PhD-FHSE-2024-022
The Faculty of Humanities, Education and Social Science
DISSERTATION

Defence held on 25/11/2023 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN SCIENCES SOCIALES
DOCTEUR DE UNIVERSITY OF INSUBRIA
EN SCIENCES ECONOMIQUES

by

Carlotta MONTORSI
Born on 23 January 1996 in Turin (Italy)

EMPIRICAL ESSAYS ON WELL-BEING

Dissertation defence committee

Dr. Alessio FUSCO, Dissertation supervisor
Senior Research Scientist, Luxembourg Institue of Socio-Economic Research (LISER)

Prof. Chiara GIGLIARANO, Dissertation supervisor
Professor, Liuc University

Prof. Philippe VAN KERM, Chairman
Professor, University of Luxembourg

Prof. Daniela SONEDDA, Vice-chair
Associate Professor, University of Insubria

Prof. Stéphane BORDAS, Expert in advisory capacity
Professor, University of Luxembourg

Prof. Climent QUINTANA-DOMEQUE, member
Professor, University of Exeter

Prof. Stéphane MUSSARD, member
Professor, University of Nîmes

Acknowledgements

I embarked on this PhD journey amid a global pandemic, followed by a war that shocked economies and geopolitical stability and another war that showed the fragility of international agreements. Meanwhile, the world underwent a Fourth Industrial Revolution that reshaped the labor markets, led by artificial intelligence (AI) technologies and powerful computing capabilities. As to say, it was not an ordinary walk, let alone a predictable one. Against difficulties and changes, I wouldn't be where I am without the support and guidance of many great people accompanying me along the way. I would need more space to thank you as much as I would like, but here is a modest attempt.

First and foremost, I would like to express my gratitude to my PhD advisors, Alessio Fusco and Chiara Gigliarano. Despite the 700 kilometers separating them, they could perfectly complement each other in many aspects. Alessio, thank you for keeping the paramount objective of my research in focus, pushing me to do better with your "good exercise", and always being there to solve my doubts. Chiara, thank you for the encouragement, opportunities, and trust you have shown me since we first met when I was a freshly graduated 23-year-old student; very much uncertain about my future career and capabilities. Without your support, I would probably not have pursued a PhD, and I will forever be grateful for that. Their combined guidance has been fundamental throughout these years, as I am sure it will be for the ones that will follow.

I would also like to thank the other members of my thesis supervisory committee, Philippe Van Kerm and Stéphane Bordas. Philippe's insightful feedback, to-the-point questions, and "holistic" vision have been fundamental in better framing my research questions and orienting the overall narrative of this thesis. Stéphane's sparkling and positive personality, paired with the opportunities to interact with the Computer Science department, has been highly beneficial for developing machine learning skills. I am also grateful to Daniela Sonedda, Climent Quintana-Domeque, and Stéphane Mussard for kindly agreeing to be part of my PhD Jury.

I am also indebted to my sponsor, Dario Sansone, for the opportunity to visit the Economics Department at the University of Exeter during my PhD. Dario and the great professors and PhD students who welcomed me with open arms made those four months a window of enthusiasm, improvement, and expansion of my research

interests—an experience I would repeatedly redo.

I am grateful to the Fonds National de la Recherche (FNR) that funded my PhD, which resulted in this dissertation, Grant – 12252781-DTU-DRIVEN, to the University of Luxembourg and Insubria, the Luxembourg Institute for Socio-Economic Research, and the Living Conditions department, for giving me the perfect environment to conduct my research. Here's to the friendship and camaraderie with my peers, friends, and co-authors, Jules, Lucas, Martin, Etienne, Felix, Julio, Juliette, and Ivana: growing up with you along this journey has been invaluable to me and an extraordinary memory for the rest of my life. Your shared experiences, advice, and laughter have made this period more enjoyable and smoothed out the various obstacles. My hope goes to our future, where I see ourselves discussing and toasting at conferences and events, celebrating our future successes, and supporting us for any failures. This is not a farewell but a goodbye.

To Alberto, if I am thanking you now, it means not only that we are actually at the end of a journey that we pushed each other to start, aware of the distance it would create between us. But it also means that these four years have not divided us, and we are now closer than ever. Thank you because you have always been a source of inspiration for me.

Il ringraziamento finale per sempre sarà per la mia famiglia: mamma e papà, le mie sorelle, Marghe e Cate, mio fratello Pietro. Siete stati e sarete per sempre le persone a cui chiedero il primo consiglio, di vita e di carriera. Il vostro amore incondizionato e il vostro incoraggiamento perenne hanno attraversato come onde sonore la distanza di questi quattro anni, facendovi sentire vicini anche se lontani. Siete la mia rappresentazione di amore. Un ringraziamento speciale anche alle mie due nonne, Carla e Rita, e alla mia zia Paola, altre donne della mia vita che, per diverse aspetti, hanno plasmato la donna che sono oggi.

Contents

Abstract

This doctoral dissertation addresses a spectrum of research topics, unified by the general objective of unfolding factors that shape well-being at the national and individual levels. The ambition is that findings from this study might improve policy decision-making and, therefore, boost societal and individual well-being. However, it is crucial to perceive each chapter as a standalone article. As such, they delve into unique research questions, each requiring distinct data and methodological approaches.

Chapter 1 – Predicting Depression in Old Age: Combining Life Course Data with Machine Learning. *Published in Economics & Human Biology*. With ageing populations, understanding life course factors that raise the risk of depression in old age may help anticipate needs and reduce healthcare costs in the long run. In this Chapter, we estimate the risk of depression in old age by combining adult life biographies and childhood conditions in supervised machine learning algorithms. Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), we implement and compare the performance of six alternative machine learning algorithms. We analyse the performance of the algorithms using different life-course data configurations. While we obtain similar predictive abilities between algorithms, we achieve the highest predictive performance when employing semi-structured representations of life courses using sequence data. We use the Shapley Additive Explanations method to extract the most decisive predictive patterns. Age, health, childhood conditions, and low education predict most depression risk later in life. Still, we identify new predictive patterns in indicators of life course instability and low utilization of dental care services.

Chapter 2 – The Old Folks at Home: Parental Retirement and Adult Children Well-Being. *Submitted*. In this Chapter, we appeal to changes in the UK State Pension eligibility age to establish the causal effect of parental retirement on adult children's well-being. In a Fuzzy Regression Discontinuity Design analysis, maternal retirement increases adult children's life and income satisfaction by 0.20 standard deviations in the short run. In Differences-in-Difference regressions, fathers' delayed retirement increases adult sons' life and income satisfaction by 0.14 and 0.12 standard deviations. These impacts are stronger for adult children with lower incomes, with young dependents of their own, and who live close to their retired parents. We emphasize the critical role of intergenerational time transfers from retired mothers in enhancing their adult children's well-being.

