.$$

Finally, in order to complete the definition of $\psi_q(u,\sigma_q)$ in (5.2), we take $\sigma_q = \sigma_Y$ for $0<q<1$, and we use the Huber function

$$\psi(u) = uI_{(-c_\psi,c_\psi)}(u) + c_\psi\,\text{sgn}(u)I_{(-\infty,-c_\psi]\bigcup[c_\psi,\infty)}(u),\ u\in\mathbb{R};\quad c_\psi>0. \tag{5.3}$$

MQ regression uses bounded influence functions that reduce the influence of outlier observations (Huber, 1981). A widely accepted choice for the influence function ψ is the Huber function, defined in (5.3), although other options are also possible. Note that its choice is of little importance for the calculation of small area estimates and does not merit special attention (Chambers and Tzavidis, 2006). In any case, it is assumed that ψ depends on a predetermined tuning constant $c_\psi \geqslant 0$. It is customary to set $c_\psi = 1.345$ because it guarantees 95% efficiency when the errors are normal, and still offers protection against outliers (Holland and Welsch, 1977). Huber functions are presumably used here.

To extend the MQ functions to regression models, the argument inside the (q,σ_q,ψ)-check function is replaced by standardized residuals, as we will explained below.

5.2 Two-fold M-quantile linear regression for SAE

GLMMs and LMMs are the simpler and most commonly used statistical models for SAE. However, in the presence of atypical data or violation of the strong parametric assumptions of the previous models (normality of errors and random effects, among others), they can lead to incorrect modelling. Fortunately, the semi-parametric approach of Chambers and Tzavidis (2006) to SAE considerably reduces the necessary assumptions and allows to capture the variability between areas, modelled with random effects in GLMMs and LMMs, by fitting a different MQ regression surface for each area. It is also a robust option against atypical unit-level and area-level data. For ease of exposition, we will explain their approach step-by-step and finally generalise it to temporal data.

Let U be a finite population of size N hierarchically partitioned in domains U_d of sizes N_d, $d=1,\ldots,D$. Let s and s_d be the corresponding sampled subsets of sizes n and n_d, respectively. For $0<q<1$, the two-level M-quantile linear regression (MQ2) models are

$$y_{dj} = \boldsymbol{x}'_{dj}\boldsymbol{\beta}_\psi(q) + e_{\psi,dj}(q), \quad d=1,\ldots,D,\ j=1,\ldots,N_d, \tag{5.4}$$

where $\boldsymbol{x}'_{dj} = (x_{dj1}, \ldots, x_{djp})$, $\boldsymbol{\beta}_\psi(q) = (\beta_{\psi 1}(q), \ldots, \beta_{\psi p}(q))'$, $p \geqslant 1$, and $e_{\psi,dj}(q)$ are independent model errors with unknown c.d.f. $F_q(u) = P(e_{\psi,dj}(q) \leqslant u|\boldsymbol{x}_{dj})$, $u \in \mathbb{R}$. In addition they satisfy, by definition, that $Q_q(e_{\psi,dj}(q); \sigma_q, \psi|\boldsymbol{x}_{dj}) = 0$ and, although no explicit parametric assumptions are being made, the homoscedasticity assumption $\sigma_q = \text{var}^{1/2}(e_{\psi,dj}(q)) = \sigma_\psi(\boldsymbol{\beta}_\psi(q))$ is imposed. One of the advantages of models (5.4) over LMMs is that we have not specified a formal structure for the model errors, let alone required them to follow a normal distribution. Without going into further detail, the MQ2 linear models are generalised to three levels of hierarchy in Section 5.3.

In practice, $\widehat{\boldsymbol{\beta}}_\psi(q)$ and $\widehat{\sigma}_q$ are estimated using the IRLS algorithm, which ensures convergence to a unique solution (Bianchi and Salvati, 2015). This algorithm is explained step-by-step in Appendix C for estimating the regression parameters of the TWMQ linear models, and its adaptation to the MQ2 linear models is straightforward.

5.2.1 Two-fold M-quantile approach for inter-area variability

The MQ2 linear models have been used to model inter-area variability. The idea is to non-parametrically capture the variability of the population, beyond what is explained by the auxiliary variables, using the so-called MQ coefficients (Breckling and Chambers, 1988). This approach avoids distributional assumptions, as well as problems associated with the specification of the random effects in LMMs, allowing differences between areas to be characterised by the variation in area-specific MQ coefficients. It is therefore more robust to atypical data than the inclusion of random effects, as quantiles and MQs are more robust than the mean of a random variable (Koenker, 2005).

For $j = 1, \ldots, N_d$, the unit-level MQ coefficients of models (5.4) are

$$q_{dj} = \underset{0<q<1}{\text{solution}} \left\{ Q_q(y_{dj}; \sigma_q, \psi|\boldsymbol{x}_{dj}) = y_{dj} \right\}, \; Q_q(y_{dj}; \sigma_q, \psi|\boldsymbol{x}_{dj}) = \boldsymbol{x}'_{dj}\boldsymbol{\beta}_\psi(q).$$

For each target variable y_{dj}, $j = 1, \ldots, N_d$, we calculate the quantile q_{dj} for which the model error $e_{dj}(q_{dj})$ would be equal to zero if the β-coefficients were known.

By definition, it holds that

$$y_{dj} = Q_{q_{dj}}(y_{dj}; \sigma_{q_{dj}}, \psi|\boldsymbol{x}_{dj}) = \boldsymbol{x}'_{dj}\boldsymbol{\beta}_\psi(q_{dj}), \quad \sigma_{q_{dj}} = \sigma_\psi(\boldsymbol{\beta}_\psi(q_{dj})).$$

The unit-level MQ coefficient q_{dj} is the "most likely" quantile of unit j of area d. That is, of all the MQ2 linear models that vary by $0 < q < 1$, the model with $q = q_{dj}$ would predict y_{dtj} without error if $\boldsymbol{\beta}_\psi(q)$ was known. Since the target variables y_{dj}, $j = 1, \ldots, n_d$, are observed for all units in the sampled subsets s_d, the unit-level MQ coefficients can be estimated in these units. More specifically, an estimator of q_{dj}, $j = 1, \ldots, n_d$, is

$$\widehat{q}_{dj} = \underset{0<q<1}{\text{solution}} \left\{ \widehat{Q}_q(y_{dj}; \widehat{\sigma}_q, \psi|\boldsymbol{x}_{dj}) = y_{dtj} \right\}, \quad \widehat{Q}_q(y_{dj}; \widehat{\sigma}_q, \psi|\boldsymbol{x}_{dj}) = \boldsymbol{x}'_{dj}\widehat{\boldsymbol{\beta}}_\psi(q).$$

The idea is to use the unit-level MQ coefficients of each individual to then estimate the average in each area and fit the MQ2 linear models to that q-value. In LMMs we include random

effects to capture the variability between areas and with the MQ approach to SAE we capture this variability by fitting different MQ2 linear regression surfaces for each area.

The domain population and sample means of unit-level MQ coefficients are

$$\theta_d \triangleq \bar{q}_{d.} = \frac{1}{N_d}\sum_{j=1}^{N_d} q_{dj}, \quad \widehat{\theta}_d \triangleq \widehat{\bar{q}}_{d.} = \frac{1}{n_d}\sum_{j=1}^{n_d} \widehat{q}_{dj}, \quad d = 1, \ldots, D,$$

respectively. By definition, $\widehat{\theta}_d$ is the area-level average of the estimated unit-level MQ coefficients of the units in the sample. Therefore, we expect that the MQ2 linear model with $q = \widehat{\theta}_d$ will be the one that provides the best predictions in the domain d. Overall, if we are predicting linear domain parameters, such as population means. That is to say, in order to predict y_{dj} in the unobserved part of domain d, we will choose the MQ2 linear model with $q = \widehat{\theta}_d$. Thus, prediction in MQ2 linear models for SAE involves fitting D models, each one with the most appropriate quantile $q = \widehat{\theta}_d$ for the domain d it represents.

5.2.2 Robust predictors for two-fold M-quantile models

The MQ2 linear models are used to predict domain quantities including, but not only, population means. Specifically, we calculate predictors of addictive quantities $G_d = g_d(y_{d1}, \ldots, y_{dN_d})$, where $g_d : \mathbb{R}^{N_d} \to \mathbb{R}$ is a continuous function. Therefore, its applicability overlaps with the EBP methodology based on NER models of Rao and Molina (2010). For example, we can predict poverty rates and poverty gaps, but this research focuses on predicting linear domain-dependent population values and, in particular, population means:

$$\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}, \ d = 1, \ldots, D.$$

As the first option, a plug-in type predictor is calculated for the population means. Under regularity assumptions, a Taylor series expansion of $\boldsymbol{\beta}_\psi(\theta_d)$ around $\widehat{\theta}_d$ yields to

$$\boldsymbol{\beta}_\psi(\theta_d) \approx \boldsymbol{\beta}_\psi\big(\widehat{\theta}_d\big) + \frac{\partial \boldsymbol{\beta}_\psi(q)}{\partial q}\Big|_{q=\widehat{\theta}_d}(\theta_d - \widehat{\theta}_d), \quad d = 1, \ldots, D.$$

Let $r_d = U_d - s_d$ be the non sampled subset of U_d, $d = 1, \ldots, D$. If we assume that the sum of the residuals $e_{\psi,dj}(\theta_d)$ in r_d is close to zero, then it holds that

$$\begin{aligned}
\overline{Y}_d &= \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj} = \frac{1}{N_d}\Big\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \boldsymbol{x}'_{dj}\boldsymbol{\beta}_\psi(\theta_d) + \sum_{j\in r_d} e_{\psi,dj}(\theta_d)\Big\} \\
&\approx \frac{1}{N_d}\Big\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \boldsymbol{x}'_{dt}\boldsymbol{\beta}_\psi\big(\widehat{\theta}_d\big)\Big\} + \frac{1}{N_d}\sum_{j\in r_d} \boldsymbol{x}'_{dj}\frac{\partial \boldsymbol{\beta}_\psi(q)}{\partial q}\Big|_{q=\widehat{\theta}_d}(\theta_d - \widehat{\theta}_d).
\end{aligned}$$

Typically, the second summand of the last expression is much smaller than the first one. Therefore, we define the MQ predictor of $\overline{Y}_d$ as

$$\widehat{\overline{Y}}_d^{mq} = \frac{1}{N_d}\Big\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \boldsymbol{x}'_{dj}\widehat{\boldsymbol{\beta}}_\psi\big(\widehat{\theta}_d\big)\Big\}. \tag{5.5}$$

Given that $Q_{\theta_d}(e_{\psi,dj}(\theta_d);\sigma_{\theta_d},\psi|\boldsymbol{x}_{dj})=0$, it follows that

$$\sum_{j\in r_d} e_{\psi,dj}(\theta_d)\approx 0 \quad \text{if} \quad \theta_d\in(1/2-\varepsilon,1/2+\varepsilon),$$

for some small $\varepsilon>0$, but not otherwise. It is therefore desirable to know the sign of the bias of (5.5) and reduce its magnitude, as the MQ predictor is expected to be a biased predictor of the population mean. If $\theta_d<1/2-\varepsilon$, the event $\sum_{j\in r_d} e_{\psi,dj}(\theta_d)>0$ will occur with greater probability than the opposite event and the MQ predictor will tend to have negative bias. If $\theta_d>1/2-\varepsilon$, however, the MQ predictor will tend to have positive bias.

We define the residuals and standardized residuals of the MQ2 linear models as

$$\widehat{e}_{\psi,dj}(\widehat{\theta}_d)=y_{dj}-\boldsymbol{x}'_{dj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d),\quad \widehat{u}_{\psi,dj}(\widehat{\theta}_d)=\widehat{\sigma}^{-1}_{\widehat{\theta}_d}\widehat{e}_{\psi,dj},\quad j=1,\ldots,n_d.$$

The previous calculations involve the estimation of both the $\widehat{\boldsymbol{\beta}}_\psi$-coefficients and the area-level M-quantile coefficients $\widehat{\theta}_d$. Provided

$$\frac{1}{n_d}\sum_{j\in s_d}\boldsymbol{x}'_{dj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d)\approx\frac{1}{N_d}\sum_{j\in U_d}\boldsymbol{x}'_{dj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d),$$

the bias of (5.5) is estimated as

$$\widehat{B}\big(\widehat{\overline{Y}}^{mq}_d\big)=\widehat{E}\big[\widehat{\overline{Y}}^{mq}_d-\overline{Y}_d\big]=-\frac{1}{n_d}\Big(1-\frac{n_d}{N_d}\Big)\sum_{j\in s_d}\widehat{e}_{\psi,dj}(\widehat{\theta}_d).$$

Based on a robustification of $\widehat{e}_{\psi,dj}(\widehat{\theta}_d)$, the robust bias-corrected (BMQ) predictor of (5.5) is

$$\widehat{\overline{Y}}^{bmq}_d=\widehat{\overline{Y}}^{mq}_d+\frac{1}{n_d}\Big(1-\frac{n_d}{N_d}\Big)\sum_{j\in s_d}\widehat{\sigma}_{\widehat{\theta}_d}\phi\big(\widehat{u}_{\psi,dj}(\widehat{\theta}_d)\big),\tag{5.6}$$

where ϕ is an influence function with robustness parameter $c_\phi\geqslant 0$. The last summand on the right-hand side of (5.6) has the role of controlling the potential bias. The characterization of ϕ is worthy of comment. By setting the value of c_ϕ, it is possible to trade robustness for efficiency in MQ2 linear models. If $c_\phi=0$, the BMQ predictor reduces to the MQ predictor. As c_ϕ increases, more weight is given to larger residuals, biased by the MQ predictor. Consequently, larger values of c_ϕ lead to less bias, but also less robustness and more variability. It is safe to say that it is crucial to propose optimality criteria for a proper selection of c_ϕ. For brevity, this is presented in Section 5.4.4 for predictors derived from the TWMQ linear models but an analogous argument is applied to the BMQ predictor derived from the MQ2 linear models.

5.3 Three-fold M-quantile linear regression for SAE

A review of the three-level MQ linear (MQ3) models (Marchetti et al., 2018) is necessary to present the TWMQ linear models. Let U be a finite population of size N hierarchically partitioned in domains U_d and subdomains U_{dt} of sizes N_d and N_{dt}, respectively, $d=1,\ldots,D$,

$t = 1, \ldots, T$. Let s, s_d and s_{dt} be the corresponding sampled subsets of sizes n, n_d and n_{dt}, respectively. Throughout the theoretical part, we assume that a vector of $p \geqslant 1$ unit-level auxiliary variables $\boldsymbol{x}'_{dtj} = (x_{dtj1}, \ldots, x_{dtjp})$ is known for all individuals in U_{dt} and the variable of interest, y_{dtj}, is observed for all individuals in s_{dt}.

For $0 < q < 1$, the MQ3 linear model is

$$y_{dtj} = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q) + e_{\psi,dtj}(q), \quad d = 1, \ldots, D,\ t = 1, \ldots, T,\ j = 1, \ldots, N_{dt}, \tag{5.7}$$

where $\boldsymbol{\beta}_\psi(q) = (\beta_{\psi 1}(q), \ldots, \beta_{\psi p}(q))'$ is the vector of model parameters and $e_{\psi,dtj}(q)$ are independent model errors with unknown c.d.f. $F_q(u) = P(e_{\psi,dtj}(q) \leqslant u | \boldsymbol{x}_{dtj})$, $u \in \mathbb{R}$.

It is worth noting that $\boldsymbol{\beta}_\psi(q)$ only varies with the order of the quantile, $0 < q < 1$. Indeed, models (5.7) can be expressed with two subscripts, d and i, where i runs through all combinations of values of the indexes t and j. However, the adopted notation is necessary from Section 5.3.3 onwards, where the time component will be crucial.

For the MQ function, we assume that $Q_q(e_{\psi,dtj}(q); \sigma_q, \psi | \boldsymbol{x}_{dtj}) = 0$, $0 < q < 1$, and, although no explicit parametric assumptions are being made, the homoscedasticity assumption $\sigma_q = \text{var}^{1/2}(e_{\psi,dtj}(q)) = \sigma_\psi(\boldsymbol{\beta}_\psi(q))$ is imposed. One of the advantages of models (5.7) over LMMs is that we have not specified a formal structure for the model errors, let alone required them to follow a normal distribution.

In practice, $\widehat{\boldsymbol{\beta}}_\psi(q)$ and $\widehat{\sigma}_q$ are estimated using the IRLS algorithm, which ensures convergence to a unique solution (Bianchi and Salvati, 2015). This algorithm is explained step-by-step in Appendix C for estimating the regression parameters of the TWMQ linear models, and its adaptation to the MQ3 linear models is straightforward.

5.3.1 Three-fold M-quantile approach for inter-area variability

The MQ3 linear models have recently been used to model inter-area variability (Marchetti et al., 2018). The script is parallel to the one in Section 5.2.1. To start with, we have to introduce the new notation adapted to the structure of the subdomains.

For $j = 1, \ldots, N_{dt}$, the unit-level MQ coefficients of models (5.7) are

$$q_{dtj} = \underset{0<q<1}{\text{solution}} \left\{ Q_q(y_{dtj}; \sigma_q, \psi | \boldsymbol{x}_{dtj}) = y_{dtj} \right\},\ Q_q(y_{dtj}; \sigma_q, \psi | \boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q).$$

For each target variable y_{dtj}, $j = 1, \ldots, N_{dt}$, we calculate the quantile q_{dtj} for which the model error $e_{dtj}(q_{dtj})$ would be equal to zero if the β-coefficients were known.

By definition, it holds that

$$y_{dtj} = Q_{q_{dtj}}(y_{dtj}; \sigma_{q_{dj}}, \psi | \boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q_{dtj}), \quad \sigma_{q_{dtj}} = \sigma_\psi(\boldsymbol{\beta}_\psi(q_{dtj})).$$

The unit-level MQ coefficient q_{dtj} is the "most likely" quantile of unit j of area t and time period t. That is, of all the MQ3 linear models that vary by $0 < q < 1$, the model with $q = q_{dtj}$ would predict y_{dtj} without error if $\boldsymbol{\beta}_\psi(q)$ was known. Since the target variables

y_{dtj}, $j = 1, \ldots, n_{dt}$, are observed for all units in the sampled subsets s_{dt}, the unit-level MQ coefficients can be estimated in these units.

More specifically, an estimator of q_{dtj}, $j = 1, \ldots, n_{dt}$, is

$$\widehat{q}_{dtj} = \underset{0<q<1}{\text{solution}} \left\{ \widehat{Q}_q(y_{dtj}; \widehat{\sigma}_q, \psi | \boldsymbol{x}_{dtj}) = y_{dtj} \right\}, \quad \widehat{Q}_q(y_{dtj}; \widehat{\sigma}_q, \psi | \boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi(q).$$

The idea is to use the unit-level MQ coefficients of each individual to then estimate the average in each area and fit the MQ3 linear models to that q-value.

The domain population and sample means of unit-level MQ coefficients are

$$\theta_d \triangleq \bar{q}_{d..} = \frac{1}{N_d} \sum_{t=1}^{T} \sum_{j=1}^{N_{dt}} q_{dtj}, \quad \widehat{\theta}_d \triangleq \bar{\widehat{q}}_{d..} = \frac{1}{n_d} \sum_{t=1}^{T} \sum_{j=1}^{n_{dt}} \widehat{q}_{dtj}, \quad d = 1, \ldots, D, \tag{5.8}$$

respectively. By definition, $\widehat{\theta}_d$ is the area-level average of the estimated unit-level MQ coefficients of the units in the sample. Therefore, we expect that the MQ3 linear model with $q = \widehat{\theta}_d$ will be the one that provides the best predictions in the domain d. Overall if we predict linear domain parameters, such as population means. That is to say, in order to predict y_{dtj} in the unobserved part of domain d we will choose the MQ3 linear model with $q = \widehat{\theta}_d$. Thus, prediction in MQ3 linear models for SAE involves fitting D models, each one with the most appropriate quantile $q = \widehat{\theta}_d$ for the domain d it represents.

5.3.2 Robust predictors for three-fold M-quantile models

The MQ3 linear models are used to predict subdomain quantities including, but not only, population means. Specifically, we calculate predictors of addictive quantities $G_{dt} = g_{dt}(y_{dt1}, \ldots, y_{dtN_{dt}})$, where $g_{dt} : \mathbb{R}^{N_{dt}} \to \mathbb{R}$ is a continuous function. Therefore, its applicability overlaps with the EBP methodology based on NER models of Rao and Molina (2010). For example, we can predict poverty rates and poverty gaps, but this research focuses on predicting linear domain-dependent population values and, in particular, population means:

$$\overline{Y}_{dt} = \frac{1}{N_{dt}} \sum_{j=1}^{N_{dt}} y_{dtj}, \; d = 1, \ldots, D, \; t = 1, \ldots, T.$$

As the first option, a plug-in type predictor is calculated for the population means. The idea here comes from generalizing the developments described in Section 5.2.2. Let $r_{dt} = U_{dt} - s_{dt}$ be the non sampled subset of U_{dt}, $d = 1, \ldots, D$, $t = 1, \ldots, T$. If the sum of the residuals $e_{\psi,dtj}(\theta_d)$ in r_{dt} is close to zero, then

$$\begin{aligned} \overline{Y}_{dt} &= \frac{1}{N_{dt}} \sum_{j=1}^{N_{dt}} y_{dtj} = \frac{1}{N_{dt}} \left\{ \sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \boldsymbol{\beta}_\psi(\theta_d) + \sum_{j \in r_{dt}} e_{\psi,dtj}(\theta_d) \right\} \\ &\approx \frac{1}{N_{dt}} \left\{ \sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \boldsymbol{\beta}_\psi\big(\widehat{\theta}_d\big) \right\} + \frac{1}{N_{dt}} \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \frac{\partial \boldsymbol{\beta}_\psi(q)}{\partial q}\Big|_{q=\widehat{\theta}_d} (\theta_d - \widehat{\theta}_d). \end{aligned} \tag{5.9}$$

We define the residuals and standardized residuals of the MQ3 linear models as

$$\widehat{e}_{\psi,dtj}(\widehat{\theta}_d) = y_{dtj} - \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d), \quad \widehat{u}_{\psi,dtj}(\widehat{\theta}_d) = \widehat{\sigma}^{-1}_{\widehat{\theta}_d}\widehat{e}_{\psi,dtj}, \quad j = 1,\dots,n_{dt}.$$

The previous calculations involve the estimation of both the $\widehat{\boldsymbol{\beta}}_\psi$-coefficients and the area-level M-quantile coefficients $\widehat{\theta}_d$.

We define the MQ predictor of $\overline{Y}_{dt}$ as

$$\widehat{\overline{Y}}^{mq}_{dt} = \frac{1}{N_{dt}}\left\{ \sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi\big(\widehat{\theta}_d\big) \right\} \tag{5.10}$$

and the robust BMQ predictor of (5.10) as

$$\widehat{\overline{Y}}^{bmq}_{dt} = \widehat{\overline{Y}}^{mq}_{dt} + \frac{1}{n_{dt}}\Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \sum_{j\in s_{dt}} \widehat{\sigma}_{\widehat{\theta}_d}\phi\big(\widehat{u}_{\psi,dtj}(\widehat{\theta}_d)\big), \tag{5.11}$$

where ϕ is an influence function with robustness parameter $c_\phi \geqslant 0$. For the sake of brevity, the optimality criterion for the data-driven selection of c_ϕ is only given in Section 5.4.4 for predictors derived from the TWMQ linear models. Applying the developments to this case, however, is straightforward.

5.3.3 Residual analysis and inter-period weights

In this section we introduce the necessary notation to define the TWMQ linear models. The idea will be to capture underlying temporal dependencies in the MQ3 linear models, incorporating this dependency structure into the new models proposed in Section 5.4.

Let us define the subdomain-level residuals as

$$r_{\psi,dt} = \frac{1}{n_{dt}} \sum_{j\in s_{dt}} \widehat{e}_{\psi,dtj}(\widehat{\theta}_d), \quad d = 1,\dots,D, \quad t = 1,\dots,T,$$

and assume that there exists some unknown subdomain-level temporal dependency between the target variables $y_{dt_1j_1}$ and $y_{dt_2j_2}$, $t_1, t_2 = 1,\dots,T$, $j_1 = 1,\dots,N_{dt_1}$, $j_2 = 1,\dots,N_{dt_2}$. In such case, the subdomain-level temporal dependency will remain in the subdomain-level residuals r_{ψ,dt_1} and r_{ψ,dt_2}, $t_1, t_2 = 1,\dots,T$.

As already pointed out, the MQ3 linear models are able to non-parametrically capture area-level variability by fitting different regression surfaces

$$\widehat{Q}_{\widehat{\theta}_d}(y_{dtj};\widehat{\sigma}_q,\psi|\boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\widehat{\theta}_d),\ d = 1,\dots,D,\ t = 1,\dots,T,\ j = 1\dots,N_{dt},$$

but they are not able to model temporal dependencies. Intuition motivates us to fit a seasonal autoregressive $AR(P)_D$ model with period D and order $0 \leqslant P \leqslant T$ to $\{r_{\psi,dt} : d = 1,\dots,D,\ t = 1,\dots,T\}$. Here the domains play the role of seasons and P measures how the auto-correlation decays over time (Bugallo et al., 2024e). The extreme cases are $P = 0$, where there is no autoregressive dependence structure, and $P = T$ which includes all possible lags. Our idea is to define inter-period weights $w_{t_1t_2}$ that measure the dependency between r_{ψ,dt_1} and r_{ψ,dt_2}, $t_1, t_2 = 1,\dots,T$, based on the estimated coefficients of the fitted seasonal autoregressive $AR(P)_D$ model. It is necessary to differentiate between two cases:

(1) If $1 \leqslant t \leqslant P$, there is past information from the first period up to time t, so the weights will be distributed over $\{1, \ldots, t\}$. Accordingly, if $t = 1$ there is no past information, i.e., only data from the first period is available.

(2) If $P < t \leqslant T$, there is past information from the first period up to time t, but we only assign positive weights to the last P delays, i.e. $\{t - P, \ldots, t\}$.

Some information is included below. First, the equation of an $AR(P)_D$ process $\{z_i\}_{i\in\mathbb{Z}}$ is

$$z_i = \phi_0 + \phi_1 z_{i-D} + \phi_2 z_{i-2D} + \cdots + \phi_P z_{i-PD} + a_i,$$

where $\{a_i\}_{i\in\mathbb{Z}}$ is a collection of i.i.d. normal variables, of zero mean and finite variance σ_a^2, and $\phi_0, \phi_1, \ldots, \phi_P \in \mathbb{R}$ fulfil that $\phi_P \neq 0$ and $1 - \phi_1 u - \phi_2 u^2 - \cdots - \phi_P u^P \neq 0$, $\forall u \in \mathbb{C},\ |u| \leqslant 1$ (stationarity condition). Let be $S_t = \sum_{p=1}^{t} |\phi_p|$, $1 \leqslant t \leqslant P$, the sum of the first t autoregressive coefficients. The set of past time periods that produce a dependency in the distributions of the target variables at time period t are

$$\mathcal{T}_t = \{\widetilde{t}, \ldots, t\}, \quad \widetilde{t} = \widetilde{t}(t, P) = \begin{cases} 1 & \text{if } 1 \leqslant t \leqslant P, \\ t - P & \text{if } t > P, \end{cases}$$

so the vector of inter-period weights is $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tT})$, where

$$w_{ti} = \frac{|\phi_{t+1-i}|}{S_t} \text{ if } \widetilde{t} \leqslant i \leqslant t, \text{ and } w_{ti} = 0 \text{ if } i > t \text{ or } i < \widetilde{t}, \quad t = 1, \ldots, T. \tag{5.12}$$

For each $t = 1, \ldots, T$, the population and samples sizes are $N_{.t} = \sum_{d=1}^{D} N_{dt}$ and $n_{.t} = \sum_{d=1}^{D} n_{dt}$, respectively, and the relevant subsets at time period t and corresponding sizes, are

$$\begin{aligned} s_{d(t)} &= \bigcup_{i\in\mathcal{T}_t} s_{di}, \quad U_{d(t)} = \bigcup_{i\in\mathcal{T}_t} U_{di}; \quad n_{d(t)} = \sum_{i\in\mathcal{T}_t} n_{di}, \quad N_{d(t)} = \sum_{i\in\mathcal{T}_t} N_{di}, \\ s_{(t)} &= \bigcup_{d=1}^{D} s_{d(t)}, \quad U_{(t)} = \bigcup_{d=1}^{D} U_{d(t)}; \quad n_{(t)} = \sum_{d=1}^{D} n_{d(t)}, \quad N_{(t)} = \sum_{d=1}^{D} N_{d(t)}. \end{aligned} \tag{5.13}$$

The heuristic of inter-period weights attribution given by equation (5.12) and the notation just included in (5.13) will be quite relevant in Section 5.4. Indeed, the subsets $s_{(t)}$ determine, at time period t, the set of time-dependent observations that will be assigned positive weights, according to (5.12), in the fitting algorithm of the new TWMQ linear models.

5.4 Time-Weighted M-quantile statistical methodology

LMMs with time-dependent structures rely on strong distributional assumptions, and it is also necessary to formally specify the dependence structure of the random effects. Therefore, there is a need for time-dependent SAE models that are robust against atypical data and have fewer parametric assumptions. To cover this gap, this section provides a detailed description of the proposed Time-Weighted M-quantile (TWMQ) statistical methodology. It follows Bugallo

et al. (2024e) as a reference point. First, the model formulation is presented. Section 5.4.1 derives two robust plug-in type predictors of small area means and time periods. Sections 5.4.2 and 5.4.3 focus on the estimation of the MSE and, directly related, Section 5.4.4 on that of the robustness parameter for bias correction. Appendix C outlines the steps of the fitting algorithm, an adaptation of the IRLS algorithm. Appendix D contains the mathematical proofs of Theorems 1 and 2.

For $0 < q < 1$, $t = 1, \ldots, T$, the time-weighted MQ linear regression (TWMQ) models are

$$y_{dij} = \boldsymbol{x}'_{dij}\boldsymbol{\beta}_{\psi}(q, \boldsymbol{w}_t) + e_{\psi,dij}(q, \boldsymbol{w}_t), \quad d = 1, \ldots, D,\ i = 1, \ldots, T,\ j = 1, \ldots, N_{di}, \tag{5.14}$$

where $\boldsymbol{\beta}_{\psi}(q, \boldsymbol{w}_t) = (\beta_{\psi 1}(q, \boldsymbol{w}_t), \ldots, \beta_{\psi p}(q, \boldsymbol{w}_t))'$ is the vector of time-varying model parameters, $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tT})'$ is the vector of known non-negative inter-period weights measuring the time dependency between observation of times t and i, and $e_{\psi,dij}(q, \boldsymbol{w}_t)$ are independent model errors with unknown c.d.f. $F_{qt}(u) = P(e_{\psi,dij}(q, \boldsymbol{w}_t) \leqslant u | \boldsymbol{x}_{dij})$, $u \in \mathbb{R}$. The new models feature time-varying parameters $\boldsymbol{\beta}_{\psi}(q, \boldsymbol{w}_t)$, $0 < q < 1$, $t = 1, \ldots, T$, and allow for non-parametric modelling of time dependence structures for each probability q.

It is satisfied by definition that

$$Q_q(e_{\psi,dij}(q, \boldsymbol{w}_t); \sigma_{qt}, \psi | \boldsymbol{x}_{dij}) = 0,\ 0 < q < 1,$$

and, although no explicit parametric assumptions are being made, the homoscedasticity assumption $\sigma_{qt} = \mathrm{var}^{1/2}(e_{\psi,dij}(q, \boldsymbol{w}_t)) = \sigma_{\psi}(\boldsymbol{\beta}_{\psi}(q, \boldsymbol{w}_t))$ is imposed.

Our choice is to set the vector of weights $\boldsymbol{w}_t$, $t = 1, \ldots, T$, according to (5.12), so the time-dependent subsets are defined in (5.13). As in the case of the MQ2 and MQ3 linear models, we do not need to specify the distribution of the model errors. As an inherited property of MQ models, our proposal avoids distributional assumptions and allows characterizing differences between areas, as well as time dependencies, through data-driven estimation of the model parameters. Therefore, not only do the new models have time-varying parameters, but they are also distribution-free for both areas and time.

Here it is important to note the clear differences of the TWMQ linear models (5.14) compared to the MQ Geographically Weighted Regression (Salvati et al., 2012), where the weights are symmetric, i.e. where something like $w_{ti} = w_{it}$, $i, t = 1, \ldots, T$, should be imposed. For spatial dependencies, adding new locations does not necessarily imply that those already considered lose relevance. In contrast, in the case of temporal dependencies, the first lags lose relevance as more recent information becomes available. One of the great advantages of the TWMQ linear models is that recent data can be easily incorporated into the fitting process. In fact, it manages to attribute only positive weights to the closest temporal data through the sets $\mathcal{T}_t$, $t = 1, \ldots, T$. As a result, the computational cost of fitting the TWMQ linear models is not scalable if we include more recent observations. That is, the model parameters $\boldsymbol{\beta}_{\psi}(q, \boldsymbol{w}_t)$, $0 < q < 1$, $t = 1, \ldots, T$, vary over time, using only the nearest data to estimate them. The relationship between the response variable and the covariates is characterized by local rather than global parameters, where local is defined as time specific.

In light of the above, the TWMQ linear models allows us to define the fitted regression surfaces

$$\widehat{Q}_q(y_{dtj}; \sigma_{qt}, \psi | \boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}(q, \boldsymbol{w}_t),\ d = 1, \ldots, D,\ t = 1, \ldots, T,\ j = 1, \ldots, N_{dt},$$

that uses information across quantiles $0 < q < 1$ and over time.

5.4.1 Robust predictors for Time-Weighted M-quantile models

The TWMQ linear models are used to predict subdomain quantities including, but not only, population means. In fact, they can be used to predict the same type of predictors as the MQ3 linear models in Section 5.3.2, but with this new methodology we can capture underlying time dependencies. For the purpose of using the TWMQ linear models in SAE, we use $\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t)$, where the estimation of θ_d, $d = 1, \ldots, D$, comes from the MQ3 linear models (5.7). Indeed, it has been given in (5.8) so we finally estimate $\boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)$. We follow a robust-projective approach based on plug-in robust prediction, i.e. the optimal, but outlier-sensitive, parameter estimates are replaced by outlier-robust versions.

In the following, we include some mathematical notation and developments to derive the small area predictors. For the quantile $q = \theta_d$, we write

$$\sigma_{\theta_d t} = \sigma_\psi(\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t)),\ d = 1, \ldots, D,\ t = 1, \ldots, T,$$

and introduce a simplified notation for model errors and standardized errors, i.e.

$$e_{\psi,dtj} = y_{dtj} - \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t),\ u_{\psi,dtj} = \sigma^{-1}_{\theta_d t} e_{\psi,dtj}(\theta_d, \boldsymbol{w}_t),\ j = 1, \ldots, N_{dt}. \tag{5.15}$$

For $j = 1, \ldots, n_{di}$, $i = 1, \ldots, T$, the model residuals are

$$\widehat{e}_{\psi,dij}(q, \boldsymbol{w}_t) = y_{dij} - \boldsymbol{x}'_{dij}\widehat{\boldsymbol{\beta}}_\psi(q, \boldsymbol{w}_t),\ 0 < q < 1.$$

For $j = 1, \ldots, N_{dt}$, $0 < q < 1$, we define the pseudo-residuals and standardized pseudo-residuals

$$\begin{aligned}
\widetilde{e}_{\psi,dtj}(q) &= \boldsymbol{x}'_{dtj}\big(\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\theta_d, \boldsymbol{w}_t)\big), \quad \widetilde{u}_{\psi,dtj}(q) = \sigma^{-1}_{\theta_d t}\,\widetilde{e}_{\psi,dtj}(q), \\
\widehat{e}_{\psi,dtj}(q) &= \boldsymbol{x}'_{dtj}\big(\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big), \quad \widehat{u}_{\psi,dtj}(q) = \sigma^{-1}_{\theta_d t}\widehat{e}_{\psi,dtj}(q), \\
\breve{e}_{\psi,dtj}(q) &= \boldsymbol{x}'_{dtj}\big(\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t) - \boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big), \quad \breve{u}_{\psi,dtj}(q) = \sigma^{-1}_{\theta_d t}\breve{e}_{\psi,dtj}(q).
\end{aligned} \tag{5.16}$$

For $j = 1, \ldots, N_{dt}$, the unit-level MQ coefficients of models (5.14) are

$$q_{dtj} = \underset{0<q<1}{\text{solution}} \left\{ Q_q(y_{dtj}; \sigma_{qt}, \psi | \boldsymbol{x}_{dtj}) = y_{dtj} \right\},\ Q_q(y_{dtj}; \sigma_{qt}, \psi | \boldsymbol{x}_{dtj}) = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t), \tag{5.17}$$

so it holds that

$$y_{dtj} = \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t), \quad \sigma_{q_{dtj}} = \sigma_\psi(\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t)).$$

The interpretation of the unit-level MQ coefficients of models (5.14) is the same as for the MQ2 and MQ3 linear models, but in this case the estimation of $\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)$ and $\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t)$ allow for past time dependencies to be taken into account.

For the quantile $q = q_{dtj}$, the pseudo-residuals and standardized pseudo-residuals are

$$\begin{aligned}
\widetilde{e}_{\psi,dtj} \triangleq \widetilde{e}_{\psi,dtj}(q_{dtj}) &= y_{dtj} - \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi(\theta_d, \boldsymbol{w}_t), \quad \widetilde{u}_{\psi,dtj} = \sigma^{-1}_{\theta_d t}\,\widetilde{e}_{\psi,dtj}(q_{dtj}), \\
\widehat{e}_{\psi,dtj} \triangleq \widehat{e}_{\psi,dtj}(q_{dtj}) &= y_{dtj} - \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t), \quad \widehat{u}_{\psi,dtj} = \sigma^{-1}_{\theta_d t}\widehat{e}_{\psi,dtj}(q_{dtj}), \\
\breve{e}_{\psi,dtj} \triangleq \breve{e}_{\psi,dtj}(q_{dtj}) &= y_{dtj} - \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t), \quad \breve{u}_{\psi,dtj} = \sigma^{-1}_{\theta_d t}\breve{e}_{\psi,dtj}(q_{dtj}).
\end{aligned} \tag{5.18}$$

In future developments, such as the proof of Theorem 1, the previous notation will be useful.

As the first option, a plug-in type predictor is calculated for the population means. Under regularity assumptions, a Taylor series expansion of $\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t)$ around $\widehat{\theta}_d$ yields to

$$\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) \approx \boldsymbol{\beta}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big) + \frac{\partial \boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)}{\partial q}\Big|_{q=\widehat{\theta}_d} (\theta_d - \widehat{\theta}_d), \quad d = 1, \ldots, D.$$

If we assume that the sum of model errors $e_{\psi,dtj}$ is close to zero, we have

$$\begin{aligned}
\overline{Y}_{dt} &= \frac{1}{N_{dt}} \Bigg\{ \sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) + \sum_{j \in r_{dt}} e_{\psi,dtj} \Bigg\} \\
&\approx \frac{1}{N_{dt}} \Bigg\{ \sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \boldsymbol{\beta}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big) \Bigg\} + \frac{1}{N_{dt}} \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \frac{\partial \boldsymbol{\beta}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big)}{\partial \widehat{\theta}_d} (\theta_d - \widehat{\theta}_d).
\end{aligned}$$

Typically, the second summand of the last expression is much smaller than the first one. Therefore, we define the TMQ predictor of $\overline{Y}_{dt}$ as

$$\widehat{\overline{Y}}^{tmq}_{dt} = \frac{1}{N_{dt}} \Bigg\{ \sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big) \Bigg\}. \tag{5.19}$$

Given that $Q_{\theta_d}(e_{\psi,dtj}; \sigma_{\theta_d t}, \psi | \boldsymbol{x}_{dtj}) = 0$, it follows that $\sum_{j \in r_{dt}} e_{\psi,dtj} \approx 0$ if $\theta_d \in (1/2 - \varepsilon, 1/2 + \varepsilon)$ for some small $\varepsilon > 0$, but not otherwise. It is therefore desirable to know the sign of the bias of (5.19) and reduce its magnitude, as the TMQ predictor is expected to be a biased predictor of the population mean. If $\theta_d < 1/2 - \varepsilon$, the event $\sum_{j \in r_{dt}} e_{\psi,dtj} > 0$ will occur with greater probability than the opposite event and the TMQ predictor will tend to have negative bias. If $\theta_d > 1/2 - \varepsilon$, however, the TMQ predictor will tend to have positive bias. If

$$\frac{1}{n_{dt}} \sum_{j \in s_{dt}} \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big) \approx \frac{1}{N_{dt}} \sum_{j \in U_{dt}} \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi\big(\widehat{\theta}_d, \boldsymbol{w}_t\big),$$

the bias of (5.19) is estimated as

$$B\big(\widehat{\overline{Y}}^{tmq}_{dt}\big) = E\big[\widehat{\overline{Y}}^{tmq}_{dt} - \overline{Y}_{dt}\big] = -\frac{1}{n_{dt}} \Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \sum_{j \in s_{dt}} \widehat{e}_{\psi,dtj}.$$

Based on a robustification of $\widehat{e}_{\psi,dtj}$, the robust bias-corrected (BTMQ) predictor of (5.19) is

$$\widehat{\overline{Y}}^{btmq}_{dt} = \widehat{\overline{Y}}^{tmq}_{dt} + \frac{1}{n_{dt}} \Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \widehat{B}^{btmq}_{dt}, \quad \widehat{B}^{btmq}_{dt} = \sum_{j \in s_{dt}} \sigma_{\theta_d t} \phi\big(\widehat{u}_{\psi,dtj}\big) \tag{5.20}$$

where ϕ is an influence function with robustness parameter $c_\phi \geqslant 0$. For the first time in the literature, in Section 5.4.4 we propose an optimality criterion for a proper selection of c_ϕ.

5.4.2 Mean squared error estimation for temporal M-quantile predictors

In this section we address the analytical estimation of the MSE of the TMQ predictor. Let $d = 1, \ldots, D$, $t = 1, \ldots, T$. Let us assume that $q = \widehat{\theta}_d$ is known. We define the $n(t) \times 1$

vector indicating the units of $s(t)$ that belongs to domain d and time period t, i.e.

$$\boldsymbol{\varepsilon}_{dt} = \underset{1\leqslant g\leqslant D}{\text{col}}\big(\delta_{gd}\underset{i\in\mathcal{T}_t}{\text{col}}(\delta_{it}\underset{1\leqslant j\leqslant n_{gi}}{\text{col}}(1))\big) = \underset{1\leqslant g\leqslant D}{\text{col}}\big(\underset{i\in\mathcal{T}_t}{\text{col}}(\underset{1\leqslant j\leqslant n_{gi}}{\text{col}}(\varepsilon_{dt,gij}))\big), \quad \delta_{ab} = \begin{cases} 1 & \text{if } a=b, \\ 0 & \text{otherwise.} \end{cases}$$

Then we write the TMQ predictor in the linear form

$$\begin{aligned} \widehat{\overline{Y}}_{dt}^{tmq} &= \frac{1}{N_{dt}}\left\{\boldsymbol{\varepsilon}'_{dt}\boldsymbol{y}_{s(t)} + \left(\sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\right)\big(X'_{s(t)}W_{s(t)}(\widehat{\theta}_d,\boldsymbol{w}_t)X_{s(t)}\big)^{-1}X'_{s(t)}W_{s(t)}(\widehat{\theta}_d,\boldsymbol{w}_t)\boldsymbol{y}_{s(t)}\right\} \\ &= \frac{1}{N_{dt}}\boldsymbol{a}'_{dt}\boldsymbol{y}_{s(t)} = \frac{1}{N_{dt}}\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}} a_{dt,gij}y_{gij}, \end{aligned}$$

where $\boldsymbol{a}'_{dt} = \boldsymbol{\varepsilon}'_{dt} + \boldsymbol{z}'_{dt} = \underset{1\leqslant g\leqslant D}{\text{col}}\big(\underset{i\in\mathcal{T}_t}{\text{col}}(\underset{1\leqslant j\leqslant n_{gi}}{\text{col}}(a_{dt,gij}))\big)$, $a_{dt,gij} = \varepsilon_{dt,gij} + z_{dt,gij}$ and

$$\boldsymbol{z}'_{dt} = \underset{1\leqslant g\leqslant D}{\text{col}}\big(\underset{i\in\mathcal{T}_t}{\text{col}}(\underset{1\leqslant j\leqslant n_{gi}}{\text{col}}(z_{dt,gij}))\big) = \left(\sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\right)\big(X'_{s(t)}W_{s(t)}(\widehat{\theta}_d,\boldsymbol{w}_t)X_{s(t)}\big)^{-1}X'_{s(t)}W_{s(t)}(\widehat{\theta}_d,\boldsymbol{w}_t).$$

We have written in TMQ predictor as a sum of the target variables multiplied by certain weights, stored in the vector $\boldsymbol{a}_{dt}$. Let us define the $N(t)\times 1$ vector indicating the units of $U(t)$ that belongs to domain d and time period t and the vectors $\mathbf{1}_{dt} = \underset{1\leqslant j\leqslant N_{dt}}{\text{col}}(1)$, $\boldsymbol{y}_{dt} = \underset{1\leqslant j\leqslant N_{dt}}{\text{col}}(y_{dtj})$, $\mathbf{1}_{r_{dt}} = \underset{j\in r_{dt}}{\text{col}}(1)$, $\boldsymbol{y}_{r_{dt}} = \underset{j\in r_{dt}}{\text{col}}(y_{dtj})$.

The prediction error and MSE of $\widehat{\overline{Y}}_{dt}^{tmq}$ are, respectively,

$$\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt} = \frac{1}{N_{dt}}\big(\boldsymbol{a}'_{dt}\boldsymbol{y}_{s(t)} - \mathbf{1}'_{dt}y_{dt}\big) = \frac{1}{N_{dt}}\big(\boldsymbol{z}'_{dt}\boldsymbol{y}_{s(t)} - \mathbf{1}'_{r_{dt}}\boldsymbol{y}_{r_{dt}}\big)$$

and

$$MSE(\widehat{\overline{Y}}_{dt}^{tmq}) = E\Big[\big(\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt}\big)^2\Big] = \text{var}\big(\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt}\big) + B^2\big(\widehat{\overline{Y}}_{dt}^{tmq}\big),$$

where

$$\text{var}\big(\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt}\big) = \frac{1}{N_{dt}^2}\text{var}\big(\boldsymbol{z}'_{dt}\boldsymbol{y}_{s(t)} - \mathbf{1}'_{r_{dt}}\boldsymbol{y}_{r_{dt}}\big), \quad B\big(\widehat{\overline{Y}}_{dt}^{tmq}\big) = \frac{1}{N_{dt}}E\big[\boldsymbol{z}'_{dt}\boldsymbol{y}_{s(t)} - \mathbf{1}'_{r_{dt}}\boldsymbol{y}_{r_{dt}}\big].$$

Following Chambers and Tzavidis (2006), we first derive a first-order approximation of $MSE(\widehat{\overline{Y}}_{dt}^{tmq})$ and then propose several estimators. In general, the procedure described below could be applied when the quantity to be predicted can be expressed as a linear combination of the values taken by the target variable in all units of the sample (Chambers et al., 2011). These pseudo-linear MSE estimators assume that the weights $\boldsymbol{a}_{dt}$ are fixed quantities, and thus ignore their contribution to the MSE, derived from the estimation of the θ_d coefficients using the MQ3 linear models. In practice, the latter should not be a major problem, as this variability is expected to be rather small (Schirripa Spagnolo et al., 2021).

Firstly, the variance is approximated as

$$\begin{aligned}\operatorname{var}\big(\widehat{\overline{Y}}_{dt}^{tmq}-\overline{Y}_{dt}\big) &\approx \frac{1}{N_{dt}^2}\Bigg(\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}} z_{dt,gij}^2\operatorname{var}(y_{gij})+\sum_{j\in r_{dt}}\operatorname{var}(y_{dtj})\Bigg) \qquad (5.21)\\ &= \frac{1}{N_{dt}^2}\Bigg(\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}} z_{dt,gij}^2\operatorname{var}(y_{gij})+\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{N_{gi}} \varepsilon_{rdt,gij}\operatorname{var}(y_{gij})\Bigg),\end{aligned}$$

where $\boldsymbol{\varepsilon}_{rdt}=\underset{1\leqslant g\leqslant D}{\operatorname{col}}\big(\delta_{gd}\underset{i\in\mathcal{T}_t}{\operatorname{col}}(\delta_{it}\operatorname{col}(\underset{j\in s_{gi}}{\operatorname{col}}(0),\underset{j\in r_{gi}}{\operatorname{col}}(1))))=\underset{1\leqslant g\leqslant D}{\operatorname{col}}\big(\underset{i\in\mathcal{T}_t}{\operatorname{col}}(\underset{1\leqslant j\leqslant N_{gi}}{\operatorname{col}}(\varepsilon_{rdt,gij}))\big)$.

To estimate $\operatorname{var}(y_{gij})$ and $\operatorname{var}(y_{dtj})$, we consider two approaches.

(1) *Median approach*: the variance estimators are based on the median model, and not on the area quantile coefficient models. Accordingly, we use the median estimator $\widehat{\operatorname{var}}(y_{gij})=(y_{gij}-\boldsymbol{x}_{gij}'\widehat{\boldsymbol{\beta}}_\psi(0.5,\boldsymbol{w}_t))^2=\widehat{e}_{\psi,gij}^2(0.5,\boldsymbol{w}_t)$ for the variance of the sample observations in $s(t)$ and estimate the second summand in (5.21) using

$$\sum_{j\in r_{dt}}\widehat{\operatorname{var}}(y_{dtj})=\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}}\varepsilon_{dt,gij}\frac{N_{dt}-n_{dt}}{n_{d(t)}-1}\,\widehat{e}_{\psi,gij}^2(0.5,\boldsymbol{w}_t).$$

Therefore, the median estimator of $\operatorname{var}\big(\widehat{\overline{Y}}_{dt}^{tmq}-\overline{Y}_{dt}\big)$ is

$$\widehat{V}_{11,dt}^{tmq}=\frac{1}{N_{dt}^2}\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}}\lambda_{dt,gij}\,\widehat{e}_{\psi,gij}^2(0.5,\boldsymbol{w}_t),\quad \lambda_{dt,gij}=z_{dt,gij}^2+\frac{N_{dt}-n_{dt}}{n_{d(t)}-1}\,\varepsilon_{dt,gij}.$$

(2) *Area quantile coefficient approach*: the variance is estimated according to the representative quantile of the area g from which the observation is drawn. Consequently, we use the area quantile coefficient estimator $\widehat{\operatorname{var}}(y_{gij})=(y_{gij}-\boldsymbol{x}_{gij}'\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_g,\boldsymbol{w}_t))^2=\widehat{e}_{\psi,gij}^2(\widehat{\theta}_g,\boldsymbol{w}_t)$ for the variance of the sample observations in $s(t)$ and estimate the second summand in (5.21) using

$$\sum_{j\in r_{dt}}\widehat{\operatorname{var}}(y_{dtj})=\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}}\varepsilon_{dt,gij}\frac{N_{dt}-n_{dt}}{n_{d(t)}-1}\,\widehat{e}_{\psi,gij}^2(\widehat{\theta}_g,\boldsymbol{w}_t).$$

Therefore, the area quantile coefficient estimator of $\operatorname{var}\big(\widehat{\overline{Y}}_{dt}^{tmq}-\overline{Y}_{dt}\big)$ is

$$\widehat{V}_{12,dt}^{tmq}=\frac{1}{N_{dt}^2}\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{gi}}\lambda_{dt,gij}\,\widehat{e}_{\psi,gij}^2(\widehat{\theta}_g,\boldsymbol{w}_t).$$

The selection of robustness parameters from a minimum MSE is expected to give better results (Bugallo et al., 2024e), so we also look for a variance estimate that allows us to separate the terms derived from the bias correction from the terms common to the TMQ and BTMQ predictors. By using a different formula for the prediction error of $\widehat{\overline{Y}}_{dt}^{tmq}$, we give two alternative estimators of the variance. For this sake, we define the scalars and vectors

$$\bar{y}_{sdt}=\frac{1}{n_{dt}}\sum_{j\in s_{dt}}y_{dtj},\quad \bar{y}_{rdt}=\frac{1}{N_{dt}-n_{dt}}\sum_{j\in r_{dt}}y_{dtj},\quad \bar{\boldsymbol{x}}_{rdt}'=\frac{1}{N_{dt}-n_{dt}}\sum_{j\in r_{dt}}\boldsymbol{x}_{dtj}'.$$

From equation (5.19), we write the TMQ predictor as

$$\widehat{\overline{Y}}_{dt}^{tmq} = \frac{1}{N_{dt}}\Big\{n_{dt}\bar{y}_{sdt} + (N_{dt}-n_{dt})\bar{\boldsymbol{x}}'_{rdt}\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_d, \boldsymbol{w}_t)\Big\}$$

and the population mean as

$$\overline{Y}_{dt} = \frac{1}{N_{dt}}\Big\{n_{dt}\bar{y}_{sdt} + (N_{dt}-n_{dt})\bar{y}_{rdt}\Big\}$$

so the prediction error is

$$\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt} = \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\Big\{\bar{\boldsymbol{x}}'_{rdt}\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_d, \boldsymbol{w}_t) - \bar{y}_{rdt}\Big\}.$$

An estimator of the prediction error variance is

$$\widehat{\mathrm{var}}\big(\widehat{\overline{Y}}_{dt}^{tmq} - \overline{Y}_{dt}\big) = \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2\bar{\boldsymbol{x}}'_{rdt}\widehat{V}_{\beta}\bar{\boldsymbol{x}}_{rdt} + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2\widehat{\mathrm{var}}(\bar{y}_{rdt}), \tag{5.22}$$

where $\widehat{V}_{\beta} = \widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_d, \boldsymbol{w}_t))$. Based on the sandwich approach (Street et al., 1988) to estimate the asymptotic variance of the vector of model parameters in MQ linear models (Bianchi and Salvati, 2015), we derive an estimator of V_{β} to be plugged into (5.22). Under assumptions (A1)-(A8) in Appendix D.1, an estimator of V_{β} is

$$\widehat{V}_{\beta} = \frac{n_{(t)}^2\sigma^2_{\theta_{dt}}}{n_{(t)}-p}\frac{\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j\in s_{gi}}\psi^2_{\widehat{\theta}_{dt}}\big(\widehat{e}_{\psi,gij}(\widehat{\theta}_d,\boldsymbol{w}_t),\sigma_{\theta_{dt}}\big)}{\Big(\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j\in s_{gi}}\dot{\psi}_{\widehat{\theta}_{dt}}\big(\widehat{e}_{\psi,gij}(\widehat{\theta}_d,\boldsymbol{w}_t),\sigma_{\theta_{dt}}\big)\Big)^2}\left(\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j\in s_{gi}}\boldsymbol{x}'_{gij}\boldsymbol{x}_{gij}\right)^{-1}, \tag{5.23}$$

where $\dot{\psi}_{\widehat{\theta}_d,t}$ is the partial derivative of $\psi_{\widehat{\theta}_d,t}$ with respect to the first argument.

In order to estimate $\mathrm{var}\big(\bar{y}_{rdt}\big)$ in (5.22), we use

$$\widehat{\mathrm{var}}_1\big(\bar{y}_{rdt}\big) = \frac{\sum_{j\in s_{dt}}\widehat{e}^2_{\psi,dtj}}{(N_{dt}-n_{dt})(n_{d(t)}-1)} \quad \text{or} \quad \widehat{\mathrm{var}}_2\big(\bar{y}_{rdt}\big) = \frac{\sum_{g=1}^{D}\sum_{j\in s_{gt}}\widehat{e}^2_{\psi,gtj}}{(N_{dt}-n_{dt})(n-D)}. \tag{5.24}$$

By substituting $\widehat{V}_{\beta}$ and $\widehat{\mathrm{var}}_1\big(\bar{y}_{rdt}\big)$ or $\widehat{\mathrm{var}}_2\big(\bar{y}_{rdt}\big)$ in (5.22), we obtain the estimators $\widehat{V}^{tmq}_{21,dt}$ and $\widehat{V}^{tmq}_{22,dt}$, respectively. Secondly, the bias $B\big(\widehat{\overline{Y}}_{dt}^{tmq}\big)$ is estimated by

$$\widehat{B}_{dt} = \frac{1}{N_{dt}}\left(\sum_{g=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j\in s_{gi}}a_{dt,gij}\boldsymbol{x}'_{gij}\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_g,\boldsymbol{w}_t) - \sum_{j\in U_{dt}}\boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_d,\boldsymbol{w}_t)\right), \tag{5.25}$$

where the terms $a_{dt,gij}$'s are the components of $\boldsymbol{a}_{dt}$ appearing in the definition of the prediction error of $\widehat{\overline{Y}}_{dt}^{tmq}$. All in all, we have four estimators of $MSE(\widehat{\overline{Y}}_{dt}^{tmq})$. They are $mse^{tmq}_{11,dt} = \widehat{V}^{tmq}_{11,dt} + \widehat{B}^2_{dt}$, $mse^{tmq}_{12,dt} = \widehat{V}^{tmq}_{12,dt} + \widehat{B}^2_{dt}$, $mse^{tmq}_{21,dt} = \widehat{V}^{tmq}_{21,dt} + \widehat{B}^2_{dt}$ and $mse^{tmq}_{22,dt} = \widehat{V}^{tmq}_{22,dt} + \widehat{B}^2_{dt}$. By taking squared roots of the above estimators, we define $rmse^{tmq}_{11,dt}$, $rmse^{tmq}_{12,dt}$, $rmse^{tmq}_{21,dt}$ and $rmse^{tmq}_{22,dt}$, respectively. They are estimators of $RMSE(\widehat{\overline{Y}}_{dt}^{tmq}) = \big(MSE(\widehat{\overline{Y}}_{dt}^{tmq})\big)^{1/2}$.

5.4.3 Mean squared error estimation for bias-corrected temporal M-quantile predictors

In this section we address the analytical estimation of the MSE of the BTMQ predictor. We focus on the prediction of the population means $\overline{Y}_{dt}$, which have been defined in (5.9). To start with, the BTMQ predictor of $\overline{Y}_{dt}$ can be written as

$$\widehat{\overline{Y}}_{dt}^{btmq} = \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}\big(\widehat{\theta}_d, \boldsymbol{w}_t\big)\right\} + \frac{1}{n_{dt}}\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\widehat{B}_{dt}^{btmq},$$

where $\widehat{B}_{dt}^{btmq}$ has been defined in (5.20).

We derive a first-order approximation of the MSE of the BTMQ predictor based on a decomposition of it to account for the variability arising from the estimation of q_{dtj}, $j = 1,\ldots,n_{dt}$, and $\boldsymbol{\beta}_{\psi}\big(\theta_d, \boldsymbol{w}_t\big)$. To do so, we first define the following auxiliary notation in relation to the BTMQ predictor:

$$\begin{aligned}
\overline{Y}_{dt}^{btmq} &= \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_{\psi}(\theta_d, \boldsymbol{w}_t)\right\} + \frac{1}{n_{dt}}\Big(1-\frac{n_{dt}}{N_{dt}}\Big)B_{dt}^{btmq},\\
\widetilde{\overline{Y}}_{dt}^{btmq} &= \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}(\theta_d, \boldsymbol{w}_t)\right\} + \frac{1}{n_{dt}}\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\widetilde{B}_{dt}^{btmq},\\
B_{dt}^{btmq} &= \sum_{j\in s_{dt}} \sigma_{\theta_{dt}}\phi(u_{\psi,dtj}), \quad \widetilde{B}_{dt}^{btmq} = \sum_{j\in s_{dt}} \sigma_{\theta_{dt}}\phi(\widetilde{u}_{\psi,dtj}),
\end{aligned}$$

where $u_{\psi,dtj}$ and $\widetilde{u}_{\psi,dtj}$, $j = 1,\ldots,n_{dt}$, have been defined in (5.15) and (5.18), respectively.

It holds that

$$MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) = E\big[\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big)^2\big] = \mathrm{var}\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big) + \big(E\big[\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big]\big)^2.$$

Under the assumptions listed in Appendix D.1.1, the prediction error of $\widehat{\overline{Y}}_{dt}^{btmq}$ is:

$$\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt} = \big(\widehat{\overline{Y}}_{dt}^{btmq} - \widetilde{\overline{Y}}_{dt}^{btmq}\big) + \big(\widetilde{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}^{btmq}\big) + \big(\overline{Y}_{dt}^{btmq} - \overline{Y}_{dt}\big) = \overline{Y}_{dt}^{(3)} + \overline{Y}_{dt}^{(2)} + \overline{Y}_{dt}^{(1)}. \quad (5.26)$$

To simplify the notation, we define the variance and the expected prediction difference of $\overline{Y}_{dt}^{(k)}$, for $k = 1, 2, 3$, as $V_{dt}^{(k)}$, and $E_{dt}^{(k)}$, respectively.

Based on the decomposition in equation (5.26), we write:

$$\begin{aligned}
\mathrm{var}\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big) &= V_{dt}^{(1)} + V_{dt}^{(2)} + V_{dt}^{(3)} + 2\mathrm{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(2)}\big) + 2\mathrm{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(1)}\big) + 2\mathrm{cov}\big(\overline{Y}_{dt}^{(2)}, \overline{Y}_{dt}^{(1)}\big),\\
E\big[\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big] &= E_{dt}^{(1)} + E_{dt}^{(2)} + E_{dt}^{(3)} = E_{dt}^{(1)} + o(1).
\end{aligned}$$

The covariances are $\mathrm{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(2)}\big) = E\big[\overline{Y}_{dt}^{(3)}\overline{Y}_{dt}^{(2)}\big] + o(1)$, $\mathrm{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(1)}\big) = E\big[\overline{Y}_{dt}^{(3)}\overline{Y}_{dt}^{(1)}\big] + o(1)$ and $\mathrm{cov}\big(\overline{Y}_{dt}^{(2)}, \overline{Y}_{dt}^{(1)}\big) = E\big[\overline{Y}_{dt}^{(2)}\overline{Y}_{dt}^{(1)}\big] + o(1)$. Under regularity assumptions, the expectations of the previous cross-products should be $o(1)$. We define the set $\mathcal{G}_{dt} = \Big\{j \in s_{dt} : \big|u_{\psi,dtj}\big| < c_{\phi}\Big\}$. The following theorem summarizes the final approximation of $MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big)$.

Theorem 1. Under assumptions (Φ1), (N1)-(N2), (Q1), (A1)-(A9), (B1)-(B4), (C1)-(C2), (D1)-(D2), (E1)-(E3) in Appendix D.1.1, a first-order approximation of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$ is

$$\begin{aligned}
MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) &= V_{dt}^{(1)} + V_{dt}^{(2)} + V_{dt}^{(3)} + E_{dt}^{(1)2} + o(1) = \\
&= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\mathcal{G}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\Big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d,\boldsymbol{w}_t)\Big)^2 \xi_{dt}^2 \\
&+ \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\widetilde{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\boldsymbol{x}'_{dtj}\mathrm{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\theta_d,\boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj} \\
&+ \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\widehat{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\boldsymbol{x}'_{dtj}\mathrm{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj} \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \Big(\frac{c_\phi}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} E\big[\mathrm{sgn}(e_{\psi,dtj})\big] + \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j\in\mathcal{G}_{dt}} E[R_{dtj}]\Big)^2 + o(1).
\end{aligned}$$

Proof. The proof, by Bugallo et al. (2024e), is reported in Appendix D.1. □

The estimator of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$ is given by

$$\begin{aligned}
mse_{3,dt}^{btmq} &= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\widehat{\xi}_{dt}^2}{n_{dt}^2}\Big(\frac{1}{\mathrm{card}(\widehat{\mathcal{G}}_{dt})}\sum_{j\in\widehat{\mathcal{G}}_{dt}}(\widehat{q}_{dtj}-\widehat{\theta}_d)^2\Big)^{-1}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{e}_{\psi,dtj}^2 + \\
&+ \frac{N_{dt}-n_{dt}}{n_{dt}}\frac{\widehat{\xi}_{dt}^2}{N_{dt}^2}\Big(\frac{1}{n_{dt}}\sum_{j\in s_{dt}}(\widehat{q}_{dtj}-\widehat{\theta}_d)^2\Big)^{-1}\sum_{j\in s_{dt}}\widehat{e}_{\psi,dtj}^2 \\
&+ 2\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2\frac{1}{n_{dt}^2}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\boldsymbol{x}'_{dtj}\widehat{V}_\beta\boldsymbol{x}_{dtj} + \frac{1}{N_{dt}^2}\sum_{j\in r_{dt}}\boldsymbol{x}'_{dtj}\widehat{V}_\beta\boldsymbol{x}_{dtj} \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2\frac{1}{n_{dt}^2}\Big(c_\phi\sum_{j\in s_{dt}-\widehat{\mathcal{G}}_{dt}}\mathrm{sgn}(\widehat{e}_{\psi,dtj}) + \frac{1}{2\sigma_{\theta_d t}}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{e}_{\psi,dtj}^2\Big)^2, \qquad (5.27)
\end{aligned}$$

where $\widehat{V}_\beta$ is given in (5.23), $\widehat{\mathcal{G}}_{dt} = \Big\{j \in s_{dt} : |\widehat{u}_{\psi,dtj}| < c_\phi\Big\}$, expression $\mathrm{card}(B)$ denotes the cardinal of a set B and $\widehat{\xi}_{dt}^2$ estimates $\mathrm{var}(q_{dtj})$, e.g. by

$$\widehat{\xi}_{dt}^2 = \widehat{\mathrm{var}}(q_{dtj}) = \frac{1}{n_{dt}-1}\sum_{j\in s_{dt}}(\widehat{q}_{dtj}-\widehat{\bar{q}}_{dt.})^2, \quad \widehat{\bar{q}}_{dt.} = \frac{1}{n_{dt}}\sum_{j\in s_{dt}}\widehat{q}_{dtj}.$$

Finally, an estimator of $RMSE(\widehat{\overline{Y}}_{dt}^{btmq}) = \big(MSE(\widehat{\overline{Y}}_{dt}^{btmq})\big)^{1/2}$ is $rmse_{3,dt}^{btmq} = \big(mse_{3,dt}^{btmq}\big)^{1/2}$.