Chapter 3 – Small Pictures, Big Biases: The Adverse Effect of an Airbnb Anti-discrimination Policy. Using scraped data from the Airbnb platform in New York City alongside state-of-the-art Vision Transformers models for image classification, this Chapter investigates the magnitude of ethnic disparities in the Airbnb platform and the impact of a policy to address them. First, we show that Black hosts have a 7.2 percentage points lower occupancy rate than their White counterparts despite no differences in pricing. For Asian and Hispanic hosts, the difference from Whites is small and mostly insignificant for both occupancy rate and prices. Second, using difference-in-differences and event studies approaches, we show that the Airbnb anti-discrimination policy, which reduced the size of users' profile pictures on the platform, unexpectedly increased the Black-White disparity by about 4 percentage points. As a reaction to the adverse impact of the new policy, Black hosts start offering more basic amenities for their listings. We argue that a potential mechanism for the increase in Black-White disparity stems from the increasing guests' uncertainty in discerning facial features that positively correlate with occupancy rates from the smaller profile pictures. As a result, guests focus more on skin color.

Chapter 4 – Spatial Comprehensive Well-Being Composite Indicators based on Bayesian Latent Factor Model: Evidence from Italian provinces. *Published in Social Indicators Research*. This Chapter proposes spatial comprehen-

sive composite indicators to evaluate the well-being levels and ranking of Italian provinces with data from the Equitable and Sustainable Well-Being dashboard. We use a method based on Bayesian latent factor models, which allow us to include spatial dependence across the units of analysis, quantify uncertainty in the resulting estimates, and estimate data-driven weights for elementary indicators. The results reveal that our data-driven approach changes the resulting composite indicator rankings compared to traditional composite indicator approaches. Estimated social and economic well-being is unequally distributed among southern and northern Italian provinces. In contrast, the environmental dimension appears less spatially clustered, and its composite indicators also reach above-average levels in the southern provinces. The time series of well-being composite indicators of Italian macro-areas shows clustering and macro-areas discrimination on larger territorial units.

Chapter 5 – Quality of Government for Environmental Well-Being? Subnational Evidence from European Regions. *Submitted*. This Chapter investigates the relationship between quality of government and environmental well-being in European regions at the NUTS-2 level. First, we quantified a significant spatial correlation in subnational environmental data. Therefore, we construct a set of composite indicators of environmental well-being through Bayesian spatial factor analysis. Finally, we use these composite indicators in spatial regression analysis and find that institutional quality is a key determinant of environmental well-being. We also find heterogeneity in the institutions-environment nexus across dimensions of environmental wellbeing —institutions matter especially for air and soil quality. Policymakers should be aware that environmental destruction can be tackled by building more effective regional institutions.

Co-author Statement

This dissertation includes both first-author papers and collaborative works. I provide a detailed account of my specific contributions to each paper for transparency and clarity.

- **Chapter 1 – Predicting Depression in Old Age: Combining Life Course data with Machine Learning**, co-authored with Alessio Fusco (LISER), Philippe Van Kerm (University of Luxembourg), and Stéphane Bordas (University of Luxembourg).

  I am the paper's first author. I designed the machine learning methodology and conducted the formal analysis. Alessio, Philippe, and I conceptualized the paper and wrote and reviewed the original draft. Stéphane secured the funding for the project.
- **Chapter 2 – The Old Folks at Home: Parental Retirement and Adult Children well-being**, co-authored with Andrew Clark (Paris School of Economics).

  I am the paper's first author. I secured access to the UK data and conducted the formal analysis. Andrew and I conceptualized the paper and wrote the draft.
- **Chapter 3 – Small pictures, Big Biases: The Adverse Effects of an Airbnb Anti-Discrimination policy**, co-authored with Julio Garbers (PhD student at LISER).

  This paper results from a strict collaboration between Julio and me. We share the tasks involved equally, including presentation and dissemination.
- **Chapter 4 – Spatial Comprehensive Well-Being Composite Indicators based on Bayesian latent factor model: evidence from Italian provinces**, co-authored with Chiara Gigliarano (Luic University).

Under the supervision of Principal Investigator Chiara Gigliarano, I handled the data cleaning, developed the methodology, and conducted the analysis. Chiara and I conceptualized the research question and wrote the draft.

- **Chapter 5 – Quality of government for environmental Well-Being? Subnational evidence form European Region**, co-authored with Andrea Vaccaro (first author - Oxford University) and Chiara Gigliarano.
  I am the paper's second author. I developed the methodology for the Environmental composite indicator and conducted the spatial analysis. Andrea collected the dashboard of environmental and quality of government indicators at the European level and conducted the literature review. The three authors refined the research question and wrote the final draft.

General Introduction

What This Dissertation Is About

Well-being is a multi-dimensional and expanding concept. As a consequence, well-being research is an interdisciplinary field where economics analysis pairs with psychology, moral philosophy, and, more recently, computer science and advanced statistical methods. Findings from this blend of disciplines have marked many academic milestones. Happiness does not correlate with income after a certain high threshold (Easterlin, 1973), and it tends to return to a stable set-point over time (Brickman et al., 1978). Unemployment modified this set point in the long run, affecting life satisfaction for those who remain employed (Lucas et al., 2004), and reducing air pollution improves children's health (Simeonova et al., 2021). More recently, advanced machine learning models could predict early mortality and personality traits (Savcisens et al., 2024). Yet, the field is not saturated, and researchers are still discussing what well-being entails, how to measure it, predict it, and what contributes to it. This ongoing debate underscores the need for continuing research on well-being to inform policy and improve quality of life.

This dissertation contributes to the current debate on well-being by focusing on three cross-cutting research areas: predictive analytics and machine learning in well-being research, inter-generational spillovers on well-being, and the impact of policies and institutional quality on well-being. Within these macro areas, each chapter addresses current policy priorities: mental health, population aging, ethnic discrimination, going beyond GDP measure, and environmental degradation.

In this introduction, I zoom into each of these three focal areas and highlight the distinctive contributions of this thesis. First, I explore how advanced computational methods can guide optimal resource allocation and advance the knowledge

of well-being determinants. Second, I present the relevance of spillover and intergenerational studies. Then, I discuss the crucial role of effective governance and policy design in shaping well-being outcomes. Finally, I present the outline of this dissertation.

Predictive Models in Well-Being Research

Recent advancements in predictive modeling, improved computer capabilities, and higher accessibility have boosted the expansion of machine-learning applications in all spheres, including economics. Here, machine learning tools have proven to accurately predict socio-economic outcomes, such as pupil school dropouts (Sansone, 2019) and economic developments (Ahn et al., 2023). In well-being research, there is growing yet sometimes contrasting evidence of machine learning models' abilities to predict subjective well-being outcomes, such as quality of life (Jannani et al., 2021), life satisfaction (Oparina et al., 2022), or mental health (Garriga et al., 2022). But first, one natural question arises: What is the need to obtain machine learning predictions in well-being research?

In any field of economics, the choice between traditional causal approaches and machine learning always hinges on a critical trade-off: prioritizing identifying the cause of an event or obtaining precise out-of-sample predictions of the likelihood of such events occurring. Causal approaches are essential for understanding the underlying factors that drive well-being outcomes, enabling policymakers to design interventions that precisely target specific causes. In contrast, machine learning excels in processing large, complex datasets and finding hidden patterns to accurately predict outcomes, which is invaluable for identifying targets and optimizing resource allocation (Kleinberg et al., 2015).

In the economics of well-being, machine learning models can predict which populations are most at risk of ill-being, allowing resources to be directed more efficiently toward preventive measures. Together, these approaches complement each other: causal analysis provides the ”why,” while machine learning offers the ”who,” creating a robust framework for informed policymaking and effective resource allocation.

Compared to other subjective well-being outcomes, mental health is among the most challenging but most relevant nowadays to target (WHO, 2021), especially in

old age. Poor mental health might indeed turn into severe chronic diseases, such as depression, anxiety, dementia, and many more, which are all hard to diagnose and even harder to prevent and treat. In old age, the problem of detecting who is at risk of mental illnesses is magnified by the stigma associated with aging, which leads to confounding symptoms of mental disorders with symptoms of aging. If policymakers aim to guarantee happier and healthier lives for all its citizens, including the elderly, targeting mental health is a critical starting point.

The complexity of predicting depression lies in the highly dimensional arrays of potential triggers and their interactions, both observable, such as parental divorce (Cherlin et al., 1998) and financial distress (Guan et al., 2022), and unobservable, like genetics and phenotypical traits. These underlying complexities motivate using machine learning algorithms to predict the risk of depression outbreaks and, if successful, complement human practitioners in their diagnosis. Still, a question remains open: Can depression be predicted from socioeconomic information?

Chapter 1 of this dissertation answers this question by applying machine learning techniques to predict the risk of clinical depression in old age from socioeconomic life course information. It proves that socio-economic biographical information paired with a machine learning algorithm can help identify vulnerable individuals. It also reveals new insights into the life course factors influencing mental health. These findings highlight the potential of predictive models in identifying who is at risk of potential ill-being and revealing neglected determinants of mental well-being.

Inter-Generational and Spillover Effects

Government interventions often have far-reaching consequences that extend beyond their immediate targets, influencing not only the direct recipients but also their families and broader communities (Angelucci & Di Maro, 2010). These effects, known as spillover effects, are critical for understanding the full impact of public policies. By examining these effects, researchers can uncover how policies aimed at one target population can ripple through to affect the well-being of future generations and other groups within society.

Positive and negative spillover effects of public interventions have been found across several fields. For example, deworming programs in Kenya significantly improved the educational outcomes of untreated pupils (Miguel & Kremer, 2004).

Older siblings' educational choices causally impacted the younger' (Altmejd et al., 2021). Retirement choices spillover across spouses (García-Miralles & Leganza, 2024). Most existing studies focus on spillover among siblings or peers, spouses or coworkers. Less attention has been given to inter-generational spillover effects, i.e., how public policies targeting one generation influence the youngest. Among them, only a few look at the spillover effect of old age public intervention (see, e.g., Ilciukas, 2023), and none look at subjective well-being outcomes.

This oversight is partly due to population aging becoming a prominent phenomenon in most wealthy countries only in recent decades.

Population aging, paired with decreasing fertility rates, is challenging the financial sustainability of the pension system in many OECD countries. As such, many governments in the OECD countries have gradually modified some of the elements of their pension systems, and postponing the statutory retirement age is among the most common intervention (OECD, 2023).

To fully gauge the potential impacts of such direct intervention, a crucial question is whether retirement is good or bad for retirees' well-being. Therefore, what would the impact of postponing retirement be? At the same time, policymakers must consider potential intergenerational spillover effects and how retirement affects the well-being of other family members, including their adult children in the workforce.

Chapter 2 provides answers to these questions by adopting a causal approach. It investigates the direct causal effect of parental retirement on retirees and their adult children's well-being (aged 25-45). It reveals significant retirement intergenerational spillover effects that vary by socioeconomic status and proximity to parents. These findings underscore the importance of considering the broader consequences of retirement interventions, as they often extend beyond the intended beneficiaries and can contribute to widening social inequalities.

Impact of Policies and Institutional Quality on Well-being

Well-designed policies and well-functioning institutions are essential for fostering environments where well-being can thrive. Conversely, poor institutional quality and ineffective policies can exacerbate inequalities, reduce trust in public systems, and ultimately undermine societal well-being. This section delves into these three facets and explains the specific criticality this thesis addresses.

Anti-Discrimination policies Discrimination, defined as treating someone differently based on traits like gender, ethnicity, or age, remains a crucial barrier to reaching equality in market outcomes. It limits opportunities, damages self-esteem (Jackson et al., 2019), and undermines societal efficiency by wasting talent and fostering segregation (Feagin and McKinney, 2005).

Designing effective policies to curb the socioeconomic harms of discrimination is challenging. Since Gary Becker's seminal work in 1957 (Becker, 1957), various economic models have attempted to explain the persistence and evolution of discrimination, revealing that its causes and forms are diverse and context-dependent. This diversity requires tailored anti-discrimination intervention; otherwise, such policies risk backfiring and worsening the issues they aim to address (see, e.g., Agan and Starr, 2018; Doleac and Hansen, 2020).

The rise of digital platforms like Airbnb has brought about new challenges in combating discrimination. While these platforms have implemented design changes to reduce discrimination, the core feature of sharing personal information—such as profile pictures and names—can inadvertently facilitate discriminatory behaviors.

Chapter 3 of this dissertation quantifies ethnic disparities in Airbnb market outcomes and assesses the effectiveness of an anti-discrimination policy introduced in October 2018. First, it describes a striking disparity in occupancy rates between minority hosts, especially Black hosts, and whites, which is not explained by apartment location or other observable characteristics. Second, the anti-discrimination policy evaluation surprisingly reveals backfiring effects that significantly increased the Black-White gap the 6-month after its implementation.

Composite indicators Well-being is multidimensional; the different yet related outcomes presented so far reflect this idea. Analyzing single proxies of well-being has advantages for effective policy-making, such as narrowing potential policy recommendations and facilitating analytical approaches, as most standard econometrics models assume a one-dimensional outcome. However, this uni-dimensional approach naturally lacks a holistic perspective on well-being evolving dimensions. As such, focusing on a single outcome might lead to underestimating or overestimating the actual well-being response and bias the resource allocation process in policymaking.