Two alternative estimators of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$ could be also derived. It should be noted that, from equation (5.16), we have $\widehat{u}_{\psi,dtj} - \breve{u}_{\psi,dtj} = \sigma_{\theta_d t}^{-1}\boldsymbol{x}'_{dtj}(\boldsymbol{\beta}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t)) = \widehat{u}_{\psi,dtj}(\widehat{\theta}_d)$. A Taylor series expansion of $\phi(\widehat{u}_{\psi,dtj})$ around $\phi(\breve{u}_{\psi,dtj})$ yields to

$$\phi(\widehat{u}_{\psi,dtj}) \approx \phi(\breve{u}_{\psi,dtj}) + \dot{\phi}(\breve{u}_{\psi,dtj})\frac{\boldsymbol{x}'_{dtj}(\boldsymbol{\beta}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t))}{\sigma_{\theta_d t}}, \; j = 1\ldots, n_{dt}.$$

Let us define:

$$\bar{\boldsymbol{x}}'_{\breve{\mathcal{G}}_{dt}} = \frac{1}{\text{card}(\breve{\mathcal{G}}_{dt})} \sum_{j \in \breve{\mathcal{G}}_{dt}} \boldsymbol{x}'_{dtj}, \quad \bar{\boldsymbol{x}}'_{sdt} = \frac{1}{n_{dt}} \sum_{j \in s_{dt}} x'_{dtj}, \quad \bar{e}_{rdt} = \bar{y}_{rdt} - \bar{\boldsymbol{x}}'_{rdt} \boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t).$$

If ϕ is the Huber function, $\dot{\phi}(\breve{u}_{\psi,dtj}) = 1$ if $j \in \breve{\mathcal{G}}_{dt} = \left\{ j \in s_{dt} : \left|\breve{u}_{\psi,dtj}\right| < c_\phi \right\}$ and $\dot{\phi}(\breve{u}_{\psi,dtj}) = 0$, otherwise. In a similar vein to the proposal by Chambers et al. (2014a), it holds that

$$\begin{aligned}
\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi(\widehat{u}_{\psi,dtj}) &\approx \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi(\breve{u}_{\psi,dtj}) + \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in \breve{\mathcal{G}}_{dt}} \frac{\boldsymbol{x}'_{dtj}(\boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t))}{\sigma_{\theta_d t}} \\
&= \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi(\breve{u}_{\psi,dtj}) + \frac{\text{card}(\breve{\mathcal{G}}_{dt})}{n_{dt}} \bar{\boldsymbol{x}}'_{\breve{\mathcal{G}}_{dt}}(\boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)) \\
&\approx \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi(\breve{u}_{\psi,dtj}) + \bar{\boldsymbol{x}}'_{sdt}(\boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)).
\end{aligned}$$

The prediction error of $\widehat{\overline{Y}}_{dt}^{btmq}$ is approximated as

$$\begin{aligned}
\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt} &= \frac{1}{N_{dt}} \left(\sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \sum_{j \in r_{dt}} y_{dtj} \right) + \frac{1}{n_{dt}} \left(1 - \frac{n_{dt}}{N_{dt}}\right) \widehat{B}_{dt}^{btmq} \\
&\approx \left(1 - \frac{n_{dt}}{N_{dt}}\right) \left(\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi\big(\breve{u}_{\psi,dtj}\big) + \bar{\boldsymbol{x}}'_{sdt} \boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) + (\bar{\boldsymbol{x}}_{rdt} - \bar{\boldsymbol{x}}_{sdt})' \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \bar{y}_{rdt} \right) \\
&= \left(1 - \frac{n_{dt}}{N_{dt}}\right) \left(\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi\big(\breve{u}_{\psi,dtj}\big) + (\bar{\boldsymbol{x}}_{rdt} - \bar{\boldsymbol{x}}_{sdt})' (\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t) - \boldsymbol{\beta}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)) - \bar{e}_{rdt} \right).
\end{aligned}$$

An estimator of the variance $V_{dt}^{btmq} = \text{var}\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big)$ is

$$\widehat{V}_{dt}^{btmq} = \left(1 - \frac{n_{dt}}{N_{dt}}\right)^2 \left[\left(\frac{\sigma_{\theta_d t}}{n_{dt}}\right)^2 \sum_{j \in s_{dt}} \phi^2\big(\widehat{u}_{\psi,dtj}\big) + (\bar{\boldsymbol{x}}_{rdt} - \bar{\boldsymbol{x}}_{sdt})' \widehat{V}_\beta (\bar{\boldsymbol{x}}_{rdt} - \bar{\boldsymbol{x}}_{sdt}) + \widehat{\text{var}}(\bar{e}_{rdt}) \right],$$

where the estimation of the variance matrix $\widehat{V}_\beta$ is given in (5.23) and $\text{var}(\bar{e}_{r_{dt}})$ is estimated using $\widehat{\text{var}}_1\big(\bar{y}_{rdt}\big)$ or $\widehat{\text{var}}_2\big(\bar{y}_{rdt}\big)$, given in (5.24). Depending on which formula we choose, we obtain the estimators $\widehat{V}_{21,dt}^{btmq}$ and $\widehat{V}_{22,dt}^{btmq}$ of the variance V_{dt}^{btmq}, respectively.

Against this background, note that the bias correction in (5.20) is controlled by c_ϕ, which can be either very large or close to zero. Consequently, it is not advisable to ignore the presence of a potential bias, as it may (and perhaps should) still persist. Using the bias estimator $\widehat{B}_{dt} = \widehat{B}\big(\widehat{\overline{Y}}_{dt}^{tmq}\big)$, given in (5.25), two estimators of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$ are

$$mse_{1,dt}^{btmq} = \widehat{V}_{21,dt}^{btmq} + \left(\widehat{B}_{dt} + \frac{1}{n_{dt}} \left(1 - \frac{n_{dt}}{N_{dt}}\right) \widehat{B}_{dt}^{btmq} \right)^2$$

and

$$mse_{2,dt}^{btmq} = \widehat{V}_{22,dt}^{btmq} + \left(\widehat{B}_{dt} + \frac{1}{n_{dt}} \left(1 - \frac{n_{dt}}{N_{dt}}\right) \widehat{B}_{dt}^{btmq} \right)^2.$$

By taking squared roots of the above estimators, we have $rmse_{1,dt}^{btmq}$ and $rmse_{2,dt}^{btmq}$, respectively. They are estimators of $RMSE(\widehat{\overline{Y}}_{dt}^{btmq}) = \left(MSE(\widehat{\overline{Y}}_{dt}^{btmq})\right)^{1/2}$.

5.4.4 Selection of the robustness parameter

The robustness parameter is critical in determining the improvements of the BTMQ predictor over the TMQ predictor. However, the selection of an optimal robustness parameter remains an open question for bias-corrected predictors (Chambers et al., 2014a; Dongmo-Jiongo et al., 2013). In the context of MQ predictors, a common practice is to set $c_\phi = 3$ (Salvati et al., 2012; Tzavidis et al., 2010). While this choice has yielded promising results, it is ultimately subjective. An alternative to determine an appropriate value for the robustness parameter could be to minimize an estimator of the MSE (Beaumont et al., 2013).

Thus, to strike a balance between bias and variance, we are looking for the values of $c_\phi \geqslant 0$ that minimize $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$. In practice, this means solving the minimization problems

$$\widehat{c}_{\phi,dt} = \underset{c_\phi \geqslant 0}{\operatorname{argmin}}, mse_{dt}^{btmq}(c_\phi), \quad d = 1, \ldots, D, \ t = 1, \ldots, T, \tag{5.28}$$

where mse_{dt}^{btmq} is an estimation of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$. Considering that the solutions are adaptive robustness parameters for each area and time period, calculable e.g. by using grid search methods, we call them area-time specific robustness parameters.

The following theorem states the existence and uniqueness of solutions of the minimization problems (5.28) for $mse_{dt}^{btmq} \in \{mse_{1,dt}^{btmq}, mse_{2,dt}^{btmq}\}$ (Bugallo et al., 2024e).

Theorem 2. Let ϕ be the Huber function, defined in (5.3). Let $d = 1, \ldots, D$, $t = 1, \ldots, T$. For $mse_{dt}^{btmq} \in \{mse_{1,dt}^{btmq}, mse_{2,dt}^{btmq}\}$, it exists an unique solution $\widehat{c}_{\phi,dt}$ of the minimization problem (5.28) belonging to the interval $\left[0, \max_{j \in s_{dt}} |\widehat{u}_{\psi,dtj}|\right]$ and its explicit expression is calculable.

Proof. The proof is reported in Appendix D.2. □

5.5 Model-based simulations

The experimental design is inspired by Chambers et al. (2014a) and Bugallo et al. (2024e), but here we have also included a time reference to all observations. The outline of the simulations is described below, including the scenarios for the incorporation of unit-level and area-level outliers, the cases for the time dependency random effects and the number of iterations, $S = 500$. Population data are generated for $D = 40$ areas and $T = 10$ time periods. From each population, the sample data have been selected by simple random sampling without replacement within each intersection between areas and time periods. Population and sample sizes are fixed at $N_{dt} = 100$ and $n_{dt} = 5$, respectively. Both the MQ3 and TWMQ linear models are fitted using the IRLS algorithm. In addition, the projective influence function ψ is the Huber function with tuning constant $c_\psi = 1.345$, the same as the function ϕ of the

robust bias-corrected MQ predictors BMQ and BTMQ, but their tuning constant has been selected area-time specific. So as to calculate $\widehat{c}^{(s)}_{\phi,dt}$, $s = 1, \ldots, S$, we use a fine grid from 0 to 10, with evenly spaced breaks of 0.001 width. For the fitting of the MQ models, the prediction of small area linear indicators, the estimation of the MSE and the selection of the robustness parameters, we have used a code developed by the authors. Finally, the LMMs are fitted using REML. The `R` library `nlme` (Pinheiro and Bates, 2023, 2000) has been used for this purpose and, in particular, the `lme` function.

Simulations 1 and 2 have the following steps:

1. Define $\beta_1 = 100$ and $\beta_2 = 5$. Vary q on a fine grid $G \subset [0, 1]$.

→ Choose Scenario [0,0], [e,0] or [e,u], where

[0,0] – absence of outliers, $u_{1,d} \backsim N(0, 3)$ and $e_{dtj} \backsim N(0, 6)$;

[e,0] – only individual level outliers, $u_{1,d} \backsim N(0, 3)$ and $e_{dtj} \backsim \delta N(0, 6)+(1-\delta)N(20, 150)$, where δ is an independently generated BE random variable with $P(\delta = 1) = 0.97$;

[e,u] – outliers affect both area and individual effects, $u_{1,d} \backsim N(0, 3)$ for areas $1 \leqslant d \leqslant 36$, $u_{1,d} \backsim N(9, 20)$ for areas $37 \leqslant d \leqslant 40$, and $e_{dtj} \backsim \delta N(0, 6) + (1 - \delta)N(20, 150)$.

→ Choose Case 1.1, 1.2 or 2, where

Case 1. $\boldsymbol{u}_2 = \underset{1 \leqslant t \leqslant T}{\text{col}}(u_{2,t}) \backsim N_T(\mathbf{0}, \Sigma_u)$, $\Sigma_u = \sigma_u^2 \Omega_T(\rho)$, and the correlation matrix is

$$\Omega_T(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & \ddots & \rho^{T-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho & 1 \end{pmatrix} \in \mathcal{M}_{T \times T}, \quad \rho \in (-1, 1). \tag{5.29}$$

Case 1.1: $\sigma_u = 1$, $\rho = 0.2$; and Case 1.2: $\sigma_u = 1$, $\rho = 0.8$.

Case 2. Each $u_{2,t}$ is independently generated according to a stationary $AR(3)$ model with coefficients $\phi_1 = 0.4$, $\phi_2 = 0.3$, $\phi_3 = 0.25$ and white noise variance $\sigma = 1$.

2. Repeat $S = 500$ times ($s = 1, \ldots, S$):

   (a) For $d = 1 \ldots, D$, $t = 1 \ldots, T$, $j \in U_{dt}$, generate $x^{(s)}_{dtj} \backsim \text{LogN}(1, 0.5)$, $u^{(s)}_{1,d}$ and $e^{(s)}_{dtj}$ depending on the chosen scenario, and $u^{(s)}_{2,t}$ depending on the chosen case; and

$$y^{(s)}_{dtj} = \beta_1 + x^{(s)}_{dtj}\beta_2 + u^{(s)}_{1,d} + u^{(s)}_{2,t} + e^{(s)}_{dtj}, \quad \overline{Y}^{(s)}_{dt} = \frac{1}{N_{dt}} \sum_{j=1}^{N_{dt}} y^{(s)}_{dtj}.$$

   (b) Fit the MQ3 linear models using the population data. Calculate $q^{(s)}_{dtj}$ and then $\theta^{(s)}_d$, $d = 1 \ldots, D$, $t = 1 \ldots, T$, $j \in U_{dt}$. Use the IRLS algorithm.

   (c) Randomly generating n_{dt} different positions between 1 and N_{dt}, draw a sample $s^{(s)}_{dt}$ of size n_{dt}, $d = 1 \ldots, D$, $t = 1 \ldots, T$. In what follows, only sample data are used.

(d) Calculate the Hájek estimator with equal weights, i.e. the arithmetic mean:

$$\widehat{\overline{Y}}_{dt}^{hajek} = \frac{1}{n_{dt}} \sum_{j=1}^{n_{dt}} y_{dtj}.$$

(e) Using REML, fit the area-level LMM_1 model

$$y_{dtj} = \beta_1 + x_{dtj}\beta_2 + u_{1,d} + e_{dtj},\ u_{1,d} \backsim N(0, \sigma^2_{u_1}),\ e_{dtj} \backsim N(0, \sigma^2_e),\ \ \sigma^2_{u_1}, \sigma^2_e > 0,$$

and the area-level and time-level LMM_2 model

$$y_{dtj} = \beta_1 + x_{dtj}\beta_2 + u_{1,d} + u_{2,t} + e_{dtj},\ u_{1,d} \backsim N(0, \sigma^2_{u_1}),\ u_{2,t} \backsim N(0, \sigma^2_{u_2}),$$
$$e_{dtj} \backsim N(0, \sigma^2_e),\ \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_e > 0,$$

where β_1 and β_2 are the corresponding model parameters; $u_{1,d}$ are the area-level random intercepts of LMM_1; $u_{1,d}$ and $u_{2,t}$ are the area-level and time-level random intercepts of LMM_2, respectively; and e_{dtj} are the corresponding model errors.

(f) Calculate the predictors $\widehat{\overline{Y}}_{dt}^{eblup_1}$ and $\widehat{\overline{Y}}_{dt}^{eblup_2}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$, given by

$$\widehat{\overline{Y}}_{dt}^{eblup_1} = \frac{1}{N_{dt}} \Big\{ \sum_{j \in s_{dt}^{(s)}} y_{dtj} + \sum_{j \in r_{dt}^{(s)}} (\widehat{\beta}_1 + x_{dtj}\widehat{\beta}_2 + \widehat{u}_{1,d}) \Big\},$$
$$\widehat{\overline{Y}}_{dt}^{eblup_2} = \frac{1}{N_{dt}} \Big\{ \sum_{j \in s_{dt}^{(s)}} y_{dtj} + \sum_{j \in r_{dt}^{(s)}} (\widehat{\beta}_1 + x_{dtj}\widehat{\beta}_2 + \widehat{u}_{2,d} + \widehat{u}_{2,t}) \Big\},$$

where $r_{dt}^{(s)} = U_{dt} - s_{dt}^{(s)}$ is the non sampled subset of U_{dt}; $\widehat{u}_{1,d}$ is the EBLUP of the random intercept $u_{1,d}$ for LMM_1; and $\widehat{u}_{1,d}$ and $\widehat{u}_{2,t}$ are the EBLUPs of the random intercepts $u_{1,d}$ and $u_{2,t}$ for LMM_2, respectively.

(g) Fit the MQ3 linear model, i.e. estimate $\widehat{\boldsymbol{\beta}}_{\psi}^{(s)}(q)$ and $\widehat{\sigma}_q^{(s)} = \widehat{\sigma}_{\psi}(\widehat{\boldsymbol{\beta}}_{\psi}^{(s)}(q))$, with $q \in G$.

(h) For $d = 1 \ldots, D$, $t = 1 \ldots, T$, $j \in s_{dt}^{(s)}$, estimate $\widehat{q}_{dtj}^{(s)}$ and then $\widehat{\theta}_d^{(s)}$.

(i) Calculate the predictors $\widehat{\overline{Y}}_{dt}^{mq}$ and $\widehat{\overline{Y}}_{dt}^{bmq}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$.

(j) Calculate the inter-period weights $\boldsymbol{w}_t^{(s)} = (w_{t1}^{(s)}, \ldots, w_{tT}^{(s)})$, $t = 1, \ldots, T$.

(k) For $t = 1, \ldots, T$, $d = 1, \ldots, D$, fit the TWMQ linear models with $q = \widehat{\theta}_d^{(s)}$.

(l) Calculate the predictors $\widehat{\overline{Y}}_{dt}^{tmq}$ and $\widehat{\overline{Y}}_{dt}^{btmq}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$.

(m) For the TMQ predictor, calculate

$$rmse_{dt}^{tmq} \in \{rmse_{11,dt}^{tmq}, rmse_{12,dt}^{tmq}, rmse_{21,dt}^{tmq}, rmse_{22,dt}^{tmq}\}$$

and for the BTMQ predictor, calculate

$$rmse_{dt}^{btmq} \in \{rmse_{1,dt}^{btmq}, rmse_{2,dt}^{btmq}, rmse_{3,dt}^{btmq}\},$$

$d = 1, \ldots, D$, $t = 1, \ldots, T$.

3. For $\widehat{\overline{Y}}_{dt} \in \{\widehat{\overline{Y}}^{hajek}_{dt}, \widehat{\overline{Y}}^{eblup_1}_{dt}, \widehat{\overline{Y}}^{eblup_2}_{dt}, \widehat{\overline{Y}}^{mq}_{dt}, \widehat{\overline{Y}}^{bmq}_{dt}, \widehat{\overline{Y}}^{tmq}_{dt}, \widehat{\overline{Y}}^{btmq}_{dt}\}$, calculate

$$\mathrm{BIAS}_{1,dt} = \frac{1}{S}\sum_{s=1}^{S}\big(\widehat{\overline{Y}}^{(s)}_{dt} - \overline{Y}^{(s)}_{dt}\big), \quad \mathrm{RMSE}_{1,dt} = \Big(\frac{1}{S}\sum_{s=1}^{S}\big(\widehat{\overline{Y}}^{(s)}_{dt} - \overline{Y}^{(s)}_{dt}\big)^2\Big)^{1/2}. \tag{5.30}$$

For $rmse_{dt} \in \{rmse^{tmq}_{dt}, rmse^{btmq}_{dt}\}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$, calculate

$$\mathrm{BIAS}_{2,dt} = \frac{1}{S}\sum_{s=1}^{S}(rmse^{(s)}_{dt} - \mathrm{RMSE}_{1,dt}),\ \mathrm{RMSE}_{2,dt} = \Big(\frac{1}{S}\sum_{s=1}^{S}\big(rmse^{(s)}_{dt} - \mathrm{RMSE}_{1,dt}\big)^2\Big)^{1/2},$$

where $\mathrm{RMSE}_{1,dt}$ is taken from (5.30) for $\widehat{\overline{Y}}_{dt} \in \{\widehat{\overline{Y}}^{tmq}_{dt}, \widehat{\overline{Y}}^{btmq}_{dt}\}$.

Consistent with the later notation, write

$$\mathrm{RMSE}_{1,dt} \in \{\mathrm{RMSE}^{tmq}_{dt}, \mathrm{RMSE}^{btmq}_{dt}\},$$

$d = 1, \ldots, D$, $t = 1, \ldots, T$, and

$$\mathrm{RMSE}_1 = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\mathrm{RMSE}_{1,dt} \in \{\mathrm{RMSE}^{tmq}, \mathrm{RMSE}^{btmq}\}.$$

4. For $d = 1, \ldots, D$, $t = 1, \ldots, T$, calculate the relative performance measures

$$\mathrm{RBIAS}_{1,dt} = \frac{100 \cdot \mathrm{BIAS}_{1,dt}}{\overline{Y}^{*}_{dt}}, \quad \mathrm{RRMSE}_{1,dt} = \frac{100 \cdot \mathrm{RMSE}_{1,dt}}{\overline{Y}^{*}_{dt}}, \quad \overline{Y}^{*}_{dt} = \frac{1}{S}\sum_{s=1}^{S}\overline{Y}^{(s)}_{dt},$$
$$\mathrm{RBIAS}_{2,dt} = \frac{100 \cdot \mathrm{BIAS}_{2,dt}}{\mathrm{RMSE}_{1,dt}}, \quad \mathrm{RRMSE}_{2,dt} = \frac{100 \cdot \mathrm{RMSE}_{2,dt}}{\mathrm{RMSE}_{1,dt}},$$

and the average relative performance measures

$$\mathrm{ARBIAS}_l = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|\mathrm{RBIAS}_{l,dt}|,\ \mathrm{RRMSE}_l = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\mathrm{RRMSE}_{l,dt},\ l = 1, 2.$$

Finally, to measure overestimates (underestimates) of the several methods of MSE estimation, we define the proportion of subdomains in which the proposed estimates are higher (lower) than the empirical values. In line with the above, let be

$$P_+ = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T} I(\mathrm{BIAS}_{2,dt} \geqslant 0), \quad P_- = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T} I(\mathrm{BIAS}_{2,dt} < 0) = DT - P_+.$$

5.5.1 Simulation 1

In this section we assess the performance of the TMQ and BTMQ predictors from the TWMQ linear models in estimating the domain means and compare them with others taken

from the literature. We also show how the selection of area-time-specific robustness parameters could be employed as a diagnostic tool for outlier detection. Table 5.1 presents the performance measures for Case 1.1, Case 1.2 and Case 2, and the simulation scenarios and predictors mentioned above. Top performers are highlighted in bold.

	[0,0] ARBIAS	[0,0] 1–40 RRMSE	[e,0] ARBIAS	[e,0] 1–40 RRMSE	[e,u] ARBIAS	[e,u] 1–40 RRMSE
Case 1.1	$\boldsymbol{u}_2 \backsim N_T(\mathbf{0}, \Sigma_u)$:		$\Sigma_u = \sigma_u^2 \Omega_T(\rho)$,		$\sigma_u = 1$,	$\rho = 0.2$
$\widehat{\overline{Y}}_{dt}^{hajek}$	0.117	3.223	0.133	3.549	0.132	3.523
$\widehat{\overline{Y}}_{dt}^{eblup_1}$	0.039	0.833	0.047	0.992	0.067	0.999
$\widehat{\overline{Y}}_{dt}^{mq}$	0.041	0.956	0.416	1.105	0.400	1.093
$\widehat{\overline{Y}}_{dt}^{bmq}$	0.036	0.794	0.410	0.964	0.406	0.973
$\widehat{\overline{Y}}_{dt}^{eblup_2}$	0.027	0.655	**0.042**	0.940	**0.064**	0.947
$\widehat{\overline{Y}}_{dt}^{tmq}$	0.023	0.701	0.415	0.864	0.398	0.884
$\widehat{\overline{Y}}_{dt}^{btmq}$	**0.019**	**0.553**	0.409	**0.752**	0.396	**0.795**
Case 1.2	$\boldsymbol{u}_2 \backsim N_T(\mathbf{0}, \Sigma_u)$:		$\Sigma_u = \sigma_u^2 \Omega_T(\rho)$,		$\sigma_u = 1$,	$\rho = 0.8$
$\widehat{\overline{Y}}_{dt}^{hajek}$	0.117	3.223	0.132	3.548	0.132	3.523
$\widehat{\overline{Y}}_{dt}^{eblup_1}$	0.039	0.954	0.046	1.096	0.066	1.101
$\widehat{\overline{Y}}_{dt}^{mq}$	0.041	1.066	0.414	1.200	0.399	1.191
$\widehat{\overline{Y}}_{dt}^{bmq}$	0.035	0.897	0.407	1.051	0.403	1.053
$\widehat{\overline{Y}}_{dt}^{eblup_2}$	0.027	0.682	**0.042**	0.995	**0.063**	1.001
$\widehat{\overline{Y}}_{dt}^{tmq}$	0.025	0.693	0.413	0.878	0.397	0.897
$\widehat{\overline{Y}}_{dt}^{btmq}$	**0.020**	**0.539**	0.407	**0.760**	0.396	**0.800**
Case 2	$u_{2,t} \backsim AR(3)$:		$\phi_1 = 0.4$,	$\phi_2 = 0.3$,	$\phi_3 = 0.25$,	$\sigma = 1$
$\widehat{\overline{Y}}_{dt}^{hajek}$	0.115	3.238	0.122	3.571	0.121	3.545
$\widehat{\overline{Y}}_{dt}^{eblup_1}$	0.021	0.838	0.031	1.001	0.064	1.007
$\widehat{\overline{Y}}_{dt}^{mq}$	0.027	0.962	0.414	1.114	0.396	1.101
$\widehat{\overline{Y}}_{dt}^{bmq}$	0.022	0.800	0.408	0.971	0.402	0.979
$\widehat{\overline{Y}}_{dt}^{eblup_2}$	0.018	0.652	**0.030**	0.945	**0.062**	0.951
$\widehat{\overline{Y}}_{dt}^{tmq}$	0.022	0.698	0.416	0.874	0.394	0.892
$\widehat{\overline{Y}}_{dt}^{btmq}$	**0.017**	**0.548**	0.409	**0.758**	0.394	**0.800**

Table 5.1: Assessment of the absolute performance of the predictors of small area population means by calculating average measures. ARBIAS and RRMSE values (in %).

First, if time effects are normal, as defined in Case 1, the improvement of the TWMQ linear models is, as expected, subject to the definition of the covariance matrix. These models do not aim to capture random effects over time, but rather well-founded relationships of time dependency. Provided that the normality assumptions are met and there are no outliers, the

values of ρ and σ^2 are crucial. If the variance is large enough, the LMM_2 provides acceptable ARBIAS and RRMSE values. If the correlation is higher and/or the variance is lower, the latter is not achieved, and the best options are the TWMQ linear models. The same is true for the presence of outliers. It should be mentioned, however, that the LMM_2 is not the model that generates the target variables as it does not take into account the correlation structure of the time-level random intercepts. As far as we are aware, the available correlation structures in `R` refer to the model errors e_{dtj}, and not to the random effects $u_{2,t}$. In addition, if the time effects follow an autoregressive model, as defined in Case 2, the TWMQ linear models are better in terms of RRMSE, although not in terms of ARBIAS if there are area-level outliers. In fact, they are expected to be robust in the presence of atypical data, inheriting the well-known robustness properties of the MQ regression. In that case, the ARBIAS of the EBLUP_2 is smaller and the main source of contribution to its RRMSE comes from the variance.

Regarding the TMQ and BTMQ predictors, a proper selection of c_ϕ ensures that the latter is much better than the former, correcting for bias but also mitigating variability. The same applies to the comparison of predictors MQ and BMQ. However, the flexibility of the TWMQ linear models leads to less bias correction without severely affecting the variance. Rather than merely accepting a default value as adequate, area-time specific values of c_ϕ are far preferable and very promising, being able to outperform the EBLUP_2 based on the LMM_2. Moreover, an unexpected advantage is that the set $\left\{\widehat{c}_{\phi,dt} : \ d = 1,\ldots,D,\ t = 1,\ldots,T\right\}$ can be used as a diagnostic tool for outlier detection (see below) and, not least important, computational effort and execution times are not affected either if we use optimum values of c_ϕ. The BTMQ predictor is, indeed, a plug-in type predictor, so it is easy to program and fast to calculate.

Moving on to our second contribution, we will illustrate how the area-time-specific robustness parameters for bias correction is used for outlier detection. The discussion focuses on Case 1.1, although the conclusions are similar for the remaining two cases. Let us define the average value, across simulations, of $\widehat{c}_{\phi,dt}$, given by $\overline{\widehat{c}}_{\phi,dt} = \frac{1}{S}\sum_{s=1}^{S}\widehat{c}^{(s)}_{\phi,dt}$.

First of all, some basic descriptive measures are calculated. For Scenario [0,0], $\overline{\widehat{c}}_{\phi,dt}$ ranges from 0.380 to 0.624, with a median value of 0.491. Something similar happens for Scenario [e,0], where $\overline{\widehat{c}}_{\phi,dt}$ ranges from 0.365 to 0.594, and the median is 0.474. For Scenario [e,u] and the non-atypical areas $1 \leqslant d \leqslant 36$, $\overline{\widehat{c}}_{\phi,dt}$ ranges from 0.338 to 0.492, and the median is 0.433, but ranges from 1.736 to 2.075, with a median of 1.887, for the atypical areas $37 \leqslant d \leqslant 40$. Bearing all of the above in mind, $\overline{\widehat{c}}_{\phi,dt}$ is unaffected by the presence of unit-level outliers but a very different picture emerges with the presence of area-level outliers. The latter can be observed in Figure 5.1, where the peaks corresponding to areas $37 \leqslant d \leqslant 40$ are quite revealing. Compared to Scenario [0,0], the presence of individual-level outliers stops the improvement in bias earlier, to avoid overfitting. If outliers are also at the area level, as in Scenario [e,u], it is possible to give more strength to the bias correction without an increase of the variance.

After reviewing the literature, the results are supported by statistical tests adapted to our problem. First, Friedman's test (Friedman, 1937) is used for one-way repeated measures analysis of variance under different experimental conditions. The repeated measures are the values of $\overline{\widehat{c}}_{\phi,dt}$ selected for an area d, $d = 1,\ldots,D$, over all time periods, $t = 1,\ldots,T$ (which are the experimental conditions). It is assumed that each $\overline{\widehat{c}}_{\phi,dt}$ is equally distributed except at most in terms of location, which may vary according to the experimental condition and

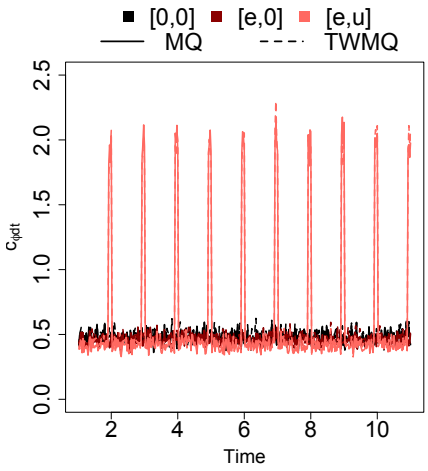


Figure 5.1: Assessment of the area-time specific robustness parameters for the detection of atypical domains. Scartterplot of $\overline{\widehat{c}}_{\phi,dt}$ sorted by area and time period for Case 1.1.

area. We write

$$\overline{\widehat{c}}_{\phi,dt} = c + \delta_{1,t} + \delta_{2,d} + e_{dt},$$

where c is the global mean, independent of both area and time; $\delta_{1,t}$ is the average effect of time period t, $t = 1, \ldots, T$; $\delta_{2,d}$ measures the average effect of the d-th area, $d = 1, \ldots, D$; and each e_{dt} i.i.d. follows an unknown zero-mean distribution F. The objective is to test

$$H_0 : \delta_{2,1} = \cdots = \delta_{2,D} \quad \text{vs} \quad H_1 : \exists\ d_1, d_2 \in \{1, \ldots, D\},\ d_1 \neq d_2,\ \delta_{2,d_1}, \neq \delta_{2,d_2}.$$

In a second step, we apply an Honestly Significant Difference test, or Tukey's multiple range test (Tukey, 1949), to detect which groups of areas shift in scale. If appropriate, both tests can be applied to time periods, which leads us to write

$$H_0 : \delta_{1,1} = \cdots = \delta_{1,t} \quad \text{vs} \quad H_1 : \exists\ t_1, t_1 \in \{1, \ldots, T\},\ t_1 \neq t_2,\ \delta_{1,t_1} \neq \delta_{1,t_2},$$

where the repeated measures are the values of $\overline{\widehat{c}}_{\phi,dt}$ selected for a time period t, $t = 1, \ldots, T$, over all areas, $d = 1, \ldots, D$ (which are the experimental conditions).