One growing well-being research area focuses on developing composite indicators (CIs) to address this issue. Composite Indicators are numerical measures that

simultaneously synthesize multiple dimensions. A notable example is the Human Development Index, introduced in 1990, which combines life expectancy, education, and income into a single measure, setting a precedent for development studies.

The 2009 report by the Commission on the Measurement of Economic Performance and Social Progress, led by Stiglitz, Sen, and Fitoussi, underscored the limitations of GDP as a measure of social progress, advocating for new indicators that "Going beyond GDP" to assess quality of life (Stiglitz et al., 2009). This report catalyzed the creation of new well-being frameworks, such as the Better Life Index.

Despite these advances, well-known composite indicators face limitations, including the arbitrary selection of indicators and weights, a lack of uncertainty measures, and failure to account for spatial spillovers. Addressing these issues is essential for improving the precision, transparency, and relevance of well-being measures for optimal resource allocation. Chapter 4 of this dissertation applies a new Bayesian statistical methodology to overcome these challenges and create more comprehensive and reliable well-being composite indicators that better inform policymakers.

Quality of Governance Governance quality refers to effectiveness, transparency, and accountability, including the legal framework, efficiency, and the rule of law, and is strictly related to institutions' quality (Kaufmann et al., 2009). Good governance boosts effective policies that protect individual rights, ensure equitable resource distribution, and maintain social stability, all foundational to promoting well-being. On the other hand, lousy governance often struggles with corruption, inefficiency, and poor service delivery, which can lead to adverse outcomes for well-being (Acemoglu et al., 2005).

Environmental degradation is among the key challenges many governments worldwide are attempting to solve. Here, striking heterogeneity appears within and across countries. A key question is the role of government quality in explaining these disparities and how to provide a comprehensive measure of environmental quality.

Chapter 5 explores the role of institutional quality in environmental well-being. It quantifies a robust positive correlation between effective governance and better environmental outcomes at the European subnational level. Despite the risk of reverse causality, these findings suggest that improving the quality of governance

might be a first step toward enhancing environmental well-being.

Dissertation outline

This dissertation explores three cross-cutting dimensions of well-being research: predictive analytics and machine learning, intergenerational and spillover effects, and the impact of policy design and governance quality on well-being. Developing in five different but complementary chapters, the research investigates explicitly (1) machine learning tools to predict depression in old age, highlighting the potential of predictive analytics in well-being studies, (2) parental retirement spillover effects on adult children's subjective well-being, (3) ethnic disparities and the impact of anti-discrimination policies in the digital platform context, (4) Bayesian spatial models for well-being composite indicators constructions, (5) the role of governance quality for environmental well-being.

Chapter 1, "Predicting Depression in Old Age: Combining Life Course Data with Machine Learning," assesses the predictive power of individual socioeconomic life biographies combined with machine learning algorithms in predicting depression in old age. We derive life biographies and depression measurements in later life from retrospective data collected in the Survey of Health Aging and Retirement in Europe. We operationalize adult life biographies using a *sequence analysis* approach. We then optimize six machine learning algorithms with increasing flexibility, i.e., Logistics Regression, three Regularized Regressions, Extreme Gradient Boosting, and Artificial Neural Network, on four different representations of life biographies. These various data representations reflect the increasing complexity of the input dataset given to the model.

Moreover, we stratify the sample by sex to investigate whether the predictive ability of these algorithms differs based on sex. We then apply the SHAP framework to identify the most critical predictors among the many that were used.

We show that depression is indeed predictable, better in women than in men. The Extreme Gradient Boosting performs best among the various algorithms, and the optimal performance is reached using a semi-structured data configuration. New predictors emerged for both sexes. Going *regularly to the dentist* throughout life decreases the likelihood of depression in old age, while higher *emotional life*

entropy increases the risk of depression. More broadly, these results show that depression is predictable from socioeconomic characteristics and that machine learning models can broaden the boundaries of knowledge by identifying previously neglected predictive factors.

Chapter 2, "The Old Folks at Home: Parental Retirement and Adult Children Well-being," investigates the causal effects of retirement on the well-being of older parents and their adult children in the United Kingdom. Using data from two nationally representative household surveys, the British Household Panel (BHPS) and the UK Household Longitudinal Study (UKHLS), we construct a panel of parent-child dyads to track changes in their socioeconomic characteristics and well-being around parental retirement.

We employ two causal identification strategies. The first uses the Statutory Pension Age, at 60 for women and 65 for men, as an exogenous cutoff in a Fuzzy Regression Discontinuity design. Our second-stage results indicate that retirement positively affects retirees' well-being, increasing life and leisure satisfaction and mental health while decreasing financial well-being. Maternal retirement positively impacts adult children's well-being, improving life and income satisfaction. This spillover effect is most pronounced among adult children with young children, those in lower income percentiles, and those living closer to their mothers.

The second strategy exploits two UK Pension Acts, implemented in 2010 and 2018, which raised the Statutory Pension Age from 60 to 66 for women and from 65 to 66 for men, using a difference-in-difference approach. We found no significant spillover effects of maternal retirement on adult children's well-being but observed negative effects from paternal retirement. Our key finding is that public policies can have inter-generational spillover effects with significant distributional consequences, emphasizing the importance of both financial and time transfers. These spillovers should be considered when designing policies that alter retirement behaviour.

Chapter 3, "Small Pictures, Big Biases: the Adverse Effects of an Airbnb Anti-discrimination Policy," examines the issue of ethnic discrimination against service providers in the emerging digital platforms market, using Airbnb as a case study. Combining scraped data from Airbnb with a state-of-the-art Vision Transformer for image classification, we first explore ethnic disparities in two market outcomes: occupancy rates and prices. Ordinary least square regression estimates, controlling

for a wide range of traditional and novel covariates, reveal that Black hosts have lower occupancy rates than comparable White hosts. Still, there is no significant difference in pricing. The disparity between Asian and White hosts is smaller but still significant for both outcomes, while no disparity is found for Hispanic hosts.

Next, we evaluate the impact of an anti-discrimination intervention introduced in October 2018, utilizing difference-in-differences and event study methodologies. The intervention involved reducing the size of profile photos from 225 to 109 square pixels. Surprisingly, our findings indicate a significant and robust negative causal effect of the policy, which increased the ethnic disparity in occupancy rates between Black and White hosts by approximately four percentage points. We interpret this outcome as resulting from the policy's reduction of positive signals potentially inferred by Airbnb guests from profile pictures, which inadvertently led them to rely more heavily on skin color when assessing a host's quality.

Chapter 4, "Spatial Comprehensive Well-Being Composite Indicators based on Bayesian Latent Factor Model: evidence from Italian Provinces," argues that Well-Being, as a macro indicator for policymakers, requires precise and informative statistical constructs while remaining interpretable. It, therefore, explores an innovative Bayesian approach to constructing Well-Being composite indicators that leverage the spatial correlation among the units of analysis, which improves the estimate's precision and provides a more informative estimate through uncertainty quantification.

We use this method on data from the Italian "Equitable and Sustainable well-being dashboard", which measures several elementary well-being indicators over time in the 110 Italian municipalities. We first group the elementary indicator into three macro Well-Being dimensions: Social, Economic and Environmental. Therefore, we apply the Bayesian Latent Factor model to obtain a composite indicator for each Italian municipality in each well-being dimension. We then descriptively assess the resulting provinces' well-being ranking through maps and bar plots and compare these rankings to the one produced by the widely adopted Mazziotta-Pareto approach. We show that our proposed methodology results in similar but more informative estimates with respect to this last approach. Moreover, this approach is also practical for dealing with missing data in the elementary indicators.

Chapter 5, "Quality of government for environmental Well-Being? Subnational

Evidence from European Regions", tackles some limitations of Chapter 4, namely the shortage of elementary indicators for the precise quantification of the Environmental dimension of Well-Being. It also delves into an empirical question, exploring the role of institutional quality in influencing environmental well-being at the sub-national level in Europe.

A dashboard of sixteen elementary environmental indicators divided into four macro groups is collected: air, soil, water, and energy quality, measured in 233 European regions (NUTS-2), and the quality of government index for three years: 2010, 2013, and 2017. Government quality is defined as "the extent to which states perform their required activities and administer public services in an impartial and uncorrupted manner". Therefore, we construct composite indicators of environmental well-being by adopting the same methodology as in the previous chapter for each environmental macro dimension. First, we highlight significant inequalities within countries regarding environmental well-being levels. Then, we estimate a battery of spatially lag regression models, where the dependent variable is the composite indicators of environmental well-being and the independent variable is the quality of governments measured with a lagged time. We show a robust and significant positive correlation between the quality of governments and each dimension of environmental well-being.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of economic growth*, *1*, 385–472.

Agan, A., & Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, *133*(1), 191–235.

Ahn, D., Yang, J., Cha, M., Yang, H., Kim, J., Park, S., Han, S., Lee, E., Lee, S., & Park, S. (2023). A human-machine collaborative approach measures economic development using satellite imagery. *Nature Communications*, *14*(1), 6811.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C., & Smith, J. (2021). O brother, where start thou? sibling spillovers on college and major choice in four countries. *The Quarterly Journal of Economics*, *136*(3), 1831–1886.

Angelucci, M., & Di Maro, V. (2010). Program evaluation and spillover effects: Impact-evaluation guidelines. *Washington, DC: Inter-American Development Bank*.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of personality and social psychology*, *36*(8), 917.

Cherlin, A. J., Chase-Lansdale, P. L., & McRae, C. (1998). Effects of parental divorce on mental health throughout the life course. *American Sociological Review*, 239–249.

Doleac, J. L., & Hansen, B. (2020). The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden. *Journal of Labor Economics*, *38*(2), 321–374.

Easterlin, R. A. (1973). Does money buy happiness? *The public interest*, *30*, 3.

Feagin, J. R., & McKinney, K. D. (2005). *The many costs of racism*. Rowman & Littlefield Publishers.

García-Miralles, E., & Leganza, J. M. (2024). Joint retirement of couples: Evidence from discontinuities in denmark. *Journal of Public Economics*, *230*, 105036.

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*, 1240–1248.

Guan, N., Guariglia, A., Moore, P., Xu, F., & Al-Janabi, H. (2022). Financial stress and depression in adults: A systematic review. *PloS one*, *17*(2), e0264041.

Ilciukas, J. (2023). Fertility and parental retirement. *Journal of Public Economics*, *226*, 104928.

Jackson, S. E., Hackett, R. A., & Steptoe, A. (2019). Associations between age discrimination and health and wellbeing: Cross-sectional and prospective analysis of the english longitudinal study of ageing. *The Lancet Public Health*, *4*(4), e200–e208.

Jannani, A., Sael, N., & Benabbou, F. (2021). Predicting quality of life using machine learning: Case of world happiness index. *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, 1–6.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters viii: Aggregate and individual governance indicators, 1996-2008. *World bank policy research working paper*, (4978).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2004). Unemployment alters the set point for life satisfaction. *Psychological science*, *15*(1), 8–13.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.

OECD. (2023). *Pensions at a glance 2023: OECD and G20 indicators*.

Oparina, E., Kaiser, C., Gentile, N., Tkatchenko, A., Clark, A. E., Neve, J.-E. D., & D'Ambrosio, C. (2022). Human wellbeing and machine learning.

Sansone, D. (2019). Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, *81*(2), 456–485.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2024). Using sequences of life-events to predict human lives. *Nature Computational Science*, *4*(1), 43–56.

Simeonova, E., Currie, J., Nilsson, P., & Walker, R. (2021). Congestion pricing, air pollution, and children's health. *Journal of Human Resources*, *56*(4), 971–996.

Stiglitz, J. E., Sen, A., Fitoussi, J.-P., et al. (2009). Report by the commission on the measurement of economic performance and social progress.

WHO. (2021). *Comprehensive mental health action plan 2013–2030*. World Health Organization.

Chapter 1

Predicting Depression in Old Age: Combining Life Course Data with Machine Learning

Montorsi, C., Fusco, A., Van Kerm, P., & Bordas, S. P. (2024). Predicting depression in old age: Combining life course data with machine learning. Economics & Human Biology, 52, 101331.

1.1 Introduction

Population ageing is one of the critical challenges of our times (United Nations, Department of Economic and Social Affairs, 2019). The share of the EU population above the age of 65 will reach almost 30% by 2050 (starting from 19.2 % in 2016). Understanding well-being in old age is therefore a priority. Mental health is a crucial aspect of it, with mental illness having detrimental individual consequences – such as a negative impact on productivity (Beck et al., 2011) – and bearing important costs for society – the annual cost of depression and anxiety amounts to USD 1 trillion for the global economy (OECD and European Union, 2018; The Lancet Global Health, 2020). To date however, mental health in old age surprisingly received less attention than in other age groups. Due to discrimination and stigma associated with ageing, mental disorders in old age are under-treated and under-diagnosed in primary care settings (WHO, 2017). From a policy perspective, it appears crucial to provide preventive tools to help identify at-risk populations and anticipate the onset of depression in old age.

Predicting depression is, however, a challenging task. A non-linear combination of individual biographies, predetermined genetic and epigenetic factors, and possibly cultural influences likely shape depression risks (Kennedy, 2001).