Our approach for outlier detection reports the following promising results. In Scenario [e,u], the mean of $\overline{\widehat{c}}_{\phi,dt}$ differs between areas (p-value $\simeq 0$) but not between time periods (p-value 0.356). As expected, the area-level outliers detected by Tukey's test are exactly those with index $37 \leqslant d \leqslant 40$. The same test applied to time periods detects only a single group, formed by all of them. Having said that, note that we have used the averaged values of $\overline{\widehat{c}}_{\phi,dt}$ to assess the performance of the proposed approach in future applications to real data. As $S = 500$, this is done to avoid randomness in the data generation process. For the sake of completeness, we have also applied Friedman's test to each set of selections $\left\{\widehat{c}^{(s)}_{\phi,dt} : \ d = 1, \ldots, D,\ t = 1, \ldots, T\right\}$, $s = 1, \ldots, S$. With a significance level of 1%, the equality of provincial means is correctly rejected in 100% of the samples; and the equality of temporal means is incorrectly rejected (i) at 1% significance in 0.8% of the samples, (ii) at 5% significance in 6% of the samples. The sensitivity and power of the results are more than acceptable.

From the perspective of LMMs, the detection of atypical data using the methodology described in Zewotir and Galpin (2007) does not report conclusive results because it is performed at unit-level. In particular, unit-level outliers have been found to have more impact than area-level outliers on the results of the test proposed by these authors. In our research, the generation of unit-level outliers has been random, i.e. as additional noise, making the above-mentioned test useless. In addition, it can be seen in Table 5.1 how LMMs overfit the atypical data, reducing the bias but excessively increasing the variance, so it is not accurate at all in analyzing the predicted random effects. In this sense, outliers severely affect non-robust models in the context of SAE. Indeed, it is straightforward that an atypical value that destabilizes a population estimate based on a large sample survey will greatly affect the results obtained from a small collection of data (Chambers and Tzavidis, 2006; Koenker, 2005).

5.5.2 Simulation 2

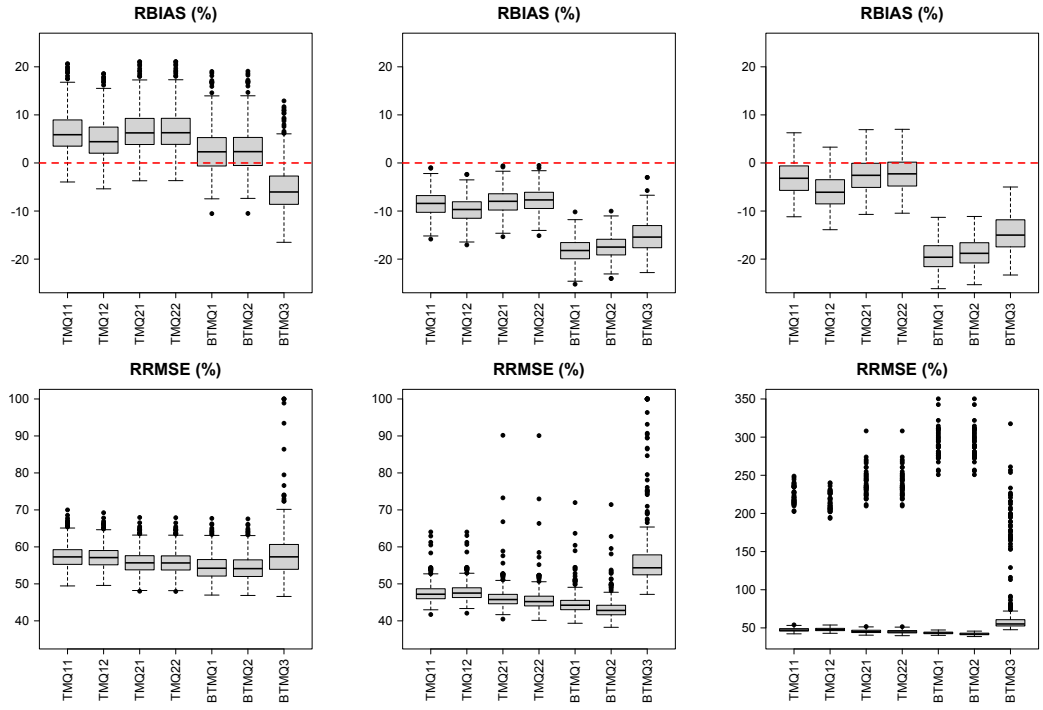
In this section we investigate the performance of several methods of MSE estimation for the TMQ and BTMQ predictors. First and foremost, Table 5.2 presents the performance measures for Case 1.1, Case 1.2 and Case 2, and the different scenarios for the generation of atypical data. We have included a third column with the proportion of subdomains in which the bias is positive and, therefore, the RMSE is overestimated.

As a first general comment, estimating RMSEs is much more difficult than estimating small area linear indicators, such as population means. Therefore, the magnitude of the results in Table 5.2 should be assessed with caution. Although it is suggested that the RMSE estimators for the TMQ predictors offer the most balanced performance in terms of ARBIAS and RRMSE, the empirical RMSEs of the BTMQ predictors are smaller. As for the sign of the RBIAS of the RMSE estimators, there are overestimates in Scenario [0,0] and underestimates in the other two scenarios, being plotting a more intuitive tool. In addition, the difference between cases for the generation of time effects is of little importance. Having said that, Figure 5.2 shows boxplots of RBIAS and RRMSE, both %, for the RMSE estimators performance in Case 1.1 and the three scenarios already considered. First, the RBIAS of the RMSE estimators of the TMQ predictors is more positive (or less negative, as appropriate) than the corresponding one for the BTMQ predictors. It can be seen how the values of column P_+ in Table 5.2 are in line with the boxplots for the bias in Figure 5.2. In terms of RRMSE, the presence of area-level outliers greatly worsens the results in these subdomains, and thus the average values in Table 5.2.

In general terms, although $rmse_{3,dt}^{btmq}$ is calculated from a first-order unbiased approximation, we feel that the numerical instability problems involved in the estimation stage are responsible for its "slightly worse" performance. As discussed in Section D.1.6, the theoretical advantage of this estimator is largely overshadowed by the highly unstable estimation of one of its variance terms. Taking into account the latter, in the application to real data in Section 5.7 we will use $rmse_{2,dt}^{btmq}$ to present error measures about the BTMQ predictor. It should be noted that the results for $rmse_{1,dt}^{btmq}$ are almost the same but slightly worse. Finally, as far as the TMQ predictor is concerned, we propose using $rmse_{2,dt}^{tmq}$.

	[0,0] 1–40			[e,0] 1–40			[e,u] 1–40		
	ARBIAS	RRMSE	P_+	ARBIAS	RRMSE	P_+	ARBIAS	RRMSE	P_+
Case 1.1	$\boldsymbol{u}_2 \backsim N_T(\mathbf{0}, \Sigma_u)$: $\Sigma_u = \sigma_u^2 \Omega_T(\rho)$, $\sigma_u = 1$, $\rho = 0.2$								
RMSE^{tmq}	0.811			1.002			1.040		
$rmse^{tmq}_{11,dt}$	6.624	57.506	0.97	8.448	47.668	0.00	18.129	64.743	0.21
$rmse^{tmq}_{12,dt}$	5.281	57.301	0.89	9.714	47.928	0.00	19.499	64.362	0.13
$rmse^{tmq}_{21,dt}$	6.977	55.923	0.98	8.036	46.331	0.00	18.064	64.639	0.25
$rmse^{tmq}_{22,dt}$	7.006	55.881	0.98	7.745	45.820	0.00	17.870	64.224	0.27
RMSE^{btmq}	0.638			0.871			0.933		
$rmse^{btmq}_{1,dt}$	4.119	54.558	0.70	18.203	44.658	0.00	35.200	68.035	0.10
$rmse^{btmq}_{2,dt}$	4.145	54.462	0.72	17.465	43.312	0.00	34.504	66.770	0.10
$rmse^{btmq}_{3,dt}$	6.467	59.178	0.14	15.335	61.613	0.00	22.095	91.702	0.10
Case 1.2	$\boldsymbol{u}_2 \backsim N_T(\mathbf{0}, \Sigma_u)$: $\Sigma_u = \sigma_u^2 \Omega_T(\rho)$, $\sigma_u = 1$, $\rho = 0.8$								
RMSE^{tmq}	0.800			1.017			1.056		
$rmse^{tmq}_{11,dt}$	6.007	57.181	0.97	10.660	46.763	0.00	19.548	63.270	0.12
$rmse^{tmq}_{12,dt}$	4.648	57.016	0.92	11.825	47.067	0.00	20.994	63.044	0.11
$rmse^{tmq}_{21,dt}$	6.406	55.626	0.98	10.254	45.521	0.00	19.383	63.263	0.14
$rmse^{tmq}_{22,dt}$	6.438	55.579	0.98	9.950	44.980	0.00	19.128	62.821	0.14
RMSE^{btmq}	0.620			0.880			0.940		
$rmse^{btmq}_{1,dt}$	3.862	54.707	0.72	19.912	44.300	0.00	36.336	67.695	0.10
$rmse^{btmq}_{2,dt}$	3.892	54.599	0.73	19.161	42.918	0.00	35.631	66.408	0.10
$rmse^{btmq}_{3,dt}$	5.620	68.705	0.17	16.238	60.570	0.00	24.483	91.384	0.10
Case 2	$u_{2,t} \backsim AR(3)$: $\phi_1 = 0.4$, $\phi_2 = 0.3$, $\phi_3 = 0.25$, $\sigma = 1$								
RMSE^{tmq}	0.802			1.012			1.049		
$rmse^{tmq}_{11,dt}$	7.335	58.195	0.97	9.572	46.851	0.00	19.109	64.197	0.18
$rmse^{tmq}_{12,dt}$	5.982	57.982	0.94	10.782	47.142	0.00	20.474	63.888	0.11
$rmse^{tmq}_{21,dt}$	7.696	56.635	0.97	9.170	45.454	0.00	18.993	63.946	0.21
$rmse^{tmq}_{22,dt}$	7.726	56.592	0.97	8.867	44.918	0.00	18.786	63.510	0.22
RMSE^{btmq}	0.630			0.878			0.938		
$rmse^{btmq}_{1,dt}$	4.668	55.316	0.80	18.947	44.205	0.00	35.512	67.396	0.10
$rmse^{btmq}_{2,dt}$	4.715	55.216	0.81	18.185	42.809	0.00	34.795	66.093	0.10
$rmse^{btmq}_{3,dt}$	6.213	69.447	0.15	15.769	58.991	0.00	22.402	72.775	0.10

Table 5.2: Performance evaluation of several methods of RMSE estimation for the TMQ and BTMQ predictors. Empirical RMSE and ARBIAS and RRMSE (in %).



(a) Scenario [0,0]. (b) Scenario [e,0]. (c) Scenario [e,u].

Figure 5.2: Boxplots of RBIAS and RRMSE in (%) for the several RMSE estimators for the TMQ and BTMQ predictors in Case 1.1.

5.6 Description of the 2013-2022 SLCS data

The proposed methodology is applied to assess changes in the average level of income in 23 provinces of Empty Spain (Pazos-Vidal, 2022; Pinilla and Saez, 2017), which refers to those provinces that have lost inhabitants between 1950 and 2019 and that also have a population density below the national average. The target population is made up of people with permanent residence in Spain and whose province, in the year of the interview, is classified as Empty Spain (see Table 5.3). In total there are $D = 23$ areas, covering 296,718 square kilometres –58% of the national territory– but they only represent the 17.2% of the Spanish population. From a socio-economic point of view, we have decided to focus on these areas because Spain has experienced a socio-economic revolution in recent decades, linked to large-scale migration movements from rural areas to large cities (Gobierno de España, 2019a,b). This leads to increasing differences between metropolis and urban and rural areas over time (Hepburn, 2016), so it could be interesting to estimate how the level of income in depopulated regions has been changing over the last few years.

Survey data are from the 2013-2022 Spanish Living Conditions Survey ($T = 10$ years) while the auxiliary variables come from the census files provided by the Spanish National Statistical Office (INE). The SLCS is designed to obtain reliable direct estimators in NUTS 2 regions, but sample sizes are quite small in NUTS 3 territories. Indeed, they range from 36 (Soria in 2014) to 1762 (Zaragoza in 2022), with a median value of 293. Unit-level data, measured

NUTS3 code	2	5	6	9	10	13	14	16
Province	Albacete	Ávila	Badajoz	Burgos	Cáceres	Ciudad Real	Córdoba	Cuenca
NUTS3 code	19	22	23	24	26	27	32	34
Province	Guadalajara	Huesca	Jaén	León	La Rioja	Lugo	Orense	Palencia
NUTS3 code	37	40	42	44	47	49	50	
Province	Salamanca	Segovia	Soria	Teruel	Valladolid	Zamora	Zaragoza	

Table 5.3: List and codification in the NUTS system of the provinces of Empty Spain.

in consecutive time periods, are hierarchically structured in provinces. Each individual (level 1: population level) is indexed according to their province (level 2: province level) and the year of the survey (time reference). The target population U is hierarchically structured in domains U_d, $d = 1, \ldots, D$, and subdomains or periods of time U_{dt}, $t = 1, \ldots, T$. The response variable is the equivalized disposable income, per person and unit of consumption, measured in thousands of euros. It is obtained by dividing the household's net income by the number of equivalent consumption units, according to the modified OECD scale (Hagenaars et al., 1994) to account for economies of scale in household consumption. For each individual $j \in U_{dt}$, the equivalized disposable income is denoted by y_{dtj}, $d = 1, \ldots, D,\ t = 1, \ldots, T$.

Auxiliary data is key to increase the effectiveness of the small area predictions. As we are dealing with unit-level data, only the following auxiliary variables are available: *sex* (*sex1*: men, *sex2*: women) and age group (*age1*: less than 25 years; *age2*: between 26 and 45 years; *age3*: between 46 and 64 years; *age4*: 65 years or older).

5.7 Application to the 2013-2022 SLCS data

5.7.1 Model fitting and validation

This section applies the developed methodology to the 2013-2022 SLCS data. We first fit the TWMQ linear models (5.14) to the target data, with sex and the four age groups (defined in Section 5.6) as auxiliary variables and *sex1*:*age4-1* as the reference category. The projective influence function ψ is the Huber function with tuning constant $c_\psi = 1.345$. The model parameters vary over time and province, which amounts to a total of $5 \cdot 23 \cdot 10 = 1150$ model parameters.

In terms of model specification, Figure 5.3 shows that the assumption of identical regression coefficients over time is not reasonable. The Student's t-test confirms that the mean of $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_D\}$ is different from 0.5 at 5% (p-value 0.034), where the normality hypothesis is verified according to the Shapiro-Wilk test (p-value 0.212). It is therefore essential to model the provincial heterogeneity through the use of $\widehat{\theta}_d$, $d = 1, \ldots, D$. To have more confidence about the fitted model as a true generating model, its validation is addressed below.

As the reader may be aware, residual analysis is widely used to assess the adequacy of a model by examining the differences between observed and predicted values. Let $d = 1 \ldots, D$,

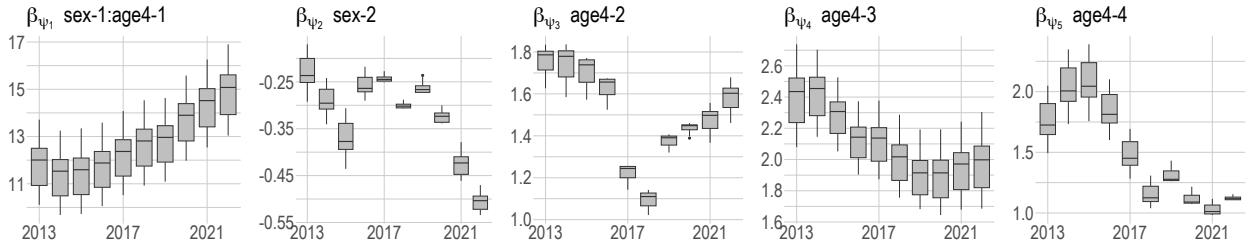


Figure 5.3: Boxplot of the model parameters of the TWMQ linear models by year for the 2013-2022 SLCS data.

$t = 1, \ldots, T$. For $j = 1, \ldots, n_{dt}$ and $q = \widehat{\theta}_d$, the model residuals are

$$\widehat{e}_{\psi,dtj} \stackrel{\Delta}{=} y_{dtj} - \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}(\widehat{\theta}_d, w_t).$$

We define the subdomain sample means of model residuals as $\overline{\widehat{e}}_{\psi,dt.} = \frac{1}{n_{dt}} \sum_{j=1}^{n_{dt}} \widehat{e}_{\psi,dtj}$, the aggregated raw residuals (ARR) as $\overline{\widehat{e}}_{\psi,dt.} - \overline{\widehat{e}}_{\psi,..}$, where $\overline{\widehat{e}}_{\psi,..} = \frac{1}{DT} \sum_{d=1}^{D} \sum_{t=1}^{T} \overline{\widehat{e}}_{\psi,dt.}$, and the aggregated standardized residuals (ASR) by dividing by the standard deviation, i.e.

$$(\overline{\widehat{e}}_{\psi,dt.} - \overline{\widehat{e}}_{\psi,..})\nu^{-1}, \quad \text{where} \quad \nu = \Big(\frac{1}{DT} \sum_{d=1}^{D} \sum_{t=1}^{T} (\overline{\widehat{e}}_{\psi,dt.} - \overline{\widehat{e}}_{\psi,..})^2\Big)^{1/2}.$$

Figure 5.4 includes boxplots of the ASRs by province (left) and year (right). As a result, most of them oscillate around $y = 0$ and lie in the interval $(-3, 3)$. Not surprisingly, the provincial variability is greater than the annual one, but neither province seems to be particularly poorly modelled. Outlier detection, based on the selection of area-time specific robustness parameters, is presented in Section 5.7.3.

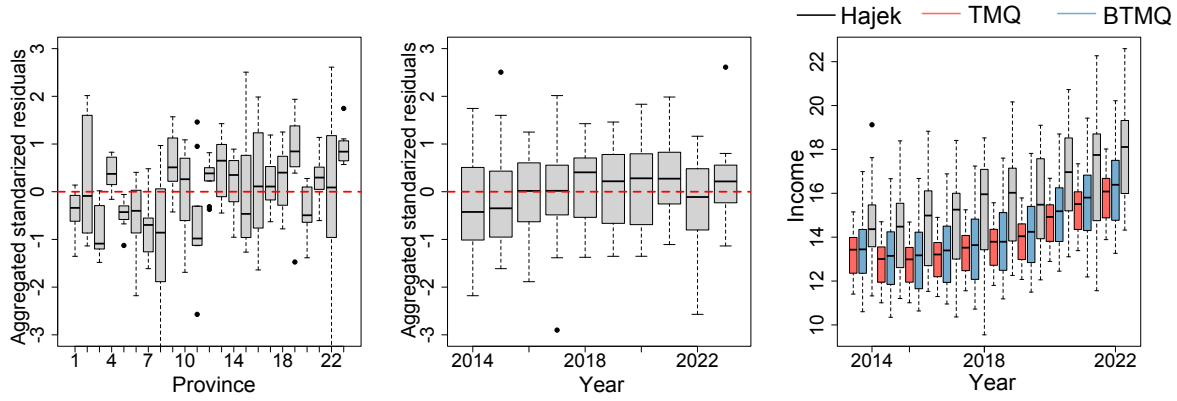


Figure 5.4: Boxplot of the ASRs of the TWMQ linear models by province (left) and year (center); and boxplots of model-based predictions and direct estimates by year (right). Data from the 2013-2022 SLCS.

5.7.2 Prediction, error measures and maps

In the following, the prediction is performed and the error measurements are calculated. As pointed out in Section 5.4.1, the TMQ predictor may introduce nonnegligible prediction biases, but the BTMQ predictor can unbalance the bias-variance trade-off of the MSE. To set the value of the robustness parameter c_ϕ, we use the selection criterion proposed in Section 5.4.4. To provide more information about the robustness parameters $\{\widehat{c}_{\phi,dt} : d = 1, \ldots, D, t = 1, \ldots, T\}$, some relevant quantiles are calculated: $\widehat{c}_{\phi,0} = \widehat{c}_{\phi,0.01} = \widehat{c}_{\phi,0.05} = 0$, $\widehat{c}_{\phi,0.25} = 0.290$, $\widehat{c}_{\phi,0.5} = 0.644$, $\widehat{c}_{\phi,0.75} = 1.016$, $\widehat{c}_{\phi,0.95} = 1.772$, $\widehat{c}_{\phi,0.99} = 2.853$ and $\widehat{c}_{\phi,1} = 3.345$. In addition, in 33 subdomains (14%) $\widehat{c}_{\phi,dt} = 0$, i.e. no bias correction is needed. An important spin-off is that the variability of these subdomains is not unnecessarily increased because of an improper bias correction. Figure 5.4 (right) plots Hájek estimates and model-based predictions for the TMQ and BTMQ predictors. The BTMQ estimator seems to smoothen the estimates better, as expected, employing a bias correction term.

Regarding the variability of the estimates, we focus on the MSE of the BTMQ predictor. Table 5.4 contains the deciles of the sample sizes n_{dt}, of the standard deviations of the Hájek estimator (see Morales et al. (2021), Chap. 3) and of the BTMQ predictor. Looking at the results in Table 5.4, the reduction in variability is evident, especially when the sample sizes are small, which supports the benefits of our proposal.

	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
n_{dt}	36	118	146	182	239	294	346	450	568	900	1762
Hájek	0.245	0.361	0.407	0.455	0.511	0.561	0.648	0.727	0.859	1.118	2.393
BTMQ	0.032	0.040	0.088	0.126	0.160	0.184	0.213	0.251	0.307	0.397	1.023

Table 5.4: Sample sizes and standard deviations of Hájek estimator and BTMQ predictor.

Lastly, and quite incidentally, the proposed estimation procedure offers the opportunity to analytically read the evolution and differences between the provinces of Empty Spain over time. Consequently, it provides valuable information for decision-making, the study of socio-economic trends and the implementation of measures related to the equitable and fair distribution of wealth. Figure 5.5 maps the equivalized disposable income for 2013 (left), 2018 (centre) and 2022 (right) obtained with the BTMQ predictor. We report these maps for the BTMQ predictor because its preference over the TMQ predictor is justified both in simulation studies and in the application to real data.

Figure 5.5 points out that there are clear differences between the northern provinces, historically richer and more developed, and those in the centre-south, where agriculture and construction predominate and the industrial sector is less promoted. The richest provinces in north-central Spain correspond to four of the atypical areas mentioned in Section 5.7.3. Finally, it is worth noting the increasing, or at least non-decreasing, trend during the study period. Figure 5.6 shows maps of the RRMSE estimates using the $rmse^{btmq}_{2,dt}$ estimator proposed in Section 5.4.3. It follows that the relative margins of error are accurate enough for a SAE problem, with RRMSE estimates lower than 9% in most domains, only exceeding it in 2 of the 23 provinces for the 3 years mapped.

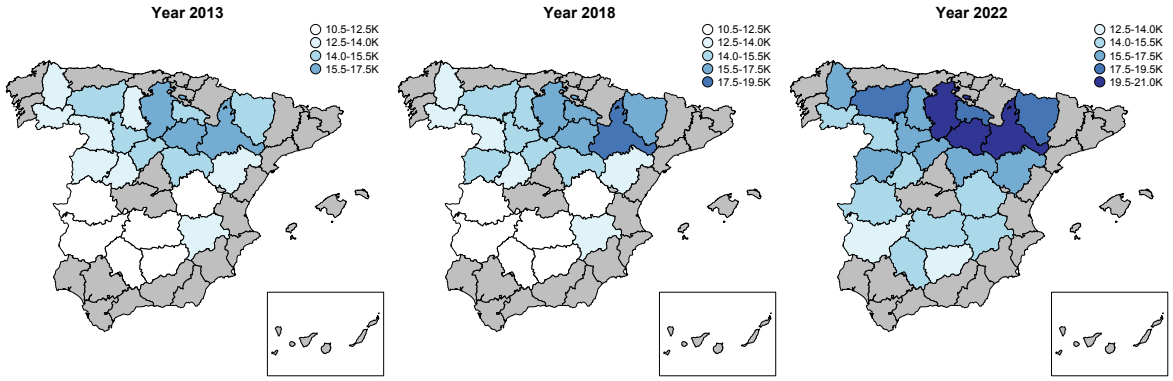


Figure 5.5: Equivalized disposable income for Empty Spain in 2013 (left), 2018 (center) and 2022 (right). The provinces in grey are those that do not belong to Empty Spain. Results for the BTMQ predictor and the 2013-2022 SLCS data.

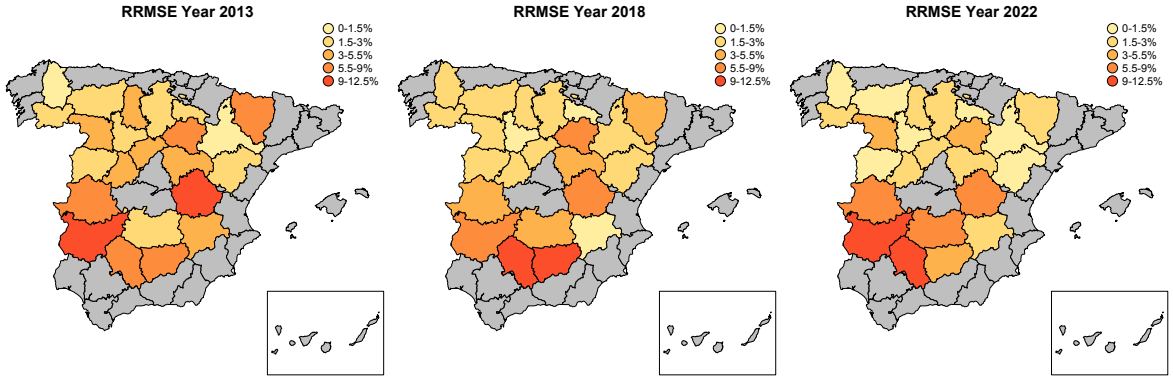


Figure 5.6: RRMSE of the equivalized disposable income for Empty Spain in 2013 (left), 2018 (center) and 2022 (right). The provinces in grey are those that do not belong to Empty Spain. Results for the BTMQ predictor and the 2013-2022 SLCS data.

5.7.3 Detection of outliers

Last but not least, in this section we use the robustness parameters to detect the outlying subdomains (cf. Section 5.5). As a starting point, outlier detection methods based on LMMs, such as that of Zewotir and Galpin (2007), cannot be used in the application to real data because they involve inverting matrices of order $n \times n$, where $n = 89971$ for the provinces of Empty Spain in the SLCS2013-2022. Moreover, if we fit a LMM with random effects in provinces and years, Cook's distances do not detect deviations. This is probably due to the strong time dependencies –and also the presence of unit-level outliers– which cloud the provincial variability of the random effects. Against this background, we propose using the set of values $\{\widehat{c}_{\phi,dt} : \ d = 1, \ldots, D, \ t = 1, \ldots, T\}$.

For a preliminary idea, Table 5.5 shows the average value of $\widehat{c}_{\phi,dt}$ by province, $\overline{\widehat{c}}_{\phi,d.} =$

$\frac{1}{T}\sum_{t=1}^{T}\widehat{c}_{\phi,dt}$, and year, $\overline{\widehat{c}}_{\phi,.t}=\frac{1}{D}\sum_{d=1}^{D}\widehat{c}_{\phi,dt}$, and the global mean, $\overline{\widehat{c}}_{\phi,..}=\frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\widehat{c}_{\phi,dt}$. It could be noted that the provincial distribution is highly variable and the annual one more uniform, which suggests potential differences between provinces.

d	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{c}_{\phi,d.}$	0.195	0.132	0.894	1.751	0.765	0.707	0.810	0.413	0.994	1.182	0.731	1.069
d	13	14	15	16	17	18	19	20	21	22	23	$\overline{\widehat{c}}_{\phi,..}$
$\widehat{c}_{\phi,d.}$	1.340	0.207	0.123	0.446	0.674	0.475	1.077	0.171	1.077	0.207	1.408	0.732

t	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	$\overline{\widehat{c}}_{\phi,..}$
$\widehat{c}_{\phi,.t}$	0.727	0.816	0.637	0.746	0.637	0.803	0.782	0.593	0.892	0.693	0.732

Table 5.5: Average value of the robustness parameters $\widehat{c}_{\phi,dt}$ by province (top) and year (bottom), compared to $\overline{\widehat{c}}_{\phi,..}$. Results for the 2013-2022 SLCS data.

To test whether there are domains that show persistent atypical behaviour over time we employ the Friedman's test (Friedman, 1937) (as in Section 5.5). Significant evidence is found for the selection of $\widehat{c}_{\phi,dt}$ between provinces (p-value $\simeq 0$), but not between years (p-value 0.358). To detect which groups of provinces shift in scale, we apply a Tukey (1949) multiple range test. It is found that Burgos ($d=4$), Huesca ($d=10$), La Rioja ($d=13$), Guadalajara ($d=19$) and Zaragoza ($d=23$) are atypicals over time, with higher values than the average. For the rest of Empty Spain's provinces, no significant differences are detected. In socio-economic terms, our findings are reasonable. Burgos, Huesca, La Rioja and Zaragoza –all four in the centre-north of Spain– are traditionally prosperous provinces and Guadalajara is very close to Madrid, serving as a "commuter province" for many workers from the capital. In short, all these regions deviate from the patterns that characterise Empty Spain.

Chapter 6

Conclusions

The aim of this chapter is to outline and assess the achievements and improvements made by the contributions described in this thesis, as well as to present the lines of future research.

6.1 Summary and discussion

It is widely acknowledged that conquering poverty and reducing social inequalities are challenges for the near future. We need to think globally and act locally. Estimates of finite population parameters for subgroups, such as geographic areas or socio-economic groups, are increasingly required for better planning and evaluation of government programs. In addition, global warming and land use are changing fire dynamics worldwide, increasing fire activity and its impact on ecosystems, livelihoods and urban settlements. The first step is to have accurate information on which to act, and thus to promote statistical research. Indeed, the lack of attention to sample sizes, and often their smallness, could lead to inaccurate results. SAE techniques help to meet the growing demand for reliable disaggregated statistics by fitting statistical models to unit-level or area-level data. In this respect, their increasing importance cannot be denied. Until recently, the contributions have been many, but the ongoing globalisation and climate change pose many challenges in this field as well. To leave no one behind, accurate information must be available to identify small communities in need and to address their problems accurately and effectively.

In the context of SAE, the contributions proposed in this thesis include the study of area-level zero-inflated mixed models in Chapter 2, the derivation of plug-in and EBP-type predictors and the estimation of MSEs. To start with, the applicability of the area-level zero-inflated mixed models has been demonstrated both for the estimation of socio-demographic indicators, such as the proportion of single-person households in small areas, and for tackling the problem of predicting the number and size of forest fires in Spain. It stands to reason that there is a need for research into tools to raise awareness of extreme fire events. In this sense, the use of zero-inflated structures has been found to be successful in modelling these data, due to the natural cause of the excess of zeros that can occur in forest fire research.

Still in the area-level modelling research, and knowing that there are no published studies

dealing with the estimation of segregation indexes in small areas, Chapter 3 contributes to this field. In particular, it fits an three-fold Fay-Herriot (FH3) model, then derives model-based predictors of Duncan Segregation Indexes (DSI) and estimates MSEs, and finally analyses the poor performance of direct estimators. Moreover, our simulation studies have shown that it is not necessary to fit measurement error models when the explanatory variables are estimated using considerably more data than those used to calculate direct estimates of the target variables. It is our belief (and hope) that the latter is a valuable starting point for the promotion of SAE in sociological studies of current interest.

Under a unit-level model-based approach, Chapter 4 studies the small area prediction of the proportions of employed, unemployed and inactive people, and of unemployment rates. In the population units (people residing in Spain aged 16 and over), the target variables are dichotomous and indicate whether they belong to the three employment status categories or not. Since these dummy variables sum up to one, they are modelled vectorially using a unit-level multinomial logit mixed model. Thus, it is assumed that the above model generates the values of the vector of target variables in all units of the population. That is, we have accepted the paradigm that the only source of randomness comes from the superpopulation model and the sample is a fixed subset. The extension of the theory to a more general probabilistic setting, where the model and sampling distributions are considered together, has not been addressed. This problem is complex and deserves further specific research.

In a general view, SAE unit-level models are powerful tools for describing target variables if the model fits the data properly. When a supporting census file is available, model-based plug-in and EBP-type predictors are expected to have a high predictive capability. Unfortunately, this is not usually the case, and the methodology loses strength when it is restricted to ANOVA-type models. The latter weakness can be partially addressed by using contextual models, i.e. fitting unit-level models with area-level auxiliary variables. Having said that, it is appealing to have statistical methods that do not have to rely on less informative aggregated data. In this sense, it is not intended to replace procedures based on area-level models, to which great contributions have been made. It is therefore up to the statistical teams to choose the methodology to be used in each case, as there is no universally better one.

The suitability of unit-level models, but also the need to avoid some strong distributional assumptions that the use of mixed models entails, brings the discussion to the next topic: the contributions to MQ regression made in Chapter 5. First and foremost, the effective use of past information and the modeling of temporal dependencies is an appealing method for borrowing strength in SAE. At this regard, MQ models that capture time-dependent relationships are proposed to avoid the strong distributional assumptions of unit-level independence and the formal specification of the hierarchical structure of the random effects. To achieved this, the MQ models have been adapted to temporal data by including weights in the fitting process and defining semiparametric temporal distance criteria. As an inherent property of MQ models, our approach avoids distributional assumptions and allows for characterizing differences between areas, as well as time dependencies, through data-driven estimation of the regression coefficients. Consequently, the new models feature time-varying parameters and remain distribution-free for both areas and time.