Previous research has shown that mental disorders threatening successful ageing may result from complex combinations of circumstances and events taking place throughout the entire life span, as well as exposure to different institutions (Colman & Ataullahjan, 2010; Currie & Almond, 2011; Falkingham et al., 2020; Layard et al., 2014; Pakpahan et al., 2017). Central to life course epidemiology theories is that health-related states are shaped by endogenous and exogenous forces interacting through time. Notably, the effect of these forces is different along the life cycle, with 'sensitive' periods of development where specific experiences may exert a marked influence over future history (Bornstein, 1989). The complexity and high dimensionality of the mechanisms at play when we examine how life course experiences influence old age outcomes challenge traditional modelling techniques.

Against this background, this paper exploits supervised machine learning algorithms (SML) to assess how much individual life course data can predict clinical

measures of depression in old age. We employ life course biographies of individuals aged 50+ contained in a dedicated module of the Survey of Health, Ageing, and Retirement in Europe (SHARE) collected in nineteen European countries. The data contain both clinical measures of depression at the time of interview and retrospective data collected in a different wave of the survey. The retrospective data records livelihood information during childhood and tracks rich biographical information related to employment and activity status, marital status and family composition, location of residence, housing status, general health conditions, and periods of financial stress.

The predictive performance of such biographical information on old-age depression is informative in two ways. First, it indicates the potential long-term impacts of (the absence of) life events on future mental health conditions. Of course, predictive performance does not imply direct causation. Nonetheless, given the relatively rich set of past life course variables (and of contemporaneous variables) that we consider and because of the way we extract potentially meaningful signals from sequences of events, we trust the predictive power of biographies on old-age depression is plausibly informative of long-term health effects of life events and conditions. Second – even in the absence of identification of causal relationships – the ability of biographical data to predict depression and identify individuals at a heightened health risk can be useful from an epidemiological and prevention perspective. This may complement existing approaches that mine, e.g., electronic health records or medical screening (Nemesure et al., 2021 and Garriga et al., 2022).

Our raw biographical data contain, for each respondent, annual status information from the ages of 15 to 49 across six domains of life (activity status, health, location of residence, home ownership status, family situation and financial situa-

There is a growing literature using SML in economics and social sciences focusing on objective outcomes: in the fields of criminal justice (Berk, 2012), economic well-being measurement at a granular level using mobile data or satellite imagery (Engstrom et al., 2016), means testing in developing countries (McBride & Nichols, 2018), high school dropouts (Sansone, 2019) or inequality of opportunity measurement (Brunori & Neidhöfer, 2021). SML models have also been used with subjective data in the context of analysis of affective forecasting (Wilson & Gilbert, 2005), prediction of happiness, health, and depression from a combination of high-frequency data and surveys (Jaques et al., 2015 and Oparina et al., 2022), daily stress prediction from mobile phone and weather conditions data (Bogomolov et al., 2014), and to predict depression among university students (Choudhury et al., 2019).

Notably, clinical information included in electronic health records are not harmonized across countries (Bincoletto, 2020) and may therefore limit possibilities to learn from cross-national population records. Medical screening may be costly and invasive. Biographical data, on the contrary, may potentially be collected relatively easily – especially when few biographical markers are of importance, as we find.

tion). In their most flexible forms, SML algorithms can be fed with raw, unstructured input data – here all respondents' 34 annual statuses in all 6 dimensions – and algorithmically extract the most predictive combination of input. Such models, however, risk producing predictions based on combinations of statuses that are hard to interpret. Therefore, we also examine the predictive performance of models run on alternative, more structured encodings of biographical information. Predictions obtained from structured encodings of biographies can lead to outcomes with clearer interpretation. If rolled out for clinical predictions, that may lead to more straightforward data collection requirements. This can also enhance the performance of simple off-the-shelf algorithms by reducing the dimensionality of the prediction task (Christodoulou et al., 2019).

Our structured data encoding strategy is based on sequence analysis (Abbott, 1995). Similarly to DNA molecules representation, this approach represents an individual life history as an ordered string of characters – a *sequence* – representing each life domain. Following the methodology proposed in Wahrendorf et al., 2013, Studer and Ritschard, 2016 and Bolano and Studer, 2020, we extract interpretable information from the sequences in two ways of increasing dimensionality: (i) by grouping similar sequences into a small number of groups (by cluster analysis), and (ii) by summarizing sequences by a set of sequence attributes (timing of events, ordering of events, duration of states and entropy of the sequence). As we show, using the latter as input to SML models outperforms both the coarse clustering approach and the fully unstructured raw input for predicting depression later in life and highlights easily interpretable sequence attributes as markers of risk.

Machine learning models have proved to excel at capturing complex non-linear interactions and generally outperform conventional linear prediction models for health outcomes (Leist et al., 2022). As mentioned above, recent studies have focused on symptoms of depression (Librenza-Garcia et al., 2021) predicted from electronic health records (EHR) or medical screening (Garriga et al., 2022; Nemesure et al., 2021), or have looked at post-therapeutic-treatment outcomes (Sajjadian et al., 2021). We therefore benchmark the predictive capacity of life course information with two different predictor sets. The first is a minimal predictor set including

We focus on these life domains because they are all the ones whose life sequences we could construct from the SHARELIFE questionnaire. In addition, we chose this specific lifetime frame to capture as much information about adulthood as possible but excluding socio-economic circumstances that might co-occur with the depression measurement.

only demographic variables, e.g., country of residence, age, birth cohort, interview year, educational level, age at first childbirth, and migration status. The second is the predictive capacity of clinical studies targeting depression risk (Garriga et al., 2022).

A shortcoming of many SML approaches is the difficulty of interpreting predictions. Lack of interpretability might result from the intrinsic black box character of SML methods such as neural networks or ensemble methods such as Gradient Boosting. To obtain insights about predictor roles, we follow the recent literature on ”interpretable machine learning” (Lundberg & Lee, 2017) and use the Shapley additive explanations (SHAP) to assess the predictive power of the input variables.

The results highlight well-known predictors of depression, such as age, health, childhood conditions, and education level, for both sexes. Additionally, we identify two biographical predictors for depression risk. The first is the entropy of the general life trajectory. The general life trajectory collects information about periods of happiness, stress, financial strain, and hunger that individuals experience at one or more moments. The entropy in the trajectory indicates how many of these periods have happened throughout the life course.

The higher the number of remarkable periods (positive or negative), the greater the likelihood of depression in later life. This reveals the potentially detrimental impact of instability in life trajectories. The literature on depression epidemiology has so far neglected this lifetime measure.

The second is low lifetime utilization of dental care services, which stands out as a key predictor of later-life depression. Although a direct causal impact is plausible – e.g., through persistent pain due to poor dental health or aesthetics impacts of poor dentition – it is likely that low lifetime dental care captures limited access to general, non-essential healthcare services or poor self-care.

On the technical side, we find that more complex SML algorithms do not systematically outperform standard logistic models in our data. We achieve the highest predictive performance when we use semi-structured input data based on life sequence attributes combined with a Gradient Boosting model. Compared to the minimal benchmark, the predictive accuracy, judged by the Area Under the Precision-Recall curve (PR-AUC), increases by around ten percentage points for the Gradient Boosting when using sequence attributes as input. This improvement confirms that

biographical data matters for predicting accuracy. Moreover, we find a difference of around five percentage points in predictive accuracy between the structured and unstructured sequences' encoding. Our results confirm that imposing some structure on the underlying input data both enhances interpretability and improves predictive performance.

Independently of the algorithm, a PR-AUC of 0.77 for females and 0.65 for males is a reliable maximum given the type of available information (Saito and Rehmsmeier, 2015). These predictive results are consistent with other studies targeting similar outcomes (Bogomolov et al., 2014; Garriga et al., 2022).

The remainder of the article is organized as follows. Section 1.2 describes the source survey, and Section 1.3 outlines the methods. Section 1.4 elaborates on the predictive findings. Section 1.5 provides a discussion of the results, while section 1.6 concludes the paper.

1.2 Data

1.2.1 The sample

Our analysis draws data from the bi-annual Survey of Health, Ageing, and Retirement in Europe (SHARE). The SHARE survey has collected individual-level data on health, socio-economic status, and social and family networks of more than 123,000 individuals aged 50+ from 2004 to 2020 (Börsch-Supan, 2019). A feature of SHARE that makes it particularly suitable for our application is the retrospective questionnaire SHARELIFE. The questionnaire was included in the third (2008 – 2009) and the seventh waves (2017) and includes modules on several individual life dimensions, such as childhood conditions, partnerships and parenting, employment trajectories, migration, housing, and financial histories. SHARELIFE collects retrospective information using the so-called "life-grid approach". The life-grid approach supplements the interviewers' questions with a graphical longitudinal representation of the respondents' lives. The interviewer fills the grid during the interview, ending with information for each respondent's age.

A potential problem of retrospective data is the recall bias if respondents systematically incorrectly remember past life events (Havari & Mazzonna, 2015). To limit the influence of recall bias, we include only individuals aged 50 to 88 when

answering the SHARELIFE interviews and exclude individuals who had difficulties responding to the retrospective questionnaire.

The countries covered are Austria, Belgium, Switzerland, Czech Republic, Germany, Denmark, Spain, France, Greece, Italy, Poland, Sweden, the Netherlands, Luxembourg, Hungary, Portugal, Slovenia, Estonia, and Croatia. The analysis is conducted by pooling these countries and stratifying the sample by sex. The final respondent sample includes 58,323 respondents (32,984 females, 25,339 males).

1.2.2 Measure of depression

The outcome of interest is a binary indicator of clinical depression. We construct this measure from the 12 EURO-D items of depressive symptoms stored in the mental health module of SHARE (waves 1 to 6, SHARELIFE waves 3 and 7 are excluded). The construct, face, and content validity and reliability of this measure have widely been validated by the literature (Prince et al., 1999; Walker & Schimmack, 2008). EURO-D scores range between a minimum of 0 and a maximum of 12, with a score of at least 4 indicating ”clinical depression” and below 4 ”no depression”.

SHARE is a bi-annual panel study. For some individuals, we have repeated observations. We define them as ‘depressed’ if they have at least one measure of depression over the observation period, i.e., we selected the individual observation where depression is positive. If they have more than one depression measurement, we randomly choose one. This selection criterion results in a relatively high depression prevalence, 49% for females and 29% for males. Figure 1.1 and Appendix 1.A illustrate the distribution of depression prevalence within the analyzed countries at NUTS3 and NUTS2 levels, stratified by sex.

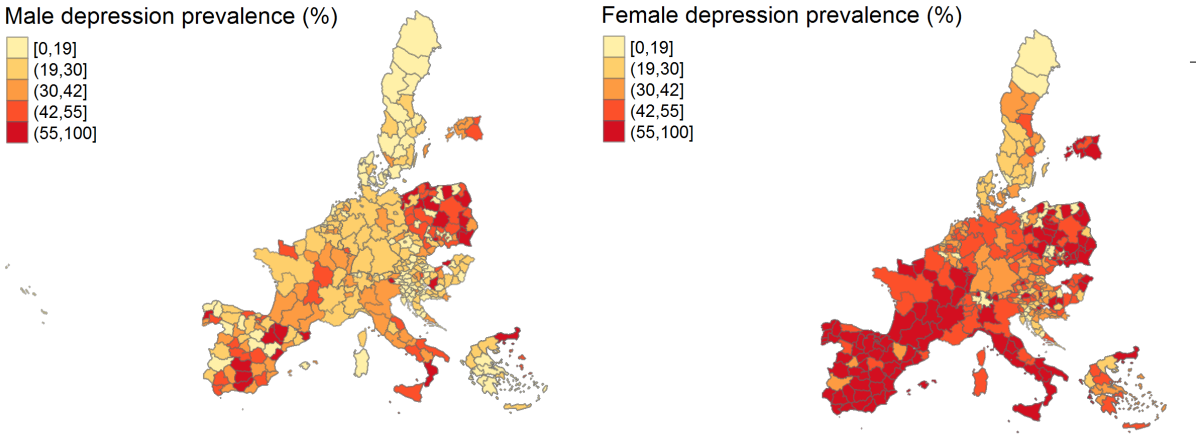
In line with the literature, the depression prevalence is unequally distributed

Another potential threat underlying retrospective data is the influence of depression on memory tasks (LeMoult & Gotlib, 2019). For example, scholars found that depressed patients recall negative episodes with disproportionate frequency than non-depressed patients (Dillon & Pizzagalli, 2018). When it comes to autobiographical memory tasks, depressed individuals tend to over-report positive autobiographical memories and recall little details (Williams et al., 2007). Studies suggest that cognitive bias may occur in depressed individuals but do not provide consistent evidence of memory accuracy loss, where patients misremember specific events. Moreover, these studies target a population of severely depressed individuals. Our reference sample is representative of the general population, characterized by mild to low depressive symptoms. The issue is also partially addressed as the measurement of depression and the collection of retrospective information occurred at years of distance (see below).

For more information on this depression threshold validation, see Prince et al. (1999). We also analyzed different threshold values, e.g., 3, 5, and 6 as a sensitivity test. We observe a decrease in models’ predictive ability at higher threshold values.

(Van de Velde et al., 2010). First, females display a higher propensity to be depressed than males. Second, except for Croatia, we observe a gradient in depression prevalence across more economically developed and less economically developed countries. Croatia, Denmark, the Netherlands, Sweden, and Switzerland have the lowest depression prevalence. At the top of the distribution are Poland, Portugal, France, Italy, and Estonia. We address these differences in the analysis by stratifying the sample by sex and including country and macro-region (NUTS1) information and country-age interactions among the predictors.