As final remarks, the model-based simulations illustrate the adequacy and, where applicable, the superiority of the new techniques. In addition, indirect estimators are often biased,

and even more so when derived from robust models, but their variance is lower than that of direct (design-based unbiased) estimators. As for the latter, and according to our simulation results, we have been able to correct for bias without increasing the variability in plug-in type predictors derived from general MQ models. Comparing the Time-Weighted M-quantile (TWMQ) models with area-time linear mixed models, the latter are successful as long as their hypotheses are correct. Namely, the strong distributional assumptions imposed on the temporal structure and the absence of outliers. Our research shows that, in a more general setup, no other predictor improve the performance measures achieved with the robust bias-corrected temporal MQ (BTMQ) predictor, with area-time-specific robustness parameters.

6.2 Further research

With all that has been investigated so far, there are many opportunities for future research. First and foremost, the inclusion of new auxiliary variables to improve the predictive performance of our models would be beneficial for future investigations of forest fire data. This could include the availability of information on the arrival of fire-fighting resources at forest fires, land use and socio-economic variables describing the demographic trends of the inhabitants of each region. Based on the statistical results presented here, the aim could be to refine the prediction of extreme, highly damaging and dangerous events. We expect to present results for the region of Galicia (northwest Spain) in the near future with improved models using this information.

Regarding the prediction of non-linear indicators that has been done in this thesis, it remains to be investigated the estimation of Duncan Segregation Indexes (DSI) derived from bivariate FH models. In that case, one component would represent the men group and the other component, the women group. As far as the multinomial logit mixed model is concerned, we are working on the inclusion of correlated random effects. The computational cost of this more complex model needs to be assessed in order to decide whether it is worthwhile to account for such correlations in academia or in the production of public statistics. Again, the methodology would be useful for estimating labour indicators, such as unemployment rates. The above two investigations will allow us to compare the methods used and, if appropriate, to improve the results reported in this thesis.

In addition, a typical modelling approach for target variables with zero inflation is based on parametric models, mixing degenerate distributions at zero and appropriate parametric distributions to model the remaining part of the distribution of the target variable (see Chapter 2 for further details). However, although classical quantile regression has recently been adapted to zero-inflated data (Ling et al., 2022), not only has it never been applied to SAE, but also zero-inflated MQ models have never been investigated. As of today, we are working on the extension of the MQ regression, which has numerous advantages over quantile regression, to zero-inflated data and the modelling of hierarchical dependency structures. The future contribution will include the proposal of zero-inflated MQs and MQ models, the study of asymptotic properties, the derivation of robust predictors, their optimal bias correction and the analytical calculation of MSEs. The methodology will be evaluated by means of model-based simulations, investigating the potential gain that the new proposal could bring

in the presence of only a few atypical data.

Another alternative to enrich the literature could be to extend the TWMQ statistical methodology to spatio-temporal data. Indeed, the derivation of small area estimators that account for both temporal and spatial correlations may yield better results. This will be treated elsewhere. We also believe that there is room for improvement in the selection of robustness parameters for bias correction and their implications for the calculation of robust predictors. Added to this, the reader should be aware that the prediction of non-linear quantities, and even more so the estimation of the MSE, requires further research and could be investigated elsewhere. Finally, the current approach is valid only for continuous outcome variables. Future work will extend the generalized MQ regression models to time-dependent data to derive small area predictors for discrete response variables.

To conclude, it would be interesting to work on open source development in the near future. As the reader may be aware, the focus here was not on the code, although the later would be quite useful for the success of this work.

6.3 Conclusions in Spanish

Las contribuciones de esta tesis incluyen el estudio de modelos mixtos inflados en el cero a nivel de área en el Capítulo 2, el cálculo de predictores *plug-in* y EBP y la estimación de errores cuadráticos medios. En la práctica su aplicabilidad ha sido demostrada tanto para la estimación de indicadores sociodemográficos, como la proporción de hogares unipersonales en áreas pequeñas, como para abordar el problema de la predicción de incendios forestales en España. Continuando con el enfoque de área, el Capítulo 3 consiste en un estudio pionero sobre la estimación de índices de segregación en áreas pequeñas. En particular, se ajusta un modelo Fay-Herriot de tres niveles (FH3) para predecir Índices de Segregación de Duncan (DSI) y se llevan a cabo estudios de simulación. Según nuestros resultados, no es necesario ajustar modelos de error de medida cuando las variables explicativas se estiman empleando muchos más datos que los utilizados para calcular las estimaciones directas de las variables objetivo. Creemos (y esperamos) que esto último promueva la estimación en áreas pequeñas (SAE) en estudios sociológicos de interés actual.

Bajo un enfoque basado en modelos a nivel de unidad, el Capítulo 4 estudia la predicción en áreas pequeñas de las proporciones de ocupados, parados e inactivos, y de las tasas de paro. En las unidades de la población, las variables objetivo son dicotómicas e indican si pertenecen o no a las tres categorías de situación laboral. Dado que estas variables ficticias suman uno, se modelizan vectorialmente mediante un modelo mixto logit multinomial. Así, se supone que el modelo anterior genera los valores del vector de variables objetivo en todas las unidades de la población. Es decir, hemos aceptado el paradigma de que la única fuente de aleatoriedad procede del modelo de superpoblación y la muestra es un subconjunto fijo. No se ha abordado la extensión de la teoría a un entorno probabilístico más general, en el que el modelo y las distribuciones muestrales se consideran conjuntamente. Este problema es complejo y merece una investigación más específica.

En general, los modelos a nivel de unidad son herramientas potentes para describir vari-

ables objetivo si se ajustan bien a los datos. Cuando se dispone de un archivo censal de apoyo, se espera que los predictores *plug-in* y EBP tengan una gran capacidad predictiva. Por desgracia, esto último no es habitual y la metodología pierde fuerza al limitarse a modelos ANOVA. Dicho esto, resulta atractivo disponer de métodos estadísticos que no tengan que basarse en datos agregados menos informativos. En este sentido, no se pretende sustituir a los procedimientos basados en modelos a nivel de área, sobre los que se han realizado grandes aportaciones. Por lo tanto, corresponde a los equipos estadísticos elegir la metodología que se utilizará en cada caso, ya que no existe una que sea universalmente mejor.

La idoneidad de los modelos a nivel de unidad, pero también la necesidad de evitar ciertas restricciones paramétricas vinculadas al uso de modelos mixtos, lleva la discusión a la última aportación: las contribuciones a la regresión M-cuantil (MQ) del Capítulo 5. En primer lugar, se proponen modelos MQ que capturan dependencias temporales para relajar la hipótesis de independencia de los errores y la especificación formal de los efectos aleatorios. Para lograrlo, los modelos MQ se han adaptado a datos temporales mediante la inclusión de ponderaciones en el proceso de ajuste y la definición de criterios semiparamétricos de distancia temporal. Nuestras simulaciones ilustran la idoneidad y, en su caso, la superioridad de las nuevas técnicas. Por otra parte, hemos conseguido corregir el sesgo sin aumentar la variabilidad de los predictores *plug-in* derivados de modelos MQ. Nuestra investigación muestra que, en escenarios generales, ningún otro predictor puede mejorar las medidas de rendimiento conseguidas con el predictor robusto MQ temporal con corrección de sesgo (BTMQ), con parámetros de robustez específicos del área y del tiempo.

Appendix A

The maximum likelihood Laplace algorithm for fitting the zero-inflated GLMMs in Chapter 2

This appendix describes the Laplace algorithm for the model log-likelihood. For more details on the ML-Laplace algorithm, see e.g., Demidenko (2013) and Kristensen et al. (2016). In this thesis, the ML-Laplace algorithm is used to calculate ML estimators of the model parameters and modal predictors of the random effects of the area-level zero-inflated mixed models[1] detailed in Chapter 2. Note that in all cases the proposed methodology is based on a mixture model with a BE distribution and a PO, NB or GA distribution, as appropiate. Although for a mixture-type model it seems to be more natural to apply an expectation-maximization algorithm, it is not recommended in our research. The reason for this is that an expectation-maximization algorithm does not provide modal predictions of the random effects (Wu, 1983), which are key to calculate plug-in type predictors of area-level small area quantities. The function `glmmTMB` of the `R` (R Development Core Team, 2024) package `glmmTMB` (Brooks et al., 2017) implements the ML-Laplace algorithm.

Let us start with the Laplace approximation of a multiple integral of a general function $\exp(h(\boldsymbol{x}))$, where $h : \mathbb{R}^m \mapsto \mathbb{R}$ is a twice continuously differentiable function with a global maximum at the column vector $\boldsymbol{x}_0$. This is to say, let us assume that $\dot{h}(\boldsymbol{x}_0) = \left.\frac{dh}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} = \mathbf{0}$ and $\ddot{h}(\boldsymbol{x}_0) = \left.\frac{d^2h}{d\boldsymbol{x}^2}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is negative definite. A Taylor series expansion of $h(\boldsymbol{x})$ around $\boldsymbol{x}_0$ yields to

$$\begin{aligned} h(\boldsymbol{x}) &= h(\boldsymbol{x}_0) + \dot{h}'(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\ddot{h}(x_0)(\boldsymbol{x}-\boldsymbol{x}_0) + o\big(\|\boldsymbol{x}-\boldsymbol{x}_0\|^2\big) \\ &\approx h(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\ddot{h}(x_0)(\boldsymbol{x}-\boldsymbol{x}_0). \end{aligned}$$

[1] The ML-Laplace algorithm has also been described in Chapter 4 to maximize the log-likelihood of a unit-level multinomial logit mixed model. Nevertheless, this is a self-contained chapter and the log-likelihood of a multinomial logit mixed model differs greatly from that of the zero-inflated models discussed here.

Using this expansion, the multivariate Laplace approximation of the integral of $\exp\left(h(\boldsymbol{x})\right)$ is

$$\begin{aligned}\int_{\mathbb{R}^m} e^{h(\boldsymbol{x})}\, d\boldsymbol{x} &\approx \int_{\mathbb{R}^m} e^{h(\boldsymbol{x}_0)} \exp\Big\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\big(-\ddot{h}(\boldsymbol{x}_0)\big)(\boldsymbol{x}-\boldsymbol{x}_0)\Big\}\, d\boldsymbol{x} \\ &= (2\pi)^{m/2}\big|-\ddot{h}(\boldsymbol{x}_0)\big|^{-1/2} e^{h(\boldsymbol{x}_0)},\end{aligned}$$

where we use that the integral of the multivariate normal p.d.f. $f(\boldsymbol{x})$ is one.

Below are the explicit expressions of the likelihood and log-likelihood functions of the zero-inflated mixed models proposed in this thesis: the area-level zero-inflated PO (aZIP13), the area-level zero-inflated NB (aZINB11) and the area-level zero-inflated GA (aZIG22) mixed models in Chapter 2. An analytical approximation of the corresponding log-likelihood functions is also included for its use in future steps.

1. The likelihood of the aZIP13 mixed model is

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{K(1+IJ)}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \int_{\mathbb{R}^{K(1+IJ)}} \exp\big\{h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})\big\}\, d\boldsymbol{u}, \tag{A.1}$$

where

$$h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \log P(y_{ijk}|\boldsymbol{u}_{ijk};\boldsymbol{\theta}) - \frac{K(1+IJ)}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\Big(u_{1,k}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J} u_{2,ijk}^2\Big).$$

The log-likelihood of the aZIP13 mixed model is approximated by

$$\log P(\boldsymbol{y};\boldsymbol{\theta},) \approx IJ\log 2\pi + h(\boldsymbol{u}^\circ) - \frac{1}{2}\log|-\ddot{h}(\boldsymbol{u}^\circ)| \triangleq g(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}^\circ).$$

2. The likelihood of the aZINB11 mixed model is

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{4JK}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \int_{\mathbb{R}^{4JK}} \exp\big\{h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})\big\}\, d\boldsymbol{u}, \tag{A.2}$$

where

$$h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \log P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) - \frac{4JK}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}(u_{1,j}^2 + u_{1,k}^2 + u_{2,j}^2 + u_{2,k}^2).$$

The log-likelihood of the aZINB11 mixed model is approximated by

$$\log P(\boldsymbol{y};\boldsymbol{\theta},) \approx 2JK\log 2\pi + h(\boldsymbol{u}^\circ) - \frac{1}{2}\log|-\ddot{h}(\boldsymbol{u}^\circ)| \triangleq g(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}^\circ).$$

3. The likelihood of the aZIG22 mixed model is

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{2JK}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \int_{\mathbb{R}^{2JK}} \exp\big\{h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})\big\}\, d\boldsymbol{u}, \tag{A.3}$$

where

$$h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \log P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) - \frac{2JK}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}(u_{1,jk}^2 + u_{2,jk}^2).$$

The log-likelihood of the aZIG22 mixed model is approximated by

$$\log P(\boldsymbol{y};\boldsymbol{\theta},) \approx 2JK\log 2\pi + h(\boldsymbol{u}^\circ) - \frac{1}{2}\log|-\ddot{h}(\boldsymbol{u}^\circ)| \triangleq g(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}^\circ).$$

To apply the Laplace algorithm to the integrals in (A.1), (A.2) or (A.3), we have to maximize $h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})$ in $\boldsymbol{u}$, given $\boldsymbol{y}$ and $\boldsymbol{\theta}$. For simplicity, we write $h(\boldsymbol{u})$. We can carry out the maximization by applying a `R` function of optimization. Alternatively, we implement a Newton-Raphson algorithm after calculating the first and second partial derivatives of h with respect to the vector of random effects $\boldsymbol{u}$, given $\boldsymbol{y}$ and $\boldsymbol{\theta}$. Let $\dot{h}$ and $\ddot{h}$ denote the vector and the matrix of first and second order partial derivatives of $h(\boldsymbol{u})$ with respect to $\boldsymbol{u}$, respectively.

The Newton-Raphson updating equation is

$$\boldsymbol{u}^{(r+1)} = \boldsymbol{u}^{(r)} - \ddot{h}^{-1}(\boldsymbol{u}^{(r)})\,\dot{h}(\boldsymbol{u}^{(r)}). \tag{A.4}$$

Let us denote by $\boldsymbol{u}^{\circ}$ the argument of maxima of the function $h(\boldsymbol{u})$. It holds $\dot{\boldsymbol{h}}(\boldsymbol{u}^{\circ}) = \boldsymbol{0}$ and the matrix $\ddot{h}(\boldsymbol{u}^{\circ})$ is negative definite. The following step is to maximize $g(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}^{\circ})$ in $\boldsymbol{\theta} \in \Theta$. For simplicity, we write $g(\boldsymbol{\theta})$. We can carry out the maximization by applying a `R` function of optimization. Alternatively, a successful option is to apply again a Newton-Raphson algorithm after calculating the first and second partial derivatives of g with respect to the components of $\boldsymbol{\theta}$, given $\boldsymbol{y}$ and $\boldsymbol{u}^{\circ}$. Let us define the size of the parameter space as $M = \dim(\Theta) = q_1 + q_2 + 2$. Let $\dot{g}$ and $\ddot{g}$ denote the $M \times 1$ vector and the $M \times M$ matrix of first and second order partial derivatives of $g(\boldsymbol{\theta})$, respectively.

The Newton-Raphson updating equation is

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \ddot{g}^{-1}(\boldsymbol{\theta}^{(r)})\,\dot{g}(\boldsymbol{\theta}^{(r)}). \tag{A.5}$$

The final ML-Laplace algorithm, used for both estimating $\boldsymbol{\theta}$ and predicting $\boldsymbol{u}$, combines the two Newton-Raphson algorithms and is summarised by the following steps:

1. Set the initial values $r = 0$, $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $\varepsilon_3 > 0$, $\varepsilon_4 > 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)} = \boldsymbol{\theta}^{(0)} + \boldsymbol{1}$, $\boldsymbol{u}^{(0)} = \boldsymbol{0}$, $\boldsymbol{u}^{(-1)} = \boldsymbol{1}$, where $\boldsymbol{0}$ and $\boldsymbol{1}$ are column vectors of zeros and ones, respectively.
2. Until $\|\boldsymbol{\theta}^{(r)} - \boldsymbol{\theta}^{(r-1)}\|_2 < \varepsilon_1$, $\|\boldsymbol{u}^{(r)} - \boldsymbol{u}^{(r-1)}\|_2 < \varepsilon_2$, do
   (a) Apply the Newton-Raphson updating equation (A.4) with seeds $\boldsymbol{u}^{(r)}$, convergence tolerance ε_3 and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r)}$ fixed. Output: $\boldsymbol{u}^{(r+1)}$.
   (b) Apply the Newton-Raphson updating equation (A.5) with seeds $\boldsymbol{\theta}^{(r)}$, convergence tolerance ε_4 and $\boldsymbol{u} = \boldsymbol{u}^{(r+1)}$ fixed. Output: $\boldsymbol{\theta}^{(r+1)}$.
   (c) $r \leftarrow r + 1$.
3. Output: $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(r)}$ and $\widehat{\boldsymbol{u}} = \boldsymbol{u}^{(r)}$.

As output from the ML-Laplace algorithm, and apart from the ML estimators of the model parameters, we obtain modal predictors, $\widehat{\boldsymbol{u}}$, of random effects and the maximized marginal log-likelihood. Given that ML estimators are consistent and asymptotically normal when the number of domains tends to infinity (see e.g. Section 3.7.2 in Jiang (2007)), the algorithm can also be used to approximate the asymptotic covariance matrix (inverse of the Fisher's information matrix) which allows the calculation of Wald statistics to test hypotheses about the model parameters. In practice, we use the sign-shifted Hessian matrix (second derivatives of the log-likelihood function changed in sign) as an approximation of the Fisher's information

matrix. In other words, the asymptotic variance matrix of $\widehat{\boldsymbol{\theta}}$, $\boldsymbol{Q}(\boldsymbol{\theta})$, is approximated by $\boldsymbol{Q}(\boldsymbol{\theta}) \approx -\ddot{g}^{-1}(\widehat{\boldsymbol{\theta}})$. Further, the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ is $N_M(\boldsymbol{\theta}, \boldsymbol{Q}(\boldsymbol{\theta}))$. Therefore, an asymptotic CI at the level $1-\alpha$ for a component θ_ℓ of $\boldsymbol{\theta}$ is

$$\widehat{\theta}_\ell \pm z_{1-\alpha/2}\, q_{\ell\ell}^{1/2},\ \ell = 1, \ldots, M,$$

where $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^\kappa$, $\boldsymbol{Q}(\boldsymbol{\theta}^\kappa) = (q_{ab})_{a,b=1,\ldots,M}$, κ is the last iteration of the ML-Laplace algorithm and z_α is the α-quantile of the $N(0,1)$ distribution. For each $\beta_{a\ell}$, $a = 1, 2$, $\ell = 1, \ldots, q_a$, we calculate asymptotic p-values to test significance. If $\widehat{\beta}_{1\ell} = \beta_0$, the p-value to test $H_0 : \beta_{1\ell} = 0$ vs $H_1 : \beta_{1\ell} \neq 0$ is

$$p\text{-value} = 2P_{H_0}(\widehat{\beta}_{1\ell} > |\beta_0|) = 2P(N(0,1) > |\beta_0|/\sqrt{q_{\ell\ell}}), \quad \ell = 1, \ldots, q_1.$$

To test $H_0 : \beta_{2\ell} = 0$ vs $H_1 : \beta_{2\ell} \neq 0$, we use $q_{q_1+\ell\, q_1+\ell}$ instead of $q_{\ell\ell}$.
The Akaike Information Criterion, commonly used to compare nested models according to the size M and its goodness-of-fit, is calculated as

$$2M - 2g(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}, \widehat{\boldsymbol{u}})$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{u}}$ are taken from the output of the ML-Laplace algorithm.

Appendix B

K-means algorithm

This appendix describes the K-means algorithm (Hartigan and Wong, 1979), the clustering method used in Section 2.4. First, a brief explanation of clustering methods is provided. Pattern Recognition deals with the construction of mechanisms capable of extracting relevant information and key patterns from sample observations. That is, the identification of regularities in the data, in order to impose a set of identity (classification, clustering, association, etc.) or dependence (regression) relationships. Cluster analysis, or simply clustering, is the task of grouping a set of observations in such a way that observations in the same group (cluster) are more similar (in some sense) to each other than to those in other groups. The aim of these techniques is to form groups in order to detect patterns or structures within the population. Clustering itself is not a specific algorithm, but the general problem to be solved.

The K-means algorithm finds $k \in \mathbb{Z}^{\geqslant 1}$ clusters (fixed value), around a given set of centres $\{\boldsymbol{m}_1^{(1)}, ..., \boldsymbol{m}_k^{(1)}\}$ which define the initial clusters $S_1^{(1)}, ..., S_k^{(1)}$, by iterating the following steps:

1. Assign each observation $\boldsymbol{\mathcal{X}}_p$ to a single cluster, being the one with the closest mean:

$$S_\ell^{(r)} = \left\{ \boldsymbol{\mathcal{X}}_p \ : \| \boldsymbol{\mathcal{X}}_p - \boldsymbol{m}_\ell^{(r)} \|_2^2 \leqslant \| \boldsymbol{\mathcal{X}}_p - \boldsymbol{m}_h^{(r)} \|_2^2, \ h = 1, ..., k \right\}, \quad \ell = 1, ..., k.$$

   It is required that $\boldsymbol{\mathcal{X}}_p$ is assigned to exactly one cluster $S_\ell^{(r)}$, although it could be in two or more.

2. For each cluster, calculate the means that will be used as centres of the new clusters:

$$\boldsymbol{m}_\ell^{(r+1)} = \frac{1}{|S_\ell^{(r)}|} \sum_{\boldsymbol{\mathcal{X}}_h \in S_\ell^{(r)}} \boldsymbol{\mathcal{X}}_h, \quad \ell = 1, ..., k.$$

3. Update $r \leftarrow r + 1$.

The algorithm converges when the assignments no longer change. However, the iterative refinement process ends when the maximum number of iterations allowed is reached.

Appendix C

The Iterative Re-weighted Least Squares algorithm for fitting the Time-Weighted M-quantile models in Chapter 5

This appendix describes an adaptation of the iterative re-weighted least squares (IRLS) algorithm used to fit the TWMQ linear models (Bugallo et al., 2024e) in Section 5.4. The reason we chose to use MQ regression models –rather than other alternative robust methods– has much to do with the advantages of the fitting process. On the one hand, standard quantile regression fitting algorithms are based on linear programming methods and do not necessarily guarantee convergence to a unique solution (Koenker and Machado, 1999). In such cases, it is typical to use the simplex method (primal, dual or primal-dual). In contrast, the simple IRLS algorithm used to fit a MQ regression model (Holland and Welsch, 1977; Street et al., 1988) guarantees convergence to a unique solution for a continuous monotone influence function (Bianchi and Salvati, 2015). The IRLS is used to fit all MQ models formulated in Chapter 5 (MQ2 linear models, MQ3 linear models and TWMQ linear models), but here we will focus on the particular case of the models proposed in this thesis: the TWMQ linear models. Even so, the template is common to all of them.

As far as the model fitting for the simulations (see Section 5.5) and the application to real data (see Section 5.7) are concerned, we have used a code developed by the authors in the programming language `R`. Nevertheless, a section for `R` codes is not included because they are not yet available in any online repository.

Let us start by recalling that the TWMQ linear models are defined around equation (5.14). For $0 < q < 1$, $t = 1, \ldots, T$, the model parameters $\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)$ are estimated as

$$\widehat{\boldsymbol{\beta}}_\psi(q, \boldsymbol{w}_t) = \underset{\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{i\in\mathcal{T}_t} w_{ti} \sum_{j=1}^{n_{di}} \rho_q\big(y_{dij} - \boldsymbol{x}'_{dij}\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t), \widehat{\sigma}_{qt}\big),$$

or as the solution of the system of p estimating equations

$$\sum_{d=1}^{D} \sum_{i\in\mathcal{T}_t} w_{ti} \sum_{j=1}^{n_{di}} \psi_q\big(y_{dij} - \boldsymbol{x}'_{dij}\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t), \widehat{\sigma}_{qt}\big) x_{dijk} = 0,\ k = 1, \ldots, p, \tag{C.1}$$

where ρ_q and ψ_q are defined in (5.1) and (5.2), respectively, ψ is the Huber function (5.3) and

$$\widehat{\sigma}_{qt} = \widehat{\mathrm{var}}^{1/2}(e_{\psi,dij}(q,\boldsymbol{w}_t)) = \widehat{\sigma}_\psi(\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)) = \mathrm{mad}_{\psi,n}(q,\boldsymbol{w}_t)/0.6745$$

is a robust estimator of σ_{qt}, for $\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)$ known, and $\mathrm{mad}_{\psi,n}(q,\boldsymbol{w}_t)$ is the median absolute deviation (MAD) of $e_{\psi,dij}(q,\boldsymbol{w}_t) = y_{dij} - \boldsymbol{x}'_{dij}\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)$, i.e.

$$\mathrm{mad}_{\psi,n}(q,\boldsymbol{w}_t) = \mathrm{median}\big\{|e_{\psi,dij}(q,\boldsymbol{w}_t) - \mathrm{med}_{\psi,n}(q,\boldsymbol{w}_t)| : d=1,\dots,D, i=1,\dots,T, j=1,\dots,n_{di}\big\},$$
$$\mathrm{med}_{\psi,n}(q,\boldsymbol{w}_t) = \mathrm{median}\big\{e_{\psi,dij}(q,\boldsymbol{w}_t) : d=1,\dots,D, i=1,\dots,T, j=1,\dots,n_{di}\big\}.$$

As $\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)$ is unknown, $e_{\psi,dij}(q,\boldsymbol{w}_t)$ cannot be calculated and therefore, neither $\widehat{\sigma}_{qt}$, $0 < q < 1$, $t = 1,\dots,T$. Consequently, we have implemented an iterative procedure to solve the system (C.1) of p non-linear equations.

We define the weights $w_{\psi dij}(q,\boldsymbol{w}_t) = \psi_q(e_{\psi,dij}(q,\boldsymbol{w}_t),\widehat{\sigma}_{qt})/e_{\psi,dij}(q,\boldsymbol{w}_t)$, and the t-relevant vectors $\boldsymbol{y}_{s(t)} = \underset{1\leqslant d\leqslant D}{\mathrm{col}}\big(\underset{i\in\mathcal{T}_t}{\mathrm{col}}\big(\underset{1\leqslant j\leqslant n_{di}}{\mathrm{col}}(y_{dij})\big)\big)$ and matrices $X_{s(t)} = \underset{1\leqslant d\leqslant D}{\mathrm{col}}\big(\underset{i\in\mathcal{T}_t}{\mathrm{col}}\big(\underset{1\leqslant j\leqslant n_{di}}{\mathrm{col}}(\boldsymbol{x}'_{dij})\big)\big)$ and $W_{s(t)}(q,\boldsymbol{w}_t) = \underset{1\leqslant d\leqslant D}{\mathrm{diag}}\big(\mathrm{diag}\big(\underset{i\in\mathcal{T}_t}{\mathrm{diag}}\big(\underset{1\leqslant j\leqslant n_{di}}{\mathrm{diag}}(w_{ti}w_{\psi dij}(q,\boldsymbol{w}_t))\big)\big)\big)$. We write equations (C.1) as

$$\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t} w_{ti}\sum_{j=1}^{n_{di}} w_{\psi dij}(q,\boldsymbol{w}_t)\big(y_{dij} - \boldsymbol{x}'_{dij}\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)\big)x_{dijk} = 0, \quad k = 1,\dots,p, \tag{C.2}$$

or in the matrix form $X'_{s(t)}W_{s(t)}(q,\boldsymbol{w}_t)\boldsymbol{y}_{s(t)} - X'_{s(t)}W_{s(t)}(q,\boldsymbol{w}_t)X_{s(t)}\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t) = 0$.

If $X'_{s(t)}W_{s(t)}(q,\boldsymbol{w}_t)X_{s(t)}$ is invertible, we write (C.2) in explicit form, i.e.

$$\begin{aligned}\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t) &= \big(X'_{s(t)}W_{s(t)}(q,\boldsymbol{w}_t)X_{s(t)}\big)^{-1}X'_{s(t)}W_{s(t)}(q,\boldsymbol{w}_t)\boldsymbol{y}_{s(t)}\\ &= \Big(\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{di}} w_{ti}w_{\psi dij}(q,\boldsymbol{w}_t)\boldsymbol{x}_{dij}\boldsymbol{x}'_{dij}\Big)^{-1}\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{di}} w_{ti}w_{\psi dij}(q,\boldsymbol{w}_t)\boldsymbol{x}_{dij}y_{dij}.\end{aligned}$$

This yields to the following IRLS algorithm to calculate $\widehat{\boldsymbol{\beta}}_\psi(q,\boldsymbol{w}_t)$.

1. Set the initial values $\widehat{\boldsymbol{\beta}}^{(0)}_\psi(q,\boldsymbol{w}_t)$ using e.g. the weighted least squares estimator

$$\widehat{\boldsymbol{\beta}}^{(0)}_\psi(q,\boldsymbol{w}_t) = \Big(\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t} w_{ti}\sum_{j=1}^{n_{di}}\boldsymbol{x}_{dij}\boldsymbol{x}'_{dij}\Big)^{-1}\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t} w_{ti}\sum_{j=1}^{n_{di}}\boldsymbol{x}_{dij}y_{dij}. \tag{C.3}$$

2. For each iteration $l = 1, 2, \dots$, do

   2.1. Calculate $\widehat{e}^{(l-1)}_{\psi,dij}(q,\boldsymbol{w}_t) = y_{dij} - \boldsymbol{x}'_{dij}\widehat{\boldsymbol{\beta}}^{(l-1)}_\psi(q,\boldsymbol{w}_t)$, $\widehat{\sigma}^{(l-1)}_{qt} = \widehat{\sigma}_\psi(\widehat{\boldsymbol{\beta}}^{(l-1)}_\psi(q,\boldsymbol{w}_t))$ and

$$w^{(l-1)}_{\psi dij}(q,\boldsymbol{w}_t) = \psi_q\big(\widehat{e}^{(l-1)}_{\psi,dij}(q,\boldsymbol{w}_t),\widehat{\sigma}^{(l-1)}_{qt}\big)/\widehat{e}^{(l-1)}_{\psi,dij}(q,\boldsymbol{w}_t),$$
$$W^{(l-1)}_{s(t)}\big(q,\boldsymbol{w}_t\big) = \underset{1\leqslant d\leqslant D}{\mathrm{diag}}\big(\mathrm{diag}\big(\underset{i\in\mathcal{T}_t}{\mathrm{diag}}\big(\underset{1\leqslant j\leqslant n_{di}}{\mathrm{diag}}(w_{ti}w^{(l-1)}_{\psi dij}(q,\boldsymbol{w}_t))\big)\big)\big).$$

   2.2. Update the estimator of $\boldsymbol{\beta}_\psi(q,\boldsymbol{w}_t)$. i.e.

$$\widehat{\boldsymbol{\beta}}^{(l)}_\psi(q,\boldsymbol{w}_t) = \Big(\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{di}} w_{ti}w^{(l-1)}_{\psi dij}(q,\boldsymbol{w}_t)\boldsymbol{x}_{dij}\boldsymbol{x}'_{dij}\Big)^{-1}\sum_{d=1}^{D}\sum_{i\in\mathcal{T}_t}\sum_{j=1}^{n_{di}} w_{ti}w^{(l-1)}_{\psi dij}(q,\boldsymbol{w}_t)\boldsymbol{x}_{dij}y_{dij}.$$

3. Repeat Step 2 until convergence.

Appendix D

Proof of Theorems 1 and 2 in Section 5.4

This appendix provides technical specifications and step-by-step proofs of Theorems 1 and 2 in Section 5.4. In Section D.1, Theorem 1 derives a first-order approximation of the MSE of the robust bias-corrected temporal MQ (BTMQ) predictor (5.20) derived from the TWMQ linear models (5.14) and proposes an analytical estimator. In Section D.2, Theorem 2 presents an optimal criterion for the selection of the robustness parameter for bias correction and proves the existence and uniqueness of the solution.

D.1 First-order approximation of the mean squared error of the bias-corrected temporal M-quantile predictor

Let $d = 1, \ldots, D$ an area and $t = 1, \ldots, T$ a time period. Section 5.4 focuses on the prediction of the population means $\overline{Y}_{dt}$, which have been defined in (5.9). To start with, the BTMQ predictor of $\overline{Y}_{dt}$ can be written as

$$\begin{aligned}
\widehat{\overline{Y}}_{dt}^{btmq} &= \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}\big(\widehat{\theta}_d, \boldsymbol{w}_t\big)\right\} + \frac{1}{n_{dt}}\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\widehat{B}_{dt}^{btmq}, \\
\widehat{B}_{dt}^{btmq} &= \sum_{j\in s_{dt}} \sigma_{\theta_{dt}}\phi\big(\widehat{u}_{\psi,dtj}\big).
\end{aligned}$$

The idea of the proof is to decompose the predictor $\widehat{\overline{Y}}_{dt}^{btmq}$ to take into account the variability derived from the estimation of q_{dtj}, $j = 1, \ldots, n_{dt}$, and $\boldsymbol{\beta}_{\psi}\big(\theta_d, \boldsymbol{w}_t\big)$. To do so, we first define the following auxiliary notation in relation to the BTMQ predictor:

$$\begin{aligned}
\overline{Y}_{dt}^{btmq} &= \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_{\psi}(\theta_d, \boldsymbol{w}_t)\right\} + \frac{1}{n_{dt}}\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)B_{dt}^{btmq}, \\
\widetilde{\overline{Y}}_{dt}^{btmq} &= \frac{1}{N_{dt}}\left\{\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_{\psi}(\theta_d, \boldsymbol{w}_t)\right\} + \frac{1}{n_{dt}}\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\widetilde{B}_{dt}^{btmq}, \\
B_{dt}^{btmq} &= \sum_{j\in s_{dt}} \sigma_{\theta_{dt}}\phi(u_{\psi,dtj}), \quad \widetilde{B}_{dt}^{btmq} = \sum_{j\in s_{dt}} \sigma_{\theta_{dt}}\phi(\widetilde{u}_{\psi,dtj}),
\end{aligned}$$

where $u_{\psi,dtj}$ and $\widetilde{u}_{\psi,dtj}$, $j = 1, \ldots, n_{dt}$, have been defined in (5.15) and (5.18), respectively.

The approximation of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$ that we propose below accounts for the randomness of the unit-level MQ coefficients q_{dtj}, $j = 1, \ldots, N_{dt}$, coming from the TWMQ linear models, but assumes that θ_d and $\widehat{\theta}_d$ are known. In fact, these values and their estimates are derived from the MQ3 linear models (5.7). Not least, the standard deviations $\sigma_{\theta_d t}$ are assumed to be known. The randomness arising from the estimation of θ_d and $\sigma_{\theta_d t}$ is of minor importance (Chambers and Tzavidis, 2006) and omitting it is a common practice when estimating the MSE of predictors derived from MQ models (Chambers et al., 2011).

Appendix D.1.1 introduces the probabilistic framework and necessary assumptions to obtain a first-order asymptotic approximation of the MSE of the BTMQ predictor. Generally, all assumptions are reasonable and can be found in the literature.

D.1.1 Assumptions

This section presents a set of assumptions necessary to obtain a first-order approximation of $MSE(\widehat{\overline{Y}}_{dt}^{btmq})$. These are all reasonable practical requirements and are met under general conditions. Many have been previously proposed in the literature (Bianchi and Salvati, 2015) or are direct adaptations of assumptions required for mixed models to MQ regression.

For the influence function $\phi(u)$, we assume that

(Φ1) ϕ is differentiable at $u = 0$, with $\phi(0) = 0$ and $\dot{\phi}(0) = 1$. If $|u| \geqslant c_\phi$, $\phi(u) = c_\phi \operatorname{sgn}(u)$,

where $\operatorname{sgn}(u) = 1$ if $u > 0$; $\operatorname{sgn}(u) = -1$ if $u < 0$ and $\operatorname{sgn}(u) = 0$ if $u = 0$.
This assumption is quite common for influence functions in the field of robust statistics (Huber, 1981). From assumption(Φ1), the non-atypical data subsets are $\mathcal{G}_{dt} = \{j \in s_{dt} : |u_{\psi,dtj}| < c_\phi\}$, $\widetilde{\mathcal{G}}_{dt} = \left\{j \in s_{dt} : |\widetilde{u}_{\psi,dtj}| < c_\phi\right\}$ and $\widehat{\mathcal{G}}_{dt} = \left\{j \in s_{dt} : |\widehat{u}_{\psi,dtj}| < c_\phi\right\}$, and the intersection subsets are $\widetilde{\mathcal{H}}_{dt} = \mathcal{G}_{dt} \cap \widetilde{\mathcal{G}}_{dt}$, $\widehat{\mathcal{H}}_{dt} = \widetilde{\mathcal{G}}_{dt} \cap \widehat{\mathcal{G}}_{dt}$ and $\mathcal{G}_{dt} \cap \widetilde{\mathcal{G}}_{dt} \cap \widehat{\mathcal{G}}_{dt} = \widetilde{\mathcal{H}}_{dt} \cap \widehat{\mathcal{H}}_{dt}$. These subsets will be useful in the calculation of expected values and variances.

From assumption (Φ1), we simplify the notation and write

$$
\begin{aligned}
B_{dt}^{btmq} &= \sum_{j \in \mathcal{G}_{dt}} \sigma_{\theta_d t} \phi(u_{\psi,dij}) + c_\phi \sum_{j \in s_{dt} - \mathcal{G}_{dt}} \operatorname{sgn}(e_{\psi,dtj}), \\
\widetilde{B}_{dt}^{btmq} &= \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \sigma_{\theta_d t} \phi(\widetilde{u}_{\psi,dtj}) + c_\phi \sum_{j \in s_{dt} - \widetilde{\mathcal{G}}_{dt}} \operatorname{sgn}\big(\widetilde{e}_{\psi,dtj}\big), \\
\widehat{B}_{dt}^{btmq} &= \sum_{j \in \widehat{\mathcal{G}}_{dt}} \sigma_{\theta_d t} \phi(\widehat{u}_{\psi,dtj}) + c_\phi \sum_{j \in s_{dt} - \widehat{\mathcal{G}}_{dt}} \operatorname{sgn}\big(\widehat{e}_{\psi,dtj}\big).
\end{aligned}
$$

Below we include additional notation for model errors and pseudo-residuals that will be necessary for the calculation of expected values and variances. Related to the variables $\operatorname{sgn}(e_{\psi,dtj})$, we define the probabilities

$$
\pi_{dtj} = P\big(\operatorname{sgn}(e_{\psi,dtj}) = -1\big), \quad 1 - \pi_{dtj} = P\big(\operatorname{sgn}(e_{\psi,dtj}) = 1\big), \quad j = 1, \ldots, N_{dt},
$$

so that

$$E[\text{sgn}(e_{\psi,dtj})] = 1 - 2\pi_{dtj}, \quad \text{var}(\text{sgn}(e_{\psi,dtj})) = 4\pi_{dtj} - 4\pi_{dtj}^2.$$

Related to the variables $\text{sgn}(\widetilde{e}_{\psi,dtj}) - \text{sgn}(e_{\psi,dtj})$, we define the probabilities

$$\widetilde{\pi}_{a,dtj} = P\big(\text{sgn}(\widetilde{e}_{\psi,dtj}) - \text{sgn}(e_{\psi,dtj}) = a\big), \quad a = -2, 0, 2, \quad j = 1, \ldots, N_{dt},$$

so that $E[\text{sgn}(\widetilde{e}_{\psi,dtj}) - \text{sgn}(e_{\psi,dtj})] = 2(\widetilde{\pi}_{2,dtj} - \widetilde{\pi}_{-2,dtj})$ and

$$\text{var}(\text{sgn}(\widetilde{e}_{\psi,dtj}) - \text{sgn}(e_{\psi,dtj})) = 4(\widetilde{\pi}_{2,dtj} + \widetilde{\pi}_{-2,dtj}) - 4(\widetilde{\pi}_{2,dtj} - \widetilde{\pi}_{-2,dtj})^2.$$

Related to the variables $\text{sgn}(\widehat{e}_{\psi,dtj}) - \text{sgn}(\widetilde{e}_{\psi,dtj})$, we define the probabilities

$$\widehat{\pi}_{a,dtj} = P\big(\text{sgn}(\widehat{e}_{\psi,dtj}) - \text{sgn}(\widetilde{e}_{\psi,dtj}) = a\big), \quad a = -2, 0, 2, \quad j = 1, \ldots, N_{dt},$$

so that $E[\text{sgn}(\widehat{e}_{\psi,dtj}) - \text{sgn}(\widetilde{e}_{\psi,dtj})] = 2(\widehat{\pi}_{2,dtj} - \widehat{\pi}_{-2,dtj})$ and

$$\text{var}(\text{sgn}(\widehat{e}_{\psi,dtj}) - \text{sgn}(\widetilde{e}_{\psi,dtj})) = 4(\widehat{\pi}_{2,dtj} + \widehat{\pi}_{-2,dtj}) - 4(\widehat{\pi}_{2,dtj} - \widehat{\pi}_{-2,dtj})^2.$$

The asymptotic theory will be developed under the following assumptions. For the sample sizes, we assume

(N1) There exist $0 < \pi_{dt} < 1$ such that $\sum_{d=1}^{D}\sum_{t=1}^{T} \pi_{dt} = 1$ and $\frac{n_{dt}}{n} \to \pi_{dt}$ as $n \to \infty$.
(N2) There exist $0 < f_{dt} < 1$ such that $\frac{n_{dt}}{N_{dt}} \to f_{dt}$ as $n \to \infty$.

The asymptotic assumption (N1) avoid the possibility of domains with zero sample size. Assumption (N2) states that sample sizes and population sizes converge to the sampling fractions reasonably far from the extremes values 0 and 1.

For the unit-level MQ coefficients, we assume

(Q1) For $i \in \mathcal{T}_t$, $j \in s_{di}$, q_{dij} are independent variables with common variance $\xi_{dt}^2 = \text{var}(q_{dij})$.

The unit-level MQ coefficients q_{dij}, defined in (5.17), play a similar role to that of random intercepts in LMMs. In such models, it is common to assume that the random effects have constant variance. Because of the parallelism that can be established between methodologies based on MQ models and mixed models, we include assumption (Q1). Thus, it is just a bridge between the MQ methodology and mixed models.

For the estimator of the vector of regression parameters $\widehat{\boldsymbol{\beta}}_\psi(q, \boldsymbol{w}_t)$, $0 < q < 1$, we adapt the conditions proposed in Bianchi and Salvati (2015) and write

(A1) $\boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t) \in \Theta \subset \mathbb{R}^p$ is a twice-differentiable continuous function in its first component, q, where Θ is a compact subset of $\mathbb{R}^p$.

(A2) ψ is continuous, bounded and with bounded derivative, except at a finite number of points.

(A3) For $i \in \mathcal{T}_t$, $j \in s_{di}$, $E[|\boldsymbol{x}_{dij}|^4] < \infty$ and $E[|e_{dij}|^4] < \infty$.

(A4) For $i \in \mathcal{T}_t$, $j \in s_{di}$, $E[\boldsymbol{x}_{dij}\boldsymbol{x}'_{dij}\dot{\psi}(e_{\psi,dij}(q,\boldsymbol{w}_t),\sigma_{qt})]$ is uniformly non singular, where $\dot{\psi}_{qt}$ is the partial derivative of ψ_{qt} with respect to the first argument.

(A5) The preliminary estimator $\widehat{\boldsymbol{\beta}}^{(0)}_{\psi}(q,\boldsymbol{w}_t)$ of $\boldsymbol{\beta}_{\psi}(q,\boldsymbol{w}_t)$, defined in (C.3), is such that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(0)}_{\psi}(q,\boldsymbol{w}_t) - \boldsymbol{\beta}_{\psi}(q,\boldsymbol{w}_t)) = O_p(1).$$

(A6) $\exists\delta_1 > 0/\forall e \in (-\delta_1,\delta_1)\ \exists\dot{F}_{qt}(\text{med}_{\psi,n}(q,\boldsymbol{w}_t)+e)$ and is continuous and positive at $e=0$. $\dot{F}_{qt}$ is the first order derivative of F_{qt}, which is the c.d.f. of $e_{\psi,dij}(q,\boldsymbol{w}_t)$, $i \in \mathcal{T}_t$, $j \in s_{di}$.

(A7) $\exists\delta_2 > 0/\forall e \in (-\delta_2,\delta_2)\ \exists\dot{F}_{qt}(\text{med}_{\psi,n}(q,\boldsymbol{w}_t) \pm 0.6745\sigma_{qt}+e)$ and is continuous at $e=0$.

(A8) $\exists\delta_3 > 0/\forall e \in (\sigma_{qt}-\delta_3,\sigma_{qt}+\delta_3)\ \exists\dot{F}_{qt}(\text{med}_{\psi,n}(q,\boldsymbol{w}_t)+e) + \dot{F}_{qt}(\text{med}_{\psi,n}(q,\boldsymbol{w}_t)-e) > 0$.

Assumption (A1) is necessary to calculate Taylor polynomials. Assumption (A2) holds for the Huber function. Assumption (A3) is a technical moment condition required for the application of the Uniform Law of Large Numbers and the asymptotic representation. Assumption (A4) is an identifiability condition. Assumptions (A5)-(A8) are needed for the Bahadur representation of the median absolute deviation (MAD) estimator (see Welsh (1986)). In the case of the Huber influence function (5.3), assumptions (A1) and (A2) are satisfied. To guarantee assumption (A4), one may require that for any $\boldsymbol{\beta}_{\psi}(q,\boldsymbol{w}_t) \in \Theta$ and $c > 0$,

$$P(\sigma_{qt}^{-1}|y_{dij} - \boldsymbol{x}'_{dij}\boldsymbol{\beta}_{\psi}(q,\boldsymbol{w}_t)| \leqslant c|\,\boldsymbol{x}_{dij}) > \varepsilon > 0,\ i \in \mathcal{T}_t,\ j \in s_{di}.$$

In practice, this is verified if most of the residuals belong to the strictly convex region of ψ. Under assumptions (A1)-(A8), it holds that $\widehat{\boldsymbol{\beta}}_{\psi}(\theta_d,\boldsymbol{w}_t) - \boldsymbol{\beta}_{\psi}(\theta_d,\boldsymbol{w}_t) = O_p(n^{-1/2})$. Finally, to complete the assumptions associated with the vector of regression parameters, we include

(A9) $\exists\delta_4 > 0/\forall\theta \in (\theta_d-\delta_4,\theta_d+\delta_4)$, $\widehat{\boldsymbol{\beta}}_{\psi}(\theta,\boldsymbol{w}_t) - \boldsymbol{\beta}_{\psi}(\theta_d,\boldsymbol{w}_t) = O_p(n^{-1/2})$.

Assumption (A9) is needed to maintain the asymptotic plausibility of assumption (A8) in a neighborhood of θ_d, as in practice θ_d is substituted by $\widehat{\theta}_d$.

For the MSE of the BTMQ predictor, we assume three groups of assumptions. The first group of assumptions concerns the model errors (5.15) of the TWMQ linear models (5.14):

(B1) $\frac{1}{N_{dt}}\sum_{j\in U_{dt}} E\big[e^2_{\psi,dtj}\big] = O(1)$.
(B2) $\frac{1}{N_{dt}}\sum_{j\in U_{dt}} \pi_{dtj} = O(1)$ and $\frac{1}{N_{dt}}\sum_{j\in U_{dt}} \pi^2_{dtj} = O(1)$.
(B3) $\frac{1}{n_{dt}}\sum_{j\in s_{dt}} E\big[e_{\psi,dtj}\big] = \frac{1}{N_{dt}-n_{dt}}\sum_{j\in r_{dt}} E\big[e_{\psi,dtj}\big] + o(1)$.
(B4) $\exists\delta_5 > 0$: $\forall\theta \in (\theta_d-\delta_5,\theta_d+\delta_5)$, $\frac{1}{N_{dt}}\sum_{j\in U_{dt}} \Big(\boldsymbol{x}'_{dtj}\frac{\partial^2\boldsymbol{\beta}_{\psi}(q,\boldsymbol{w}_t)}{\partial q^2}\Big|_{q=\theta}\Big)^2 = O(1)$.

Assumption (B1) states that the second order moment of the model errors is bounded in average. Assumption (B2) is fulfilled as π_{dtj}'s are the probabilities that the model errors are negative. Assumption (B3) states that the sample average of the expected model errors

behaves similarly in the sample and the non-sample subsets. Assumption (B4) is a technical condition required for the application of the Uniform Law of Large Numbers and the asymptotic representation.

The second group of assumptions concerns the pseudo-residuals (5.18) of the TWMQ linear models (5.14). The assumptions are

(C1) $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} E\big[(\widetilde{e}_{\psi,dtj} - e_{\psi,dtj})^4\big] = o(1)$.
(C2) $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} \widetilde{\pi}_{a,dtj} = o(1)$ and $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} \widetilde{\pi}^2_{a,dtj} = o(1)$, $a = -2, 2$.
(D1) $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} E\big[(\widehat{e}_{\psi,dtj} - \widetilde{e}_{\psi,dtj})^4\big] = o(1)$.
(D2) $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} \widehat{\pi}^b_{a,dtj} = o(1)$ and $\frac{1}{N_{dt}} \sum_{j \in U_{dt}} \widehat{\pi}^2_{a,dtj} = o(1)$, $a = -2, 2$.

Assumptions (C1) and (D1) state that the fourth order moment of the differences between pseudo-residuals and model errors is bounded in average. Assumptions (C2) and (D2) are fulfilled as $\widetilde{\pi}_{a,dtj}$'s and $\widehat{\pi}_{a,dtj}$'s are the probabilities that the pseudo-residuals are negative. The third group of assumptions concerns the independence of the model errors and pseudo-residuals of the TWMQ linear models (5.14). The assumptions are

(E1) For $j \in U_{dt}$, $e_{\psi,dtj}$ are independent random variables.
(E2) For $j \in U_{dt}$, $\widetilde{e}_{\psi,dtj}$ are independent random variables.
(E3) For $j \in U_{dt}$, $\widehat{e}_{\psi,dtj}$ are independent random variables.

It is common in mixed models and MQ models to include independence assumptions such as (E1), (E2) and (E3) for the model errors and the residuals and pseudo-residuals.

D.1.2 Part I: Dealing with the differences $\overline{Y}^{btmq}_{dt} - \overline{Y}_{dt}$

In this section, we calculate the expected value and variance of the prediction differences

$$\overline{Y}^{btmq}_{dt} - \overline{Y}_{dt}, \quad d = 1, \ldots, D, \ t = 1, \ldots, T.$$

The reasoning is based on the calculation of first-order Taylor approximations and the subsequent computation of expected values and variances.

First, a Taylor series expansion of $\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t)$ around θ_d yields to

$$\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t) = \boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) + \frac{\partial \boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)}{\partial q}\Big|_{q=\theta_d} (q_{dtj} - \theta_d) + \boldsymbol{r}_{\psi,dtj}(\theta_d), \quad j = 1, \ldots, N_{dt}, \quad \text{(D.1)}$$

where $\boldsymbol{r}_{\psi,dtj}(\theta_d) \stackrel{\Delta}{=} \boldsymbol{r}_{dtj} = (r_{dtj1}, \ldots, r_{dtjp})'$ and $r_{dtjk} = O_p\big((q_{dtj} - \theta_d)^2\big)$, $k = 1, \ldots, p$, with

$$\|\boldsymbol{r}_{dtj}\|_2 = \frac{1}{2}(q_{dtj} - \theta_d)^2 \Big\| \frac{\partial^2 \boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)}{\partial q^2}\Big|_{q=\theta^*_{dtj}} \Big\|_2 \leqslant \Big\| \frac{\partial^2 \boldsymbol{\beta}_\psi(q, \boldsymbol{w}_t)}{\partial q^2}\Big|_{q=\theta^*_{dtj}} \Big\|_2, \quad |\theta^*_{dtj} - \theta_d| < |q_{dtj} - \theta_d|.$$

We introduce the notation

$$\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t) = \big(\kappa_{\psi 1}(\theta_d, \boldsymbol{w}_t), \ldots, \kappa_{\psi p}(\theta_d, \boldsymbol{w}_t)\big)'; \quad \kappa_{\psi k}(\theta_d, \boldsymbol{w}_t) = \frac{\partial \beta_{\psi k}(q, \boldsymbol{w}_t)}{\partial q}\Big|_{q=\theta_d}, \ k = 1, \ldots, p.$$

From the Taylor series expansion (D.1), the model errors defined in (5.15) is written as

$$\begin{aligned} e_{\psi,dtj} &= y_{dtj} - \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) = \boldsymbol{x}'_{dtj}(\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t) - \boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t)) \\ &= \boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)(q_{dtj} - \theta_d) + \boldsymbol{x}'_{dtj}\boldsymbol{r}_{dtj}, \quad j = 1, \ldots, N_{dt}. \end{aligned}$$

From assumption (Q1), $\mathrm{var}(q_{dtj} - \theta_d) = \mathrm{var}(q_{dtj}) = \xi_{dt}^2$, $j = 1, \ldots, N_{dt}$, and

$$\mathrm{var}(e_{\psi,dtj}) = \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\big)^2\xi_{dt}^2 + \boldsymbol{x}'_{dtj}\mathrm{var}(\boldsymbol{r}_{dtj})\boldsymbol{x}_{dtj} + \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\big)^2\mathrm{cov}\big(q_{dtj}, q_{dtj}^2\big). \quad \text{(D.2)}$$

Assumption (B4) implies that

$$\begin{aligned} \frac{1}{n_{dt}}\sum_{j\in s_{dt}} \boldsymbol{x}'_{dtj}\mathrm{var}(\boldsymbol{r}_{dtj})\boldsymbol{x}_{dtj} &= \frac{1}{4n_{dt}}\sum_{j\in s_{dt}} \Big\{\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\boldsymbol{\kappa}'_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\boldsymbol{x}_{dtj}\Big\}\mathrm{var}(q_{dtj}^2) \\ &\leqslant \frac{1}{4n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\big)^2 = O(1), \\ \Big|\frac{1}{n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\big)^2 E[q_{dtj}^3]\Big| &\leqslant \frac{1}{n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\big)^2 \big|E[q_{dtj}]\big|^3 = O(1), \\ \Big|\frac{1}{n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\big)^2 E[q_{dtj}]\Big| &\leqslant \frac{1}{n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta^*_{dtj}, \boldsymbol{w}_t)\big)^2 \big|E[q_{dtj}]\big| = O(1). \end{aligned}$$

Collecting these results, we obtain that

$$\frac{1}{n_{dt}}\sum_{j\in s_{dt}} \mathrm{var}(e_{\psi,dtj}) = \frac{1}{n_{dt}}\sum_{j\in s_{dt}} \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\big)^2\xi_{dt}^2 + O(1). \quad \text{(D.3)}$$

The corresponding standardized model errors, defined in (5.15), can be written as

$$u_{\psi,dtj} = \sigma_{\theta_d t}^{-1}\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)(q_{dtj} - \theta_d) + \sigma_{\theta_d t}^{-1}\boldsymbol{x}'_{dtj}\boldsymbol{r}_{dtj}, \quad j = 1, \ldots, N_{dt}.$$

As $\phi(0) = 0$ and $\dot\phi(0) = 1$, a Taylor series expansion of $\phi(u_{\psi,dtj})$ around $u = 0$ yields to

$$\phi(u_{\psi,dtj}) = \phi(0) + \dot\phi(0)u_{\psi,dtj} + R_{dtj} = u_{\psi,dtj} + R_{dtj}, \quad j = 1, \ldots, N_{dt}, \quad \text{(D.4)}$$

where $R_{dtj} = \frac{1}{2}\,u^{*2}_{\psi,dtj}$, $0 < |u^*_{\psi,dtj}| < |u_{\psi,dtj}|$ and $\mathrm{var}(R_{dtj}) \leqslant \mathrm{var}(u_{\psi,dtj})$.

From assumption (B1), we have

$$\frac{1}{n_{dt}}\sum_{j\in\mathcal{G}_{dt}} E[R_{dtj}] = O(1), \quad \frac{1}{n_{dt}}\sum_{j\in\mathcal{G}_{dt}} \mathrm{var}(R_{dtj}) = O(1). \quad \text{(D.5)}$$

From the Taylor series expansion (D.4), we have

$$\begin{aligned} B_{dt}^{btmq} &= \sum_{j\in\mathcal{G}_{dt}} \sigma_{\theta_d t}\phi(u_{\psi,dtj}) + c_\phi \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \mathrm{sgn}\big(e_{\psi,dtj}\big) \\ &= \sum_{j\in\mathcal{G}_{dt}} e_{\psi,dtj} + \sigma_{\theta_d t}\sum_{j\in\mathcal{G}_{dt}} R_{dtj} + c_\phi \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \mathrm{sgn}\big(e_{\psi,dtj}\big). \end{aligned}$$

The prediction difference $\overline{Y}_{dt}^{(1)} = \overline{Y}_{dt}^{btmq} - \overline{Y}_{dt}$ is

$$\begin{aligned}
\overline{Y}_{dt}^{(1)} &= \frac{1}{N_{dt}}\Big(\sum_{j\in s_{dt}} y_{dtj} + \sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t)\Big) + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}B_{dt}^{btmq} - \frac{1}{N_{dt}}\sum_{j\in U_{dt}} \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t) \\
&= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}B_{dt}^{btmq} + \frac{1}{N_{dt}}\Big(\sum_{j\in r_{dt}} \boldsymbol{x}'_{dtj}(\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) - \boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t))\Big) \\
&= \sum_{j\in U_{dt}} \left(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{\mathcal{G}_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\right) e_{\psi,dtj} \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{c_\phi}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \operatorname{sgn}(e_{\psi,dtj}) + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j\in\mathcal{G}_{dt}} R_{dtj}. \qquad \text{(D.6)}
\end{aligned}$$

From assumptions (E1) and (B2), we obtain that

$$\begin{aligned}
\frac{1}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} E[\operatorname{sgn}(e_{\psi,dtj})] &= \frac{1}{n_{dt}}\Big(\operatorname{card}(s_{dt}-\mathcal{G}_{dt}) - 2\sum_{j\in s_{dt}-\mathcal{G}_{dt}} \pi_{dtj}\Big) = O(1), \\
\frac{1}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \operatorname{var}(\operatorname{sgn}(e_{\psi,dtj})) &= \frac{4}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \pi_{dtj}(1-\pi_{dtj}) = O(1). \qquad \text{(D.7)}
\end{aligned}$$

From (D.3), (D.6), (D.5), (D.7) and assumptions (B4) and (E1), the variance of $\overline{Y}_{dt}^{(1)}$ is

$$\begin{aligned}
V_{dt}^{(1)} &= \operatorname{var}\big(\overline{Y}_{dt}^{(1)}\big) = \sum_{j\in U_{dt}} \left(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{\mathcal{G}_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\right)^2 \operatorname{var}(e_{\psi,dtj}) \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{c_\phi^2}{n_{dt}^2} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} \operatorname{var}(\operatorname{sgn}(e_{\psi,dtj})) + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\sigma_{\theta_d t}^2}{n_{dt}^2} \sum_{j\in\mathcal{G}_{dt}} \operatorname{var}(R_{dtj}) \\
&= \sum_{j\in U_{dt}} \left(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\mathcal{G}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\right) \big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\big)^2 \xi_{dt}^2 + o(n^{-1}).
\end{aligned}$$

From (D.6), (D.5), (D.7) and assumptions (B2) and (B3), we have that

$$\begin{aligned}
E_{dt}^{(1)} &= E\big[\overline{Y}_{dt}^{(1)}\big] = \sum_{j\in U_{dt}} \left(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{\mathcal{G}_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\right) E\big[e_{\psi,dtj}\big] \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{c_\phi}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} E\big[\operatorname{sgn}(e_{\psi,dtj})\big] + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j\in\mathcal{G}_{dt}} E[R_{dtj}] = \\
&= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{c_\phi}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} E\big[\operatorname{sgn}(e_{\psi,dtj})\big] + \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j\in\mathcal{G}_{dt}} E[R_{dtj}] + o(1).
\end{aligned}$$

From (D.5) and (D.7), it holds that $E_{dt}^{(1)} = O(1)$.

D.1.3 Part II: Dealing with the differences $\widetilde{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}^{btmq}$

In this section, we calculate the expected value and variance of the prediction differences

$$\widetilde{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}^{btmq}, \quad d = 1, \ldots, D,\ t = 1, \ldots, T.$$

The reasoning is based on the calculation of first-order Taylor approximations and the subsequent computation of expected values and variances.

As $\phi(0) = 0$ and $\dot{\phi}(0) = 1$, a Taylor series expansion of $\phi(\widetilde{u}_{\psi,dtj})$ around $u = 0$ yields to

$$\phi(\widetilde{u}_{\psi,dtj}) = \phi(0) + \dot{\phi}(0)\widetilde{u}_{\psi,dtj} + \widetilde{R}_{dtj} = \widetilde{u}_{\psi,dtj} + \widetilde{R}_{dtj}, \quad j = 1, \ldots, N_{dt}, \tag{D.8}$$

where $\widetilde{R}_{dtj} = \frac{1}{2}\,\widetilde{u}^{*2}_{\psi,dtj}$, $0 < |\widetilde{u}^{*}_{\psi,dtj}| < |\widetilde{u}_{\psi,dtj}|$ and $\mathrm{var}(\widetilde{R}_{dtj}) \leqslant \mathrm{var}(\widetilde{u}_{\psi,dtj})$.

From assumption (C1), we have

$$\frac{1}{n_{dt}} \sum_{j \in \widetilde{\mathcal{G}}_{dt}} E[\widetilde{R}_{dtj} - R_{dtj}] = o(1), \quad \frac{1}{n_{dt}} \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \mathrm{var}\big((\widetilde{R}_{dtj} - R_{dtj})^2\big) = o(1). \tag{D.9}$$

From the Taylor series expansion (D.8), we have

$$\begin{aligned}
\widetilde{B}^{btmq}_{dt} &= \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \sigma_{\theta_d t} \phi\big(\widetilde{u}_{\psi,dtj}\big) + c_\phi \sum_{j \in s_{dt} - \widetilde{\mathcal{G}}_{dt}} \mathrm{sgn}\big(\widetilde{e}_{\psi,dtj}\big), \\
&= \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \widetilde{e}_{\psi,dtj} + \sigma_{\theta_d t} \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \widetilde{R}_{dtj} + c_\phi \sum_{j \in s_{dt} - \widetilde{\mathcal{G}}_{dt}} \mathrm{sgn}(\widetilde{e}_{\psi,dtj}).
\end{aligned}$$

As $\widetilde{e}_{\psi,dtj} - e_{\psi,dtj} = \boldsymbol{x}'_{dtj}\big(\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\theta_d, \boldsymbol{w}_t)\big) = \widetilde{e}_{\psi,dtj}(\theta_d)$, then $B^{(2)}_{dt} = \widetilde{B}^{btmq}_{dt} - B^{btmq}_{dt}$ is

$$\begin{aligned}
B^{(2)}_{dt} &= \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \widetilde{e}_{\psi,dtj} + \sigma_{\theta_d t} \sum_{j \in \widetilde{\mathcal{G}}_{dt}} \widetilde{R}_{dtj} + c_\phi \sum_{j \in s_{dt} - \widetilde{\mathcal{G}}_{dt}} \mathrm{sgn}(\widetilde{e}_{\psi,dtj}) \\
&- \sum_{j \in \mathcal{G}_{dt}} e_{\psi,dtj} - \sigma_{\theta_d t} \sum_{j \in \mathcal{G}_{dt}} R_{dtj} - c_\phi \sum_{j \in s_{dt} - \mathcal{G}_{dt}} \mathrm{sgn}(e_{\psi,dtj}) \\
&= \sum_{j \in \widetilde{\mathcal{H}}_{dt}} \widetilde{e}_{\psi,dtj}(\theta_d) + \sum_{j \in \widetilde{\mathcal{G}}_{dt} - \widetilde{\mathcal{H}}_{dt}} \widetilde{e}_{\psi,dtj} - \sum_{j \in \mathcal{G}_{dt} - \widetilde{\mathcal{H}}_{dt}} e_{\psi,dtj} \\
&+ \sigma_{\theta_d t}\Big(\sum_{j \in \widetilde{\mathcal{H}}_{dt}} (\widetilde{R}_{dtj} - R_{dtj}) + \sum_{j \in \widetilde{\mathcal{G}}_{dt} - \widetilde{\mathcal{H}}_{dt}} \widetilde{R}_{dtj} - \sum_{j \in \mathcal{G}_{dt} - \widetilde{\mathcal{H}}_{dt}} R_{dtj} \Big) \\
&+ c_\phi\Big(\sum_{j \in s_{dt} - \widetilde{\mathcal{H}}_{dt}} (\mathrm{sgn}(\widetilde{e}_{\psi,dtj}) - \mathrm{sgn}(e_{\psi,dtj})) \\
&+ \sum_{j \in s_{dt} - (\widetilde{\mathcal{G}}_{dt} - \widetilde{\mathcal{H}}_{dt})} \mathrm{sgn}(\widetilde{e}_{\psi,dtj}) - \sum_{j \in s_{dt} - (\mathcal{G}_{dt} - \widetilde{\mathcal{H}}_{dt})} \mathrm{sgn}(e_{\psi,dtj}) \Big).
\end{aligned}$$

The prediction difference $\overline{Y}^{(2)}_{dt} = \overline{\widetilde{Y}}^{btmq}_{dt} - \overline{Y}^{btmq}_{dt}$ is

$$\begin{aligned}
\overline{Y}^{(2)}_{dt} &= \frac{1}{N_{dt}}\Big(\sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \widehat{\boldsymbol{\beta}}_\psi(\theta_d, \boldsymbol{w}_t) \Big) + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \frac{1}{n_{dt}} \widetilde{B}^{btmq}_{dt} \\
&- \frac{1}{N_{dt}}\Big(\sum_{j \in s_{dt}} y_{dtj} + \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) \Big) - \Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \frac{1}{n_{dt}} B^{btmq}_{dt} \\
&= \Big(1 - \frac{n_{dt}}{N_{dt}}\Big) \frac{1}{n_{dt}} B^{(2)}_{dt} - \frac{1}{N_{dt}} \sum_{j \in r_{dt}} \widetilde{e}_{\psi,dtj}(\theta_d).
\end{aligned}$$

By substituting $B_{dt}^{(2)}$, we obtain

$$\begin{aligned}
\overline{Y}_{dt}^{(2)} &= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Bigg\{\sum_{j\in\widetilde{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}(\theta_d)+\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}-\sum_{j\in\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt}}e_{\psi,dtj}\\
&+ \sigma_{\theta_d t}\Big(\sum_{j\in\widetilde{\mathcal{H}}_{dt}}(\widetilde{R}_{dtj}-R_{dtj})+\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt}}\widetilde{R}_{dtj}-\sum_{j\in\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt}}R_{dtj}\Big)\\
&+ c_\phi\Big(\sum_{j\in s_{dt}-\widetilde{\mathcal{H}}_{dt}}(\mathrm{sgn}(\widetilde{e}_{\psi,dtj})-\mathrm{sgn}(e_{\psi,dtj}))+\sum_{j\in s_{dt}-(\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\\
&- \sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{sgn}(e_{\psi,dtj})\Big)\Bigg\}-\frac{1}{N_{dt}}\sum_{j\in r_{dt}}\widetilde{e}_{\psi,dtj}(\theta_d).
\end{aligned}$$

We write $\overline{Y}_{dt}^{(2)}$ in the form

$$\begin{aligned}
\overline{Y}_{dt}^{(2)} &= \sum_{j\in U_{dt}}\Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}I_{\widetilde{\mathcal{H}}_{dt}}(j)-\frac{1}{N_{dt}}I_{r_{dt}}(j)\Big)\widetilde{e}_{\psi,dtj}(\theta_d)\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sigma_{\theta_d t}\sum_{j\in\widetilde{\mathcal{H}}_{dt}}\{\widetilde{R}_{dtj}-R_{dtj}\}+c_\phi\sum_{j\in s_{dt}-\widetilde{\mathcal{H}}_{dt}}\big(\mathrm{sgn}(\widetilde{e}_{\psi,dtj})-\mathrm{sgn}(e_{\psi,dtj})\big)\Big\}\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}-\sum_{j\in\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt}}e_{\psi,dtj}+\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt}}\widetilde{R}_{dtj}-\sum_{j\in\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt}}R_{dtj}\Big\}\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{c_\phi}{n_{dt}}\Big\{\sum_{j\in s_{dt}-(\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})-\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{sgn}(e_{\psi,dtj})\Big\}.
\end{aligned}$$

Assumption (E2) implies that

$$\begin{aligned}
\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}E[\mathrm{sgn}(\widetilde{e}_{\psi,dtj})] &= 2\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}-\widehat{\pi}_{-2,dtj}),\\
\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{var}(\mathrm{sgn}(\widetilde{e}_{\psi,dtj})) &= 4\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}+\widehat{\pi}_{-2,dtj})-4\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}-\widehat{\pi}_{-2,dtj})^2.
\end{aligned}$$

From assumption (C2), we obtain that

$$\frac{1}{n_{dt}}\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}E[\mathrm{sgn}(\widetilde{e}_{\psi,dtj})]=o(1),\quad \frac{1}{n_{dt}}\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widetilde{\mathcal{H}}_{dt})}\mathrm{var}(\mathrm{sgn}(\widetilde{e}_{\psi,dtj}))=o(1). \quad \text{(D.10)}$$

From (D.9), (D.10) and assumption (E2), the variance of $\overline{Y}_{dt}^{(2)}$ is

$$\begin{aligned} V_{dt}^{(2)} &= \operatorname{var}\big(\overline{Y}_{dt}^{(2)}\big) = \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{\widetilde{\mathcal{H}}_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\Big)^2 \operatorname{var}\big(\widetilde{e}_{\psi,dtj}(\theta_d)\big) \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\sigma_{\theta_d t}^2}{n_{dt}^2} \sum_{j\in\widetilde{\mathcal{H}}_{dt}} \operatorname{var}\big(\widetilde{R}_{dtj} - R_{dtj}\big) \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{c_\phi^2}{n_{dt}^2} \sum_{j\in s_{dt}-\widetilde{\mathcal{H}}_{dt}} \operatorname{var}\big(\operatorname{sgn}(\widetilde{e}_{\psi,dtj}) - \operatorname{sgn}(e_{\psi,dtj})\big) + o(n^{-1}) \\ &= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\widetilde{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big) \operatorname{var}\big(\widetilde{e}_{\psi,dtj}(\theta_d)\big) + o(n^{-1}), \end{aligned}$$

where $\operatorname{var}\big(\widetilde{e}_{\psi,dtj}(\theta_d)\big) = \boldsymbol{x}'_{dtj}\operatorname{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\theta_d,\boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj}$, $j=1,\ldots,N_{dt}$.

From assumption (A9), it holds that

$$\frac{1}{N_{dt}} \sum_{j\in U_{dt}} E\big[\widetilde{e}_{\psi,dtj}(\theta_d)\big] = \overline{\boldsymbol{x}}'_{dt} E\big[\boldsymbol{\beta}_\psi(\theta_d,\boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\theta_d,\boldsymbol{w}_t)\big] = O(n^{-1/2}), \tag{D.11}$$

where we have defined the population mean of the vector of explanatory variables as

$$\bar{\boldsymbol{x}}'_{dt} = \frac{1}{N_{dt}} \sum_{j\in U_{dt}} x'_{dtj}.$$

From (D.11), (D.9) and (D.10), the expected prediction difference is

$$\begin{aligned} E_{dt}^{(2)} = E\big[\overline{Y}_{dt}^{(2)}\big] &= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{s_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\Big) E\big[\widetilde{e}_{\psi,dtj}(\theta_d)\big] \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sigma_{\theta_d t} \sum_{j\in\widetilde{\mathcal{H}}_{dt}} E\big[\widetilde{R}_{dtj} - R_{dtj}\big] + c_\phi \sum_{j\in s_{dt}-\widetilde{\mathcal{H}}_{dt}} E\big[\operatorname{sgn}(\widetilde{e}_{\psi,dtj}) - \operatorname{sgn}(e_{\psi,dtj})\big]\Big\} = o(1). \end{aligned}$$

D.1.4 Part III: Dealing with the differences $\widehat{\overline{\boldsymbol{Y}}}_{\boldsymbol{dt}}^{\boldsymbol{btmq}} - \widetilde{\overline{\boldsymbol{Y}}}_{\boldsymbol{dt}}^{\boldsymbol{btmq}}$

In this section, we calculate the expected value and variance of the prediction differences

$$\widehat{\overline{Y}}_{dt}^{btmq} - \widetilde{\overline{Y}}_{dt}^{btmq}, \quad d=1,\ldots,D,\ t=1,\ldots,T.$$

The reasoning is based on the calculation of first-order Taylor approximations and the subsequent computation of expected values and variances.

As $\phi(0)=0$ and $\dot{\phi}(0)=1$, a Taylor series expansion of $\phi(\widehat{u}_{\psi,dtj})$ around $u=0$ yields to

$$\phi(\widehat{u}_{\psi,dtj}) = \phi(0) + \dot{\phi}(0)\widehat{u}_{\psi,dtj} + \widehat{R}_{\psi,dtj} = \widehat{u}_{\psi,dtj} + \widehat{R}_{\psi,dtj}, \quad j=1,\ldots,N_{dt}, \tag{D.12}$$

where $\widehat{R}_{dtj} = \frac{1}{2}\,\widehat{u}^{*2}_{\psi,dtj}$, $0<|\widehat{u}^*_{\psi,dtj}|<|\widehat{u}_{\psi,dtj}|$ and $\operatorname{var}(\widehat{R}_{dtj}) \leqslant \operatorname{var}(\widehat{u}_{\psi,dtj})$.

From assumption (D1), we have

$$\frac{1}{n_{dt}}\sum_{j\in\widehat{\mathcal{G}}_{dt}} E[\widehat{R}_{dtj}-\widetilde{R}_{dtj}]=o(1),\quad \frac{1}{n_{dt}}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\mathrm{var}\big((\widehat{R}_{dtj}-\widetilde{R}_{dtj})^2\big)=o(1). \tag{D.13}$$

From the Taylor series expansion (D.12), we have

$$\begin{aligned}\widehat{B}_{dt}^{btmq} &= \sum_{j\in\widehat{\mathcal{G}}_{dt}}\sigma_{\theta_d t}\phi\big(\widehat{u}_{\psi,dtj}\big)+c_\phi\sum_{j\in s_{dt}-\widehat{\mathcal{G}}_{dt}}\mathrm{sgn}\big(\widehat{e}_{\psi,dtj}\big),\\ &= \sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{e}_{\psi,dtj}+\sigma_{\theta_d t}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{R}_{dtj}+c_\phi\sum_{j\in s_{dt}-\widehat{\mathcal{G}}_{dt}}\mathrm{sgn}(\widehat{e}_{\psi,dtj}).\end{aligned}$$

As $\widehat{e}_{\psi,dtj}-\widetilde{e}_{\psi,dtj}=\boldsymbol{x}'_{dtj}\big(\boldsymbol{\beta}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t)-\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t)\big)=\widehat{e}_{\psi,dtj}(\theta_d)$, then $B_{dt}^{(3)}=\widehat{B}_{dt}^{btmq}-\widetilde{B}_{dt}^{btmq}$ is

$$\begin{aligned}B_{dt}^{(3)} &= \sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{e}_{\psi,dtj}+\sigma_{\theta_d t}\sum_{j\in\widehat{\mathcal{G}}_{dt}}\widehat{R}_{dtj}+c_\phi\sum_{j\in s_{dt}-\widehat{\mathcal{G}}_{dt}}\mathrm{sgn}(\widehat{e}_{\psi,dtj})\\ &- \sum_{j\in\widetilde{\mathcal{G}}_{dt}}\widetilde{e}_{\psi,dtj}+\sigma_{\theta_d t}\sum_{j\in\widetilde{\mathcal{G}}_{dt}}\widetilde{R}_{dtj}+c_\phi\sum_{j\in s_{dt}-\widetilde{\mathcal{G}}_{dt}}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\\ &= \sum_{j\in\widehat{\mathcal{H}}_{dt}}\widehat{e}_{\psi,dtj}(\theta_d)+\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{e}_{\psi,dtj}-\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widetilde{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}\\ &+ \sigma_{\theta_d t}\Big(\sum_{j\in\widehat{\mathcal{H}}_{dt}}(\widehat{R}_{dtj}-\widetilde{R}_{dtj})+\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{R}_{dtj}-\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widetilde{R}_{dtj}\Big)\\ &+ c_\phi\Big(\sum_{j\in s_{dt}-\widehat{\mathcal{H}}_{dt}}(\mathrm{sgn}(\widehat{e}_{\psi,dtj})-\mathrm{sgn}(\widetilde{e}_{\psi,dtj}))\\ &+ \sum_{j\in s_{dt}-(\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widehat{e}_{\psi,dtj})-\sum_{j\in s_{dt}-(\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\Big).\end{aligned}$$

The prediction difference $\overline{Y}_{dt}^{(3)}=\overline{\widehat{Y}}_{dt}^{btmq}-\overline{\widetilde{Y}}_{dt}^{btmq}$ is

$$\begin{aligned}\overline{Y}_{dt}^{(3)} &= \frac{1}{N_{dt}}\Big(\sum_{j\in s_{dt}}y_{dtj}+\sum_{j\in r_{dt}}\boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d,\boldsymbol{w}_t)\Big)+\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\widehat{B}_{dt}^{btmq}\\ &- \frac{1}{N_{dt}}\Big(\sum_{j\in s_{dt}}y_{dtj}+\sum_{j\in r_{dt}}\boldsymbol{x}'_{dtj}\widehat{\boldsymbol{\beta}}_\psi(\theta_d,\boldsymbol{w}_t)\Big)+\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\widetilde{B}_{dt}^{btmq}\\ &= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}B_{dt}^{(3)}-\frac{1}{N_{dt}}\sum_{j\in r_{dt}}\widehat{e}_{\psi,dtj}(\theta_d).\end{aligned}$$

By substituting $B_{dt}^{(3)}$, we obtain

$$\begin{aligned}
\overline{Y}_{dt}^{(3)} &= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Bigg\{\sum_{j\in\widetilde{\mathcal{H}}_{dt}}\widehat{e}_{\psi,dtj}(\theta_d)+\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{e}_{\psi,dtj}-\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}\\
&+ \sigma_{\theta dt}\Big(\sum_{j\in\widehat{\mathcal{H}}_{dt}}(\widehat{R}_{dtj}-\widetilde{R}_{dtj})+\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{R}_{dtj}-\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widetilde{R}_{dtj}\Big)\\
&+ c_\phi\Big(\sum_{j\in s_{dt}-\widehat{\mathcal{H}}_{dt}}(\mathrm{sgn}(\widehat{e}_{\psi,dtj})-\mathrm{sgn}(\widetilde{e}_{\psi,dtj}))+\sum_{j\in s_{dt}-(\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widehat{e}_{\psi,dtj})\\
&- \sum_{j\in s_{dt}-(\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\Big)\Bigg\}-\frac{1}{N_{dt}}\sum_{j\in r_{dt}}\widehat{e}_{\psi,dtj}(\theta_d).
\end{aligned}$$

We write $Y_{dt}^{(3)}$ in the form

$$\begin{aligned}
\overline{Y}_{dt}^{(3)} &= \sum_{j\in U_{dt}}\Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}I_{\widehat{\mathcal{H}}_{dt}}(j)-\frac{1}{N_{dt}}I_{r_{dt}}(j)\Big)\widehat{e}_{\psi,dtj}(\theta_d)\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sigma_{\theta dt}\sum_{j\in\widehat{\mathcal{H}}_{dt}}\{\widehat{R}_{dtj}-\widetilde{R}_{dtj}\}+c_\phi\sum_{j\in s_{dt}-\widehat{\mathcal{H}}_{dt}}\big(\mathrm{sgn}(\widehat{e}_{\psi,dtj})-\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\big)\Big\}\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{e}_{\psi,dtj}-\sum_{j\in\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widetilde{e}_{\psi,dtj}+\sum_{j\in\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt}}\widehat{R}_{dtj}-\sum_{j\in\mathcal{G}_{dt}-\widehat{\mathcal{H}}_{dt}}R_{dtj}\Big\}\\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{c_\phi\sum_{j\in s_{dt}-(\widehat{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widehat{e}_{\psi,dtj})-c_\phi\sum_{j\in s_{dt}-(\widetilde{\mathcal{G}}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{sgn}(\widetilde{e}_{\psi,dtj})\Big\}.
\end{aligned}$$

Assumption (E3) implies that

$$\begin{aligned}
\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widehat{\mathcal{H}}_{dt})}E[\mathrm{sgn}(\widehat{e}_{\psi,dtj})] &= 2\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}-\widehat{\pi}_{-2,dtj}),\\
\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{var}(\mathrm{sgn}(\widehat{e}_{\psi,dtj})) &= 4\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}+\widehat{\pi}_{-2,dtj})-4\sum_{j\in s_{dt}-\mathcal{G}_{dt}}(\widehat{\pi}_{2,dtj}-\widehat{\pi}_{-2,dtj})^2.
\end{aligned}$$

From assumption (D2), we obtain that

$$\frac{1}{n_{dt}}\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widehat{\mathcal{H}}_{dt})}E[\mathrm{sgn}(\widehat{e}_{\psi,dtj})]=o(1),\quad \frac{1}{n_{dt}}\sum_{j\in s_{dt}-(\mathcal{G}_{dt}-\widehat{\mathcal{H}}_{dt})}\mathrm{var}(\mathrm{sgn}(\widehat{e}_{\psi,dtj}))=o(1). \quad \text{(D.14)}$$

From (D.13), (D.14) and assumption (E3), the variance of $\overline{Y}_{dt}^{(3)}$ is

$$\begin{aligned} V_{dt}^{(3)} &= \operatorname{var}\big(\overline{Y}_{dt}^{(3)}\big) = \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{\widehat{\mathcal{H}}_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\Big)^2 \operatorname{var}\big(\widehat{e}_{\psi,dtj}(\theta_d)\big) \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2}\sigma_{\theta_d t}^2 \sum_{j\in\widehat{\mathcal{H}}_{dt}} \operatorname{var}\big(\widehat{R}_{dtj} - \widetilde{R}_{dtj}\big) \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} c_\phi^2 \sum_{j\in s_{dt}-\widehat{\mathcal{H}}_{dt}} \operatorname{var}\big(\operatorname{sgn}(\widehat{e}_{\psi,dtj}) - \operatorname{sgn}(\widetilde{e}_{\psi,dtj})\big) + o(n^{-1}) \\ &= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\widehat{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big) \operatorname{var}\big(\widehat{e}_{\psi,dtj}(\theta_d)\big) + o(n^{-1}), \end{aligned}$$

where $\operatorname{var}\big(\widehat{e}_{\psi,dtj}(\theta_d)\big) = \boldsymbol{x}_{dtj}' \operatorname{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj}$, $j = 1, \ldots, N_{dt}$.

From assumption (A9), it holds that

$$\frac{1}{N_{dt}} \sum_{j\in U_{dt}} E\big[\widehat{e}_{\psi,dtj}(\theta_d)\big] = \overline{\boldsymbol{x}}_{dt}' E\big[\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) - \widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big] = O(n^{-1/2}). \tag{D.15}$$

From (D.15), (D.13) and (D.14), the expected prediction difference is

$$\begin{aligned} E_{dt}^{(3)} = E\big[\overline{Y}_{dt}^{(3)}\big] &= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}} I_{s_{dt}}(j) - \frac{1}{N_{dt}} I_{r_{dt}}(j)\Big) E\big[\widehat{e}_{\psi,dtj}(\theta_d)\big] \\ &+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)\frac{1}{n_{dt}}\Big\{\sigma_{\theta_d t} \sum_{j\in\widehat{\mathcal{H}}_{dt}} E\big[\widehat{R}_{dtj} - \widetilde{R}_{dtj}\big] + c_\phi \sum_{j\in s_{dt}-\widehat{\mathcal{H}}_{dt}} E\big[\operatorname{sgn}(\widehat{e}_{\psi,dtj}) - \operatorname{sgn}(\widetilde{e}_{\psi,dtj})\big]\Big\} = o(1). \end{aligned}$$

D.1.5 Final expression of the mean squared error

Under the necessary assumptions set out in Section D.1.1, this section collects the analytical results from Sections D.1.2, D.1.3 and D.1.4 to finally derive a first-order approximation of the MSE of the BTMQ predictor $\widehat{\overline{Y}}_{dt}^{btmq}$.

First of all, it holds that

$$MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) = E\big[\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big)^2\big] = \operatorname{var}\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big) + \big(E\big[\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big]\big)^2.$$

Based on the decomposition

$$\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt} = \big(\widehat{\overline{Y}}_{dt}^{btmq} - \widetilde{\overline{Y}}_{dt}^{btmq}\big) + \big(\widetilde{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}^{btmq}\big) + \big(\overline{Y}_{dt}^{btmq} - \overline{Y}_{dt}\big) = \overline{Y}_{dt}^{(3)} + \overline{Y}_{dt}^{(2)} + \overline{Y}_{dt}^{(1)},$$

we write

$$\begin{aligned} \operatorname{var}\big(\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big) &= V_{dt}^{(1)} + V_{dt}^{(2)} + V_{dt}^{(3)} + 2\operatorname{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(2)}\big) + 2\operatorname{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(1)}\big) + 2\operatorname{cov}\big(\overline{Y}_{dt}^{(2)}, \overline{Y}_{dt}^{(1)}\big), \\ E\big[\widehat{\overline{Y}}_{dt}^{btmq} - \overline{Y}_{dt}\big] &= E_{dt}^{(1)} + E_{dt}^{(2)} + E_{dt}^{(3)} = E_{dt}^{(1)} + o(1). \end{aligned}$$

The covariances are $\operatorname{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(2)}\big) = E\big[\overline{Y}_{dt}^{(3)}\overline{Y}_{dt}^{(2)}\big] + o(1)$, $\operatorname{cov}\big(\overline{Y}_{dt}^{(3)}, \overline{Y}_{dt}^{(1)}\big) = E\big[\overline{Y}_{dt}^{(3)}\overline{Y}_{dt}^{(1)}\big] + o(1)$ and $\operatorname{cov}\big(\overline{Y}_{dt}^{(2)}, \overline{Y}_{dt}^{(1)}\big) = E\big[\overline{Y}_{dt}^{(2)}\overline{Y}_{dt}^{(1)}\big] + o(1)$. Under regularity assumptions, the expectations of the previous cross-products should be $o(1)$.

Therefore, an approximation of $MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big)$ is

$$MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) = V_{dt}^{(1)} + V_{dt}^{(2)} + V_{dt}^{(3)} + E_{dt}^{(1)2} + o(1).$$

The following theorem (Bugallo et al., 2024e) summarizes the final MSE approximation.

Theorem 1. Under assumptions (Φ1), (N1)-(N2), (Q1), (A1)-(A9), (B1)-(B4), (C1)-(C2), (D1)-(D2), (E1)-(E3) in Appendix D.1.1, a first-order approximation of $MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big)$ is

$$\begin{aligned}
MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) &= \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\mathcal{G}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\Big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\Big)^2 \xi_{dt}^2 \\
&+ \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\tilde{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\boldsymbol{x}'_{dtj}\operatorname{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\theta_d, \boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj} \\
&+ \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\widehat{\mathcal{H}}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\boldsymbol{x}'_{dtj}\operatorname{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big)\boldsymbol{x}_{dtj} \\
&+ \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \Big(\frac{c_\phi}{n_{dt}} \sum_{j\in s_{dt}-\mathcal{G}_{dt}} E\big[\operatorname{sgn}(e_{\psi,dtj})\big] + \frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j\in\mathcal{G}_{dt}} E[R_{dtj}]\Big)^2 + o(1).
\end{aligned}$$

D.1.6 Estimation of the final expression of the mean squared error

In this section we propose an estimator of the MSE of the BTMQ predictor $\widehat{\overline{Y}}_{dt}^{btmq}$ for the first-order approximation formulated in Theorem 1. We use the simplified notation

$$MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big) = S_1 + S_2 + S_3 + S_4 + o(1).$$

The first summand of $MSE\big(\widehat{\overline{Y}}_{dt}^{btmq}\big)$ is

$$\begin{aligned}
S_1 &= S_{11} + S_{12} = \sum_{j\in U_{dt}} \Big(\Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} I_{\mathcal{G}_{dt}}(j) + \frac{1}{N_{dt}^2} I_{r_{dt}}(j)\Big)\Big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\Big)^2 \xi_{dt}^2 \\
&= \Big(1-\frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n_{dt}^2} \sum_{j\in\mathcal{G}_{dt}} \Big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\Big)^2 \xi_{dt}^2 + \frac{1}{N_{dt}^2} \sum_{j\in r_{dt}} \Big(\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t)\Big)^2 \xi_{dt}^2.
\end{aligned}$$

Provided $q_{dtj} - \theta_d \neq 0$, $j = 1, \ldots, N_{dt}$, we obtain from the definition of the TWMQ linear models (5.14) and the Taylor series expansion (D.1) that

$$\begin{aligned}
\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t) &= (q_{dtj} - \theta_d)^{-1}(\boldsymbol{\beta}_\psi(q_{dtj}, \boldsymbol{w}_t) - \boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) - \boldsymbol{r}_{dtj}), \\
\boldsymbol{x}'_{dtj}\boldsymbol{\kappa}_\psi(\theta_d, \boldsymbol{w}_t) &= (q_{dtj} - \theta_d)^{-1}(y_{dtj} - \boldsymbol{x}'_{dtj}\boldsymbol{\beta}_\psi(\theta_d, \boldsymbol{w}_t) - \boldsymbol{x}'_{dtj}\boldsymbol{r}_{dtj}) \\
&= (q_{dtj} - \theta_d)^{-1}(e_{\psi,dtj} - \boldsymbol{x}'_{dtj}\boldsymbol{r}_{dtj}).
\end{aligned}$$

If $\boldsymbol{x}'_{dtj}\boldsymbol{r}_{dtj} \approx 0$, the first term of S_1 is estimated by

$$\widehat{S}_{11,0} = \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\widehat{\xi}^2_{dt}}{n^2_{dt}} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \frac{1}{(\widehat{q}_{dtj} - \widehat{\theta}_d)^2} \widehat{e}^2_{\psi,dtj},$$

where $\widehat{\xi}^2_{dt}$ is an estimator of $\mathrm{var}(q_{dtj})$, i.e.

$$\widehat{\xi}^2_{dt} = \widehat{\mathrm{var}}(q_{dtj}) = \frac{1}{n_{dt}-1} \sum_{j \in s_{dt}} (\widehat{q}_{dtj} - \widehat{\bar{q}}_{dt.})^2, \quad \widehat{\bar{q}}_{dt.} = \frac{1}{n_{dt}} \sum_{j \in s_{dt}} \widehat{q}_{dtj}.$$

The second term of S_1 is estimated as

$$\widehat{S}_{12,0} = \frac{N_{dt} - n_{dt}}{n_{dt}} \frac{\widehat{\xi}^2_{dt}}{N^2_{dt}} \sum_{j \in s_{dt}} \frac{1}{(\widehat{q}_{dtj} - \widehat{\theta}_d)^2} \widehat{e}^2_{\psi,dtj}.$$

Therefore, we estimate $V^{(1)}_{dt}$ by $\widehat{S}_{1,0} = \widehat{S}_{11,0} + \widehat{S}_{12,0}$. Nevertheless, the differences $\widehat{q}_{dtj} - \widehat{\theta}_d$ are expected to be close to zero, leading to instability problems due to multiplication by $(\widehat{q}_{dtj} - \widehat{\theta}_d)^{-2}$, $j = 1, \ldots, n_{dt}$, both in the estimation of $S_{11,0}$ and $S_{12,0}$. For this reason, we do not recommend using them. Potential solutions include proposing the following estimators

$$\begin{aligned}
\widehat{S}_{11} &= \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\widehat{\xi}^2_{dt}}{n^2_{dt}} \Big(\frac{1}{\mathrm{card}(\widehat{\mathcal{G}}_{dt})} \sum_{j \in \widehat{\mathcal{G}}_{dt}} (\widehat{q}_{dtj} - \widehat{\theta}_d)^2 \Big)^{-1} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \widehat{e}^2_{\psi,dtj}, \\
\widehat{S}_{12} &= \frac{N_{dt} - n_{dt}}{n_{dt}} \frac{\widehat{\xi}^2_{dt}}{N^2_{dt}} \Big(\frac{1}{n_{dt}} \sum_{j \in s_{dt}} (\widehat{q}_{dtj} - \widehat{\theta}_d)^2 \Big)^{-1} \sum_{j \in s_{dt}} \widehat{e}^2_{\psi,dtj},
\end{aligned}$$

such that we finally write $\widehat{S}_1 = \widehat{S}_{11} + \widehat{S}_{12}$. Clearly, this new estimation of S_1 largely avoids the problems of numerical instability, but it may also bias the final estimation. All things considered, we strongly recommend this second approach, which provides more stable results.

On the other hand, we use estimators of S_2 and S_3 to estimate the variance terms $V^{(2)}_{dt}$ and $V^{(3)}_{dt}$, respectively. An estimator for both S_2 and S_3 is obtained as follows

$$\begin{aligned}
\widehat{V}^{(2)}_{dt} &= \widehat{V}^{(3)}_{dt} = \sum_{j \in U_{dt}} \Big(\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n^2_{dt}} I_{\widehat{\mathcal{G}}_{dt}}(j) + \frac{1}{N^2_{dt}} I_{r_{dt}}(j) \Big) \boldsymbol{x}'_{dtj} \widehat{V}_\beta \boldsymbol{x}_{dtj} \\
&= \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n^2_{dt}} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \boldsymbol{x}'_{dtj} \widehat{V}_\beta \boldsymbol{x}_{dtj} + \frac{1}{N^2_{dt}} \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \widehat{V}_\beta \boldsymbol{x}_{dtj},
\end{aligned}$$

where $\widehat{V}_\beta$ is an estimator of $V_\beta = \mathrm{var}\big(\widehat{\boldsymbol{\beta}}_\psi(\widehat{\theta}_d, \boldsymbol{w}_t)\big)$, as the one given in (5.23).

The last element S_4 of $MSE\big(\widehat{\overline{Y}}^{btmq}_{dt}\big)$ corresponds to the bias term $E^{(1)2}_{dt}$. First, we propose

$$\sum_{j \in s_{dt} - \mathcal{G}_{dt}} \widehat{E}\big[\mathrm{sgn}(e_{\psi,dtj})\big] = \sum_{j \in s_{dt} - \widehat{\mathcal{G}}_{dt}} \mathrm{sgn}(\widehat{e}_{\psi,dtj}).$$

From the Taylor expansion (D.4), $R_{dtj} = \frac{1}{2}\, u^{*2}_{\psi,dtj}$, $0 < |u^{*}_{\psi,dtj}| < |u_{\psi,dtj}|$, $j = 1, \ldots, N_{dt}$, so

$$\sum_{j \in \mathcal{G}_{dt}} \widehat{E}[R_{dtj}] \;=\; \frac{1}{2\sigma^2_{\theta_{dt}}} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \widehat{e}^2_{\psi,dtj}.$$

We derive the following estimator of $MSE\big(\widehat{\overline{Y}}^{btmq}_{dt}\big)$:

$$\begin{aligned}
mse^{btmq}_{3,dt} \;=\;& \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\widehat{\xi}^2_{dt}}{n^2_{dt}} \Big(\frac{1}{\mathrm{card}(\widehat{\mathcal{G}}_{dt})} \sum_{j \in \widehat{\mathcal{G}}_{dt}} (\widehat{q}_{dtj} - \widehat{\theta}_d)^2\Big)^{-1} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \widehat{e}^2_{\psi,dtj} + \\
&+ \frac{N_{dt} - n_{dt}}{n_{dt}} \frac{\widehat{\xi}^2_{dt}}{N^2_{dt}} \Big(\frac{1}{n_{dt}} \sum_{j \in s_{dt}} (\widehat{q}_{dtj} - \widehat{\theta}_d)^2\Big)^{-1} \sum_{j \in s_{dt}} \widehat{e}^2_{\psi,dtj} \\
&+ 2\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n^2_{dt}} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \boldsymbol{x}'_{dtj} \widehat{V}_\beta \boldsymbol{x}_{dtj} + \frac{1}{N^2_{dt}} \sum_{j \in r_{dt}} \boldsymbol{x}'_{dtj} \widehat{V}_\beta \boldsymbol{x}_{dtj} \\
&+ \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{1}{n^2_{dt}} \Big(c_\phi \sum_{j \in s_{dt} - \widehat{\mathcal{G}}_{dt}} \mathrm{sgn}(\widehat{e}_{\psi,dtj}) + \frac{1}{2\sigma_{\theta_{dt}}} \sum_{j \in \widehat{\mathcal{G}}_{dt}} \widehat{e}^2_{\psi,dtj}\Big)^2. \qquad \text{(D.16)}
\end{aligned}$$

Finally, an estimator of $RMSE\big(\widehat{\overline{Y}}^{btmq}_{dt}\big) = \big(MSE\big(\widehat{\overline{Y}}^{btmq}_{dt}\big)\big)^{1/2}$ is $rmse^{btmq}_{3,dt} = \big(mse^{btmq}_{3,dt}\big)^{1/2}$.

The first-order approximation of the MSE given in Theorem 1 leads to an MSE estimator in (D.16), but the study of its theoretical properties, such as the first-order bias, is left for future research. To the best of our knowledge, the latter has never been done for a predictor derived from an MQ model. We believe that this is an issue that should be addressed for the simplest models, i.e. those proposed by Chambers and Tzavidis (2006).

D.2 Selection of area-time specific robustness parameters

In this section we provide the proof of Theorem 2, which was formulated in Section 5.4.

Theorem 2. Let ϕ be the Huber function, defined in (5.3). For $mse^{btmq}_{dt} \in \{mse^{btmq}_{1,dt}, mse^{btmq}_{2,dt}\}$, it exists an unique solution $\widehat{c}_{\phi,dt}$ of the minimization problem

$$\widehat{c}_{\phi,dt} = \underset{c_\phi \geqslant 0}{\mathrm{argmin}}, mse^{btmq}_{dt}(c_\phi), \quad d = 1, \ldots, D,\ t = 1, \ldots, T,$$

belonging to the interval $\big[0, \max_{j \in s_{dt}} |\widehat{u}_{\psi,dtj}|\big]$ and its explicit expression is calculable.

Theorem 2 provides a local, area-time specific approach, that calculates the robustness parameter that best bounds the outlier observations in each subdomain to reduce the predictive bias, without detriment to the MSE. Therefore, it allows us to intuit the relationship between the sign of the bias and the number of large positive and negative residuals. Finally, and quite incidentally, selecting the value of $\widehat{c}_{\phi,dt}$ not only avoids subjective choices, but also reveals the atypical condition of a subdomain (Bugallo et al., 2024e).

Proof. First, we calculate the c_ϕ-dependent part of $mse_{dt}^{btmq} \in \{mse_{1,dt}^{btmq}, mse_{2,dt}^{btmq}\}$. To do so, we define $A_{dt}(c_\phi)$ and write it as a function of c_ϕ, i.e.

$$A_{dt}(c_\phi) = \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)^2 \sum_{j \in s_{dt}} \phi^2\big(\widehat{u}_{\psi,dtj}\big) + \Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \phi\big(\widehat{u}_{\psi,dtj}\big)\Big)^2.$$

Technical definitions required for the proof are given below. Let $\mathcal{U} = \{|\widehat{u}_{\psi,dtj}|_{(1)}, \ldots, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}\}$ be the ordered version of the set $U = \{|\widehat{u}_{\psi,dt1}|, \ldots, |\widehat{u}_{\psi,dtn_{dt}}|\}$, so that $|\widehat{u}_{\psi,dtj}|_{(1)} = \min\limits_{j \in s_{dt}} |\widehat{u}_{\psi,dtj}|$ and $|\widehat{u}_{\psi,dtj}|_{(n_{dt})} = \max\limits_{j \in s_{dt}} |\widehat{u}_{\psi,dtj}|$. We define $|\widehat{u}_{\psi,dtj}|_{(0)} = 0$ and

$$\Lambda_{dt,\ell} = \big\{j \in \{1, \ldots, n_{dt}\} : |\widehat{u}_{\psi,dtj}| \leqslant |\widehat{u}_{\psi,dtj}|_{(\ell+1)}\big\},$$

where $\text{card}(\Lambda_{dt,\ell}) = n_{dt} - n_{dt,\ell}$, $n_{dt,\ell} = n_{dt,\ell}^+ + n_{dt,\ell}^- \in \mathbb{N}$, $\ell = 0, \ldots, n_{dt} - 1$, and

$$\begin{aligned} n_{dt,\ell}^+ &= \text{card}\big\{|\widehat{u}_{\psi,dtj}| \in \mathcal{U} : |\widehat{u}_{\psi,dtj}| > |\widehat{u}_{\psi,dtj}|_{(\ell+1)},\ \widehat{u}_{\psi,dtj} > 0\big\} \in \mathbb{N}, \\ n_{dt,\ell}^- &= \text{card}\big\{|\widehat{u}_{\psi,dtj}| \in \mathcal{U} : |\widehat{u}_{\psi,dtj}| > |\widehat{u}_{\psi,dtj}|_{(\ell+1)},\ \widehat{u}_{\psi,dtj} < 0\big\} \in \mathbb{N}. \end{aligned}$$

The c_ϕ-dependent terms of $A_{dt}(c_\phi)$ are continuous in $c_\phi \in [0, \infty)$, piecewise quadratic in $c_\phi \in I_\ell = [|\widehat{u}_{\psi,dtj}|_{(\ell)}, |\widehat{u}_{\psi,dtj}|_{(\ell+1)})$, $\ell = 0, \ldots, n_{dt} - 1$, and constant in $c_\phi \in I_{n_{dt}} = [|\widehat{u}_{\psi,dtj}|_{(n_{dt})}, \infty)$. Therefore, we have

$$\begin{aligned} A_{dt}(c_\phi) &= \Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\Big(c_\phi(n_{dt,\ell}^+ - n_{dt,\ell}^-) + \sum_{j \in \Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}\Big)\Big)^2 \\ &+ \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\sigma_{\theta_d t}^2}{n_{dt}^2}\Big(c_\phi^2 n_{dt,\ell} + \sum_{j \in \Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}^2\Big), \quad c_\phi \in I_\ell,\ \ell = 0, \ldots, n_{dt} - 1, \\ A_{dt}(c_\phi) &= \Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}} \sum_{j \in s_{dt}} \widehat{u}_{\psi,dtj}\Big)^2 + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2 \frac{\sigma_{\theta_d t}^2}{n_{dt}^2} \sum_{j \in s_{dt}} \widehat{u}_{\psi,dtj}^2, \quad c_\phi \geqslant |\widehat{u}_{\psi,dtj}|_{(n_{dt})}. \end{aligned}$$

Let us now prove the existence and uniqueness of solution and calculate its explicit expression.

Existence. Since $A_{dt}(c_\phi)$ is a constant function in $I_{n_{dt}}$, the search for extreme values is reduced to the compact (closed and bounded) interval $\big[0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}\big]$. Further, $A_{dt}(c_\phi)$ is a piecewise function in $\big[0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}\big]$ but, as an inherited property of function $\phi(u)$, $A_{dt}(c_\phi)$ is continuous in $[0, \infty)$. By virtue of the Weierstrass Theorem, $A_{dt}(c_\phi)$ reaches its absolute maximum and minimum values in $\big[0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}\big]$. If the minimum is reached at $c_\phi = 0$, no bias correction is needed. If the minimum is reached at $c_\phi = |\widehat{u}_{\psi,dtj}|_{(n_{dt})}$, is also reached in $[|\widehat{u}_{\psi,dtj}|_{(n_{dt})}, \infty)$, and the bias correction is maximal.

Uniqueness. By definition, $A_{dt}(c_\phi)$ is an infinitely differentiable function in $c_\phi \in [0, \infty) - \mathcal{U}$. The elements of the set $\mathcal{U}$ are the avoidable discontinuities of $A_{dt}(c_\phi)$. The first and second

order derivatives of $A_{dt}(c_\phi)$ in $c_\phi \in I_\ell - \{|\widehat{u}_{\psi,dtj}|_{(\ell)}\}$, $\ell = 0, \ldots, n_{dt} - 1$, are

$$\begin{aligned}
\frac{\partial A_{dt}(c_\phi)}{\partial c_\phi} &= 2\Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\Big(c_\phi(n^+_{dt,\ell} - n^-_{dt,\ell}) + \sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}\Big)\Big) \\
&\cdot \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}(n^+_{dt,\ell} - n^-_{dt,\ell}) + 2c_\phi\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2\Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)^2 n_{dt,\ell}, \\
\frac{\partial^2 A_{dt}(c_\phi)}{\partial^2 c_\phi} &= 2\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2\Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)^2(n^+_{dt,\ell} - n^-_{dt,\ell})^2 + 2\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2\Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)^2 n_{dt,\ell} \\
&= 2\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)^2\Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)^2\Big((n^+_{dt,\ell} - n^-_{dt,\ell})^2 + n_{dt,\ell}\Big).
\end{aligned}$$

For $c_\phi \in I_\ell - \{|\widehat{u}_{\psi,dtj}|_{(\ell)}\}$, $\ell = 0, \ldots, n_{dt} - 1$, it follows that $\frac{\partial^2 A_{dt}(c_\phi)}{\partial^2 c_\phi} > 0$ if and only if $n_{dt,\ell} > 0$, so $A_{dt}(c_\phi)$ is strictly convex in $\bigcup_{l=0}^{n_{dt}}(I_\ell - \{|\widehat{u}_{\psi,dtj}|_{(\ell)}\}) = [0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}] - \mathcal{U}$. Given that a strictly convex continuous function in an open set is strictly convex at its closure, $A_{dt}(c_\phi)$ is strictly convex in the compact interval $[0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}]$. Therefore, the uniqueness of the global minimum of $A_{dt}(c_\phi)$ in $[0, |\widehat{u}_{\psi,dtj}|_{(n_{dt})}]$ is guaranteed.

Explicit expression of $\widehat{c}_{\phi,dt}$. We have already proved the existence and uniqueness of solutions, so it is now appropriate to have an explicit expression. We distinguish two cases. Case 1 is the extreme solution $\widehat{c}_{\phi,dt} = |\widehat{u}_{\psi,dtj}|_{(n_{dt})}$. Case 2 is a solution $\widehat{c}_{\phi,dt} \in I_\ell$ for some $\ell \in \{0, \ldots, n_{dt} - 1\}$, so either $\widehat{c}_{\phi,dt} = |\widehat{u}_{\psi,dtj}|_{(\ell)}$, or $\widehat{c}_{\phi,dt}$ fulfills that $\frac{\partial A_{dt}(c_\phi)}{\partial c_\phi}\Big|_{c_\phi = \widehat{c}_{\phi,dt}} = 0$, i.e.

$$\begin{aligned}
&\Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\Big(\widehat{c}_{\phi,dt}(n^+_{dt,\ell} - n^-_{dt,\ell}) + \sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}\Big)\Big)(n^+_{dt,\ell} - n^-_{dt,\ell}) \\
&+ \widehat{c}_{\phi,dt}\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\Big(\frac{\sigma_{\theta_d t}}{n_{dt}}\Big)n_{dt,\ell} = 0 \Longleftrightarrow \Big(\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}\Big)(n^+_{dt,\ell} - n^-_{dt,\ell}) \\
&+ \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\Big((n^+_{dt,\ell} - n^-_{dt,\ell})^2 + n_{dt,\ell}\Big)\widehat{c}_{\phi,dt} = 0.
\end{aligned}$$

Solving for $\widehat{c}_{\phi,dt}$ from this equation, we obtain

$$\widehat{c}_{\phi,dt} = \frac{\Big(\widehat{B}_{dt} + \big(1 - \frac{n_{dt}}{N_{dt}}\big)\frac{\sigma_{\theta_d t}}{n_{dt}}\sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj}\Big)(n^-_{dt,\ell} - n^+_{dt,\ell})}{\Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\Big((n^+_{dt,\ell} - n^-_{dt,\ell})^2 + n_{dt,\ell}\Big)}. \tag{D.17}$$

Since $\widehat{c}_{\phi,dt} \in (|\widehat{u}_{\psi,dtj}|_{(\ell)}, |\widehat{u}_{\psi,dtj}|_{(\ell+1)})$, $\widehat{c}_{\phi,dt} > 0$ and the numerator of (D.17) is strictly positive. Further, there are two possibilities: (i) $n^+_{dt,\ell} < n^-_{dt,\ell}$ and $\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj} > 0$; (ii) $n^+_{dt,\ell} > n^-_{dt,\ell}$ and $\widehat{B}_{dt} + \Big(1 - \frac{n_{dt}}{N_{dt}}\Big)\frac{\sigma_{\theta_d t}}{n_{dt}}\sum_{j\in\Lambda_{dt,\ell}} \widehat{u}_{\psi,dtj} < 0$. Therefore, if the number of negative outliers, $n^-_{dt,\ell}$, is greater than the number of positive outliers, $n^+_{dt,\ell}$, then the bias is positive. If not, the other way round. □

Bibliography

Alfo, M., Salvati, N., and Ranalli, M. G. (2017). Finite mixtures of quantile and M-quantile regression models. *Statistics and Computing*, 27(2):547–570.

Anggreyani, A., Indahwati, I., and Kurnia, A. (2015). Small area estimation for estimating the number of infant mortality using a mixed effects zero inflated Poisson model. *Indonesian Journal of Statistics*, 20(2):108–115.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). Multivariate Fay-Herriot bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society, Series A*, 180(4):1191–1209.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83:28–36.

Beaumont, J. F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100(3):555–569.

Benavent, R. and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94:372–390.

Benavent, R. and Morales, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods and Applications*, 30(1):195–222.

Berg, E. (2022). Empirical best prediction of small area means based on a unit-level Gamma-Poisson model. *Journal of Survey Statistics and Methodology*, 11(4):873–894.

Berg, E. and Fuller, W. (2012). Estimators of error covariance matrices for small area prediction. *Computational Statistics and Data Analysis*, 56(10):2949–2962.

Berntsen, J., Espelid, T., and Genz, A. (1991). An adaptive algorithm for the approximate calculation of multiple integrals. *Transactions on Mathematical Software*, 17(4):437–451.

Bianchi, A., Fabrizi, E., Salvati, N., and Tzavidis, N. (2018). Estimation and testing in M-quantile regression with applications to small area estimation. *International Statistical Review*, 86(3):541–570.

Bianchi, A. and Salvati, N. (2015). Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators. *Communications in Statistics*, 44:813–827.

Boubeta, M., Lombardía, M. J., Marey-Pérez, M., and Morales, D. (2019). Poisson mixed models for predicting number of fires. *International Journal of Wildland Fire*, 28(3).

Boubeta, M., Lombardía, M. J., and Morales, D. (2016a). Empirical best prediction under area-level Poisson mixed models. *TEST*, 25:548–569.

Boubeta, M., Lombardía, M. J., and Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics and Data Analysis*, 107:32–47.

Boubeta, M., Lombardía, M. J., and Morales, D. (2023). Small area prediction of proportions and counts under a spatial Poisson mixed model. *Statistical Methods and Applications*, 1:1–23.

Boubeta, M., Lombardía, M. J., Morales, D., González-Manteiga, W., and Marey-Pérez, M. (2016b). Burned area prediction with semiparametric models. *International Journal of Wildland Fire*, 25(6):669–678.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). `glmmTMB`: Balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Brundson, C., Fotheringham, A., and M., C. (1996). Geographically Weighted Regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28:281–298.

Bugallo, M., Esteban, M. D., Hobza, T., Morales, D., and Pérez, A. (2024a). Small area estimation of labour force indicators under unit-level multinomial mixed models. *Journal of the Royal Statistical Society, Series A*.

Bugallo, M., Esteban, M. D., and Morales, D. (2024b). Small area estimation of the proportion of single-person households: Application to the Spanish Household Budget Survey. *SORT-Statistics and Operations Research Transactions*, 48(1):125–152.

Bugallo, M., Esteban, M. D., Morales, D., and Marey-Pérez, M. (2023). Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain. *Journal of Environmental Management*, 328:116788.

Bugallo, M., Esteban, M. D., Morales, D., and Marey-Pérez, M. (2024c). Pattern recognition and modelling of virulent wildfires in Spain. *International Journal of Wildland Fire.* [Submitted].

Bugallo, M., Esteban, M. D., Pagliarella, M. C., and Morales, D. (2024d). Model-based estimation of small area dissimilarity indexes: An application to sex occupational segregation in Spain. *Social Indicators Research*, 174:473–501.

Bugallo, M., Morales, D., Salvati, N., and Schirripa, F. (2024e). Temporal M-quantile models and robust bias-corrected small area predictors. *Journal of Survey Statistics and Methodology.* [Submitted].

Burgard, J. P., Esteban, M. D., Morales, D., and Pérez, A. (2021). Small area estimation under a measurement error bivariate Fay-Herriot model. *Statistical Methods and Applications*, 30(1):79–108.

Burgard, J. P., Krause, P., and Morales, D. (2022). A measurement error Rao-Yu model for regional prevalence estimation over time using uncertain data obtained from dependent survey estimates. *TEST*, 31(1):204–234.

Cabello, E., Morales, D., and Pérez, A. (2024). Area-level model-based small area estimation of divergence indexes in the Spanish Labour Force Survey. *Journal of Survey Statistics and*

Methodology.

Cai, S. and Rao, J. (2022). Selection of auxiliary variables for three-fold linking models in small area estimation: A simple and effective method. *Statistics*, 5(1):128–138.

Cengiz, M. and Terzim, Y. (2012). Comparing models of the effect of air pollutants on hospital admissions and symptoms for chronic obstructive pulmonary disease. *Central European Journal of Public Health*, 20(4):282–6.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069.

Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2014a). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 76(1):47–69.

Chambers, R., Chandra, H., and Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, 37(2):153–170.

Chambers, R., Dreassi, E., and N., S. (2014b). Disease mapping via negative binomial regression M-quantiles. *Statistics in Medicine*, 33(27):4805–4824.

Chambers, R., Salvati, N., and Tzavidis, N. (2012). M-quantile regression for binary data with application to small area estimation. *Biometrika*.

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*, 179(2):453–479.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2):255–268.

Chandra, H. and Chambers, R. (2011). Small area estimation for skewed data in presence of zeros. *Calcutta Statistical Association Bulletin*, 63:249–252.

Chandra, H., Salvati, N., and Chambers, R. (2017). Small area prediction of counts under a non- stationary spatial model. *Spatial Statistics*, 20:30–56.

Chandra, H. and Sud, U. (2012). Small area estimation for zero inflated data. *Communications in Statistics-Simulation and Computation*, 41:632–643.

Chernick, M. R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons.

Cho, Y. and Shim, K. (2019). Differences between one-person and multi-person households on socioeconomic status, health behavior, and metabolic syndrome across gender and age groups. *Korean Journal of Family Practice*, 9:373–382.

Cohen, P. (2021). The rise of one-person households. *Socius*, 7.

Das, S. and Kotikula, A. (2019). Gender-based employment segregation: Understanding causes and policy interventions. Jobs Working Paper. *The World Bank Group*, 26.

Datta, G., Lahiri, P., and Maiti, T. (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102:83–97.

Datta, G., Lahiri, P., Maiti, T., and Lu, K. (1999). Hierarchical bayes estimation of unemployment rates for the US states. *Journal of the American Statistical Association*, 94:1074–1082.

Datta, G. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110:1735–1744.

Dawber, J. and Chambers, R. (2019). Modelling group heterogeneity for small area estimation using M-quantiles. *International Statistical Review*, 87(1):50–63.

Dawber, J., Salvati, N., Fabrizi, E., and Tzavidis, N. (2022). Expectile regression for multi-category outcomes with application to small area estimation of labour force participation. *Journal of the Royal Statistical Society, Series A*, 185(2):590–619.

Demidenko, E. (2013). *Mixed Models, Theory and Applications*. Wiley.

Díz-Rosales, N., Lombardía, M. J., and Morales, D. (2023). Poverty mapping under area-level random regression coefficient Poisson models. *Journal of Survey Statistics and Methodology*, 12(2):404–434.

Dongmo-Jiongo, V., Haziza, D., and Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100:843–858.

Duncan, O. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217.

Eklund, J., Jones, J. P., and Rasanen, M. e. a. (2022). Elevated fires during Covid-19 lockdown and the vulnerability of protected areas. *Nature Sustainability*, 5:603–609.

Erciulescu, A. L. and Fuller, W. A. (2013). Small area prediction of the mean of a Binomial random variable. *Proceedings of the Survey Research Methods*.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., and Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*, 29(3):793–818.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., and Pérez, A. (2022a). Empirical best prediction of small area bivariate parameters. *Scandinavian Journal of Statistics*, 49:1699–1727.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., and Pérez, A. (2022b). Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *Journal of Applied Statistics*, 49(1):141–168.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., and Pérez, A. (2023). Small area estimation of average compositions under multivariate nested error regression models. *TEST*, 32:651–676.

Esteban, M. D., Morales, D., Pérez, A., and Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56:2840–2855.

Faltys, O., Hobza, T., and Morales, D. (2022). Small area estimation under area-level generalized linear mixed models. *Communications in Statistics*, 51(12):7404–7426.

Fay, R. and Herriot, J. R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74:269–277.

Feng, C. and Dean, C. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics*, 23(6):493–508.

Feng, C., Li, L., and Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20(175).

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs interpretability of clas-

sifications. *Biometrics*, 21:768–769.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 200:675–701.

Genz, A. and Malik, A. (1980). An adaptive algorithm for numeric integration over an n- dimensional rectangular region. *Journal of Computational and Applied Mathematics*, 6(4):295–302.

Ghosh, M., Kim, D., Sinha, K., Maiti, T., Katzoff, M., and Parsons, V. (2009). Hierarchical and empirical bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. *Survey Methodology*, 35:53–66.

Ghosh, M., Nangia, N., and Kim, D. (1996). Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association*, 91:1423–1431.

Gobierno de España (2019a). Diagnóstico Estrategica Nacional frente al Reto Demográfico. Eje Despoblación. *Ministerio de Politica Territorial y Función Pública 2019*.

Gobierno de España (2019b). Estrategia Nacional frente al Reto Demográfico. Directrices Generales. Comisionado para el Reto Demográfico. *Ministerio de Política Territorial y Función Pública 2019*.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2002). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, 51:2720–2733.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, 52:42–52.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, 10(2):215–239.

Greitemeyer, T. (2009). Stereotypes of singles: Are singles what we think? *European Journal of Social Psychology*, 39(3):368–383.

Guadarrama, M., Morales, D., and Molina, I. (2021). Time stable empirical best predictors under a unit-level model. *Computational Statistics and Data Analysis*, 160:107226.

Hagenaars, A., de Vos, K., and Zaidi, M. (1994). *Poverty Statistics in the Late 1980s: Research Based on Micro-data*. Office for Official Publications of the European Communities. Luxembourg.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling"', Part I. *The Foundations of Survey Sampling*, 236.

Hall, D. B. (2000). Zero-inflated Poisson and Binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small-area prediction. *Journal of the Royal Statistical Society, Series B*, 68:221–238.

Hariyanto, S., Notodiputro, K., Kurnia, A., and Sadik, K. (2018). Measurement error in small area estimation: A literature review. *IOP Conference Series: Earth and Environmental Science*, 187, 012034.

Hartigan, J. and Wong, M. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society, Series C*, 28(1):100–108.

Hartono, B., Kurnia, A., and Indahwati, I. (2017). Zero inflated binomial models in small area estimation with application to unemployment data in indonesia. *International Journal of Computer Science and Network*, 6(6):746–752.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Hepburn, E. (2016). *Cohesion policy and regional mobilisation*. Handbook on Cohesion Policy in the EU. Edward Elgar Publishing.

Herrador, M., Esteban, M. D., Hobza, T., and Morales, D. (2011). A modified nested-error regression model for small area estimation. *Statistics*, 47(2):258–273.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32(3):661–669.

Hobza, T., Morales, D., and Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, 27:270–294.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively re-weighted least-squares. *Communications in Statistics*, 6(9):813–827.

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

Huang, E. and Bell, W. (2004). An empirical study on using ACS supplementary survey data in SAIPE state poverty models. pages 3677–3684.

Huber, P. (1981). *Robust Statistics*. John Wiley & Sons.

Humi, M. (2017). *Introduction to Mathematical Modelling*. Chapman & Hall.

Janicki, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to esmation of poverty rates. *Communications in Statistics*, 49(9):2264–2284.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, 15:1–96.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296–1310.

Krause, J., Burgard, J. P., and Morales, D. (2022a). L2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-defficiency. *Metrika*, 85:459–489.

Krause, J., Burgard, J. P., and Morales, D. (2022b). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika*, 87(1):344–368.

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R., Ren, W., VanDeKerckhove, W., Li, L., and Rao, J. (2020). Program for the international assessment of adult compe- tencies (PIAAC): State and county estimation methodology report. *Institute of Education Sciences,*

National Center for Education Statistics: Washington.

Krieg, S., Boonstra, H. J., and Smeets, M. (2016). Small-area estimation with zero-inflated data: A simulation study. *Journal of Official Statistics*, 32(4):963–986.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B. (2016). `TMB`: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70:1–21.

Lee, A., Zhao, Y., Yau, K., and L., X. (2010). How to analyze longitudinal multilevel physical activity data with many zeros? *Preventive Medicine*, 51(6):476–481.

Lee, S. and Lee, S. (2021). Comparative analysis of health behaviors, health status, and medical needs among one-person and multi-person household groups: Focused on the ageing population of 60 or more. *Korean Journal of Family Medicine*, 42(1):73–83.

Ling, W., Cheng, B., Wei, Y., Willey, J., and Cheung, Y. (2022). Statistical inference in quantile regression for zero-inflated outcomes. *Statistica Sinica*, 32(3):1411–1433.

Lombardía, M. J., López-Vizcaíno, E., and Rueda, C. (2021). A new approach to the gender pay gap decomposition by economic activity. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(1):219–245.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, 13(2):153–178.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Association, Series A*, 178(3):535–565.

Maiti, T., Sinha, S., and Zhong, P. (2016). Functional mixed effects model for small area estimation. *Scandinavian Journal of Statistics*, 43(3):886–903.

Marchetti, S., Beresewicz, M., Salvati, N., Szymkowiak, M., and Wawrowski, L. (2018). The use of a three-level M-quantile model to map poverty at local administrative unit 1 in Poland. *Journal of the Royal Statistical Society, Series A*, 181(4):1077–1104.

Marcis, L., Morales, D., Pagliarella, M. C., and Salvatore, R. (2023). Three-fold Fay-Herriot model for small area estimation and its diagnostics. *Statistical Methods and Applications*, 23:1563–1609.

Marcos, R., Turco, M., Bedia, J., Llasat, M. C., and Provenzale, A. (2015). Seasonal predictability of summer fires in a mediterranean environment. *International Journal of Wildland Fire*, 24(8):1076–1084.

Marhuenda, Y., Molina, I., and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58:308–325.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J. (2017). Poverty mapping in small areas under a two-fold nested error regression model. *Journal of the Royal Statistical Society, Series A*, 180(4):1111–1136.

Marhuenda, Y., Morales, D., and Pardo, M. C. (2014). Information criteria for Fay-Herriot model selection. *Computational Statistics and Data Analysis*, 70:268–280.

Marhuenda, Y., Morales, D., and Pardo, M. C. (2016). Tests for the variance parameter in the Fay-Herriot model. *Statistics*, 50(1):27–42.

Marino, M., Ranalli, M., Salvati, N., and Alfo, M. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *Annals of Applied Statistics*,

13 (2):1166–1197.

Massó, M., Golias, M., and Nogueira, J. (2021). Brecha Salarial de Género en las Universidades Públicas Espanholas: Informe Final. *Universidad de la Corunha*.

Michael, F. and Thomas, D. (2016). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Chapman & Hall.

MITECO (2023). *Common Basic Duidelines for Sustainable Forest Management*. Spanish Ministry for the Ecological Transition and the Demographic Challenge.

Moanga, D., Biging, G., Radke, J., and Butsic, V. (2020). The space-time cube as an approach to quantifying future wildfires in California. *International Journal of Wildland Fire*, 30:311–327.

Molina, I., Saei, A., and Lombardía, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society, Series A*, 170(4):975–1000.

Morales, D., Esteban, M. D., Pérez, A., and Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*. Springer Nature.

Morales, D., Pagliarella, M. C., and Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT-Statistics and Operations Research Transactions*, 39(1):19–34.

Morales, D. and Santamaría, L. (2019). Small area estimation under unit-level temporal linear mixed models. *Journal of Statistical Computation and Simulation*, 89(9):1592–1620.

Newey, W. and Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847.

NUTS (2016). *Nomenclature of Territorial Units for Statistics*. European Commission.

Park, B., Kwon, H., Ha, M., and Burm, E. (2016). A comparative study on mental health between elderly living alone and elderly couples: Focus on gender and demographic characteristics. *Journal of Korean Public Health Nursing*, 20:195–20.

Parker, P. A., Janicki, R., and Holan, S. H. (2023a). Comparison of unit-level small area estimation modeling approaches for survey data under informative sampling. *Journal of Survey Statistics and Methodology*, 11(4):858–872.

Parker, P. A., Janicki, R., and Holan, S. H. (2023b). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, 11(4):829–857.

Pazos-Vidal, S. (2022). "Emptied Spain" and the limits of domestic and EU territorial mobilisation. *Revista Galega de Economía*, 31(2):1–28.

Pereira, M., Caramelo, L., Vega-Orozco, C., Costa, R., and Tonini, M. (2015). Space-time clustering analysis performance of an aggregated dataset: The case of wildfires in Portugal. *Environmental Modelling and Software*, 72:239–249.

Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16:217–237.

Pfeffermann, D., Terryn, B., and Moura, F. (2008). Small area estimation under a two- part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, 34(2):235–249.

Pinheiro, J. and Bates, D. (2023). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-164.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Pinilla, V. and Saez, L. A. (2017). Rural depopulation in Spain: Genesis of a problem and innovative policies. *Centre for Studies on Depopulation and Development of Rural Areas*.

Porter, A., Wikle, C., and Holan, S. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Australian and New Zealand Journal of Statistics*, 57(1):15–29.

Powell, M. (2004). The NEWUOA software for unconstrained optimization without derivatives. *Proceedings of the 40th Workshop on Large Scale Nonlinear Optimization*.

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons.

Pratesi, M., Ranalli, M. G., and Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit hucs of the Northeastern US. *Environmetrics*, 19(7):687–701.

R Development Core Team (2024). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ranalli, M., Montanari, G., and Vicarelli, C. (2018). Estimation of small area counts with the benchmarking property. *METRON*, 76 (3):349–378.

Ranjbar, S., Ronchetti, E., and Sperlich, S. (2023). *Chapter "Bias Calibration for Robust Estimation in Small Areas" in Robust and Multivariate Statistical Methods*. Springer.

Rao, J. (2003). *Small Area Estimation*. John Wiley & Sons.

Rao, J. and Molina, I. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Rao, J. and Yu, M. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics*, 22:511–528.

Reluga, K., Sperlich, S., and Lombardía, M. J. (2021). Simultaneous inference for empirical best predictors with a poverty study in small areas. *Journal of the American Statistical Association*, 118(541):583–595.

Reluga, K., Sperlich, S., and Lombardía, M. J. (2023). Simultaneous inference for linear mixed model parameters with an application to small area estimation. *International Statistical Review*, 91:193–217.

Ríos-Pena, L., Cadarso-Suárez, C., Kneib, T., and Marey-Pérez, M. (2015). Applying binary structured additive regression for predicting wildfire in Galicia, Spain. *Procedia Environmental Sciences*, 27:123–126.

Ríos-Pena, L., Kneib, T., Cadarso-Suárez, C., Klein, N., and Marey-Pérez, M. (2018). Studying the occurrence and burnt area of wildfires using zero-one-inflated structured additive beta regression. *Environmental Modelling and Software*, 110:107–118.

Ríos-Pena, L., Kneib, T., Cadarso-Suárez, C., and Marey-Pérez, M. (2017). Predicting the occurrence of wildfires with binary structured additive regression models. *Journal of Environmental Management*, 187:154–165.

Rodrígues, M., De la Riva, J., and Fotheringham, S. (2014). Modeling the spatial variation of

the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography*, 48:52–63.

Russo, A., Gouveia, C. M., Pascoa, P., DaCamara, C. C., Sousa, P. M., and Trigo, R. M. (2017). Assessing the role of drought events on wildfires in the Iberian Peninsula. *Agricultural and Forest Meteorology*, 237:50–59.

Saei, A. and Taylor, A. (2012). Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics*, 66:187–201.

Salardi, P. (2016). The evolution of gender and racial occupational segregation across formal and non-formal labor markets in Brazil, 1987 to 2006. *Review of Income and Wealth*, 62(S1):68–89.

Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. (2012). Small area estimation via M-quantile Geographically Weighted Regression. *TEST*, 21:1–28.

Santi, V. M., Kurnia, A., and Sadik, K. (2019). Modelling of the number of malarias suffers in indonesia using bayesian generalized linear models. *Journal of Physics: Conference Series*.

Sarndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Schirripa Spagnolo, F., Mauro, V., and Salvati, N. (2021). Generalised M-quantile random-effects model for discrete response: An application to the number of visits to physicians. *Biometrical Journal*, 63(4):859–874.

Singh, B., Shukla, G., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, 31:183–195.

Singh, M. P., Gambino, J., and Mantel, H. J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20(1):3–22.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37(3):381–399.

Snell, K. (2017). The rise of living alone and loneliness in history. *Social History*, 42(1):2–28.

Street, J., Carrol, R., and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively re-weighted least squares. *The American Statistician*, 42:152–154.

Sugasawa, S., Kubokawa, T., and Ogasawara, K. (2017). Empirical uncertain bayes methods in area-level models. *Scandinavian Journal of Statistics*, 44(3):684–706.

Tan, Z., Carrasco, L., and Taylor, D. (2021). Corrigendum to: Spatial correlates of forest and land fires in Indonesia. *International Journal of Wildland Fire*, 30(9):732–732.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 200(6):99–114.

Tzavidis, N., Marchetti, S., and Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, 52:167–186.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., and Chambers, R. (2014). Robust small area prediction for counts. *Statistical Methods in Medical Research*, 24(3).

Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17(3):393–411.

United Nations (2015). "The 17 Sustainable Development Goals". *Dept of Economic and Social Affairs*.

Viedma, O., Urbieta, I., and Moreno, J. (2018). Wildfires and the role of their drivers are changing over time in a large rural area of west-central Spain. *Scientific Reports*, 8(1):1–13.

Vinciotti, V. and Keming, Y. (2009). M-quantile regression analysis of temporal gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–20.

Welsh, A. (1986). Bahadur representations or robust scale estimators based on regression residuals. *Annals of Statistics*, 14:1246–1251.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer.

Wu, C. (1983). On the convergence properties of the em algorithm. *Annals of Statistics*, 11(1):95–103.

Yi, M. and Nordhausen, K. (2023). *Robust and Multivariate Statistical Methods*. Springer.

You, Y. and Rao, J. (2000). Hierarchical bayes estimation of small area means using multi-level models. *Survey Methodology*, 26:173–181.

Zewotir, T. and Galpin, J. (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. *TEST*, 16(1):58–75.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.