Figure 1.1: Map of depression (%) among individuals aged 50+ at the NUTS3 level, by sex



1.2.3 Predictor sets

Drawing upon a large literature on the determinants of depression risks (e.g., Arpino et al., 2018; Atkins et al., 2020; Blazer et al., 1985; Flèche et al., 2021; Kisely, 2016; Zheng et al., 2021), our predictors can be classified into three main groups: demographic characteristics, descriptors of family background and childhood conditions, and, crucially, adulthood biographies descriptors that we construct from life sequences.

Demographic characteristics

Demographic characteristics include birth cohort, age at the time of depression measurement, sex, country and macro-region of residence, educational achievement, migration status, and age at first childbirth (see Appendix 1.B for details).

We also incorporate three predictors pertaining to adulthood that we did not

directly mine from life sequences: two indicators of health – whether the respondent has ever measured blood pressure during their lifetime and whether they regularly visited the dentist – and an indicator of an individual's socioeconomic status between 20 and 30 (Wahrendorf et al., 2013).

Childhood conditions

Childhood variables come from the childhood retrospective module of SHARELIFE. We include childhood conditions that, according to the literature, influence later-life well-being: childhood socioeconomic position, material deprivation, childhood health and family composition (Arpino et al., 2018; Clark & Lee, 2021; Flèche et al., 2021; Wood et al., 2017). See Appendix 1.B for details.

Adulthood biographies

We encodes adulthood biographical information over the ages 15 to 49 following the literature on sequence analysis. Sequences are objects of an ordered list of successive elements chosen from a finite list of states, named alphabet (Abbott, 1995). Sequences' strength relies on their holistic perspective over the life course, which enables capturing complex dynamics and life transitions. Notwithstanding the extolled potential highlighted in their first formulation, their use in scientific applications has been limited (Aisenbrey & Fasang, 2010; Liao et al., 2022; Studer & Ritschard, 2016).

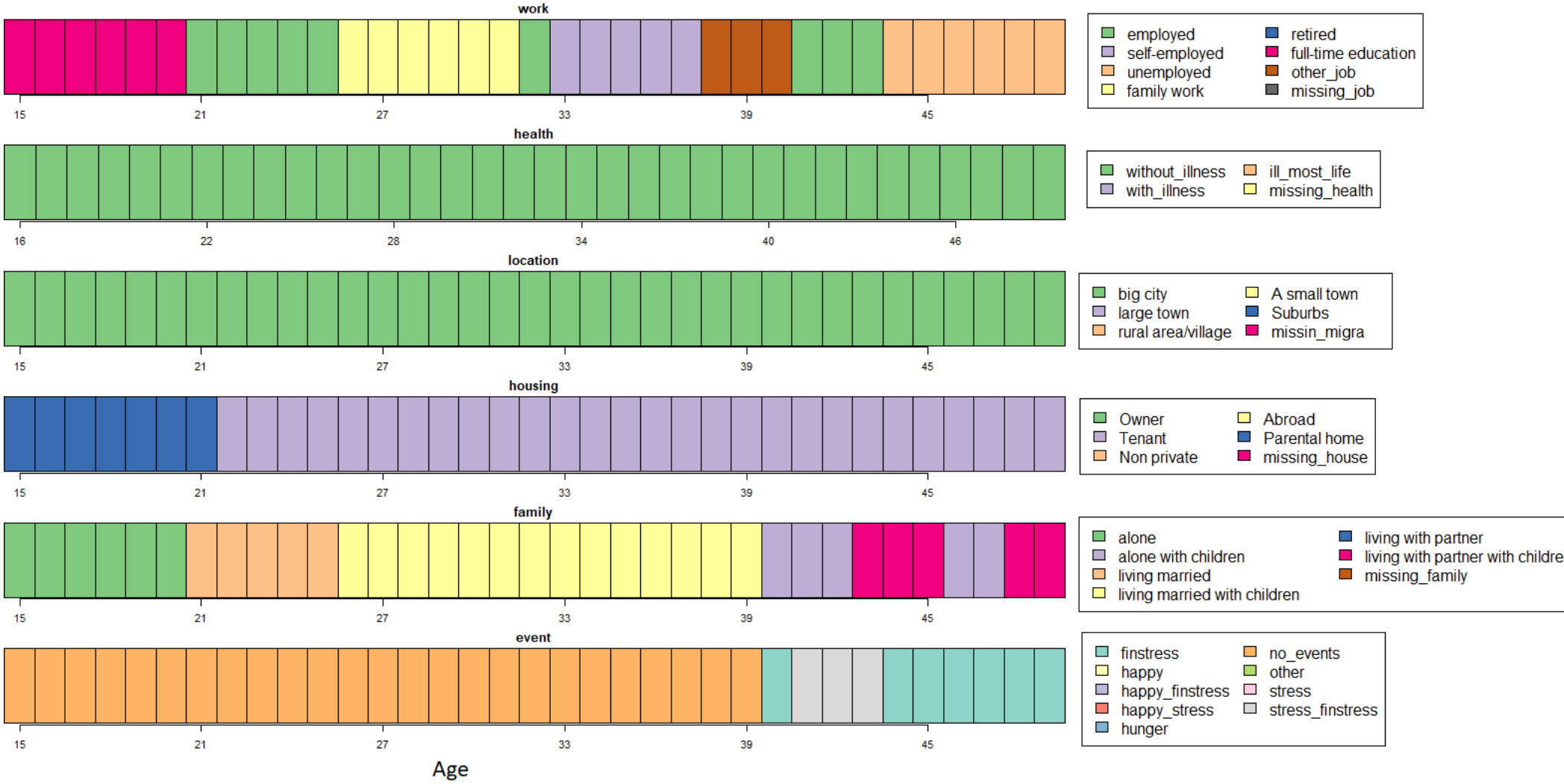
We constructed life sequences for six variables: work status, housing arrangement, family, health, residence location, and general life events. The family sequence combines information on partner history, children's history, and cohabitation history. We obtained the work, health, and housing arrangement sequences from the gateway portal harmonized sequences (Program on Global Aging, Health, and Policy, 2021). We added additional states in case of missing data. We constructed sequences and corresponding figures in R using the TraMineR package (Gabadinho et al., 2011). Figure 1.2 exemplifies the six life sequences we constructed for each

This analysis uses information from the Harmonized SHARE Life History dataset and codebook, Version B as of February 2020, developed by the Gateway to Global Aging Data in collaboration with the University of DÜsseldorf. The development of the Harmonized SHARE Life History was funded by the National Institute on Ageing (R01 AG030153, RC2 AG036619, R03 AG043052).

Missing information in the life trajectories is rare except for the location of residence. Appendix 1.C reports the alphabet and definitions for the six variables.

individual in our sample.

Figure 1.2: Representation of six life dimensions for an individual. Each rectangle represents an age and each colour represents a different state



Different ways of coding sequences can be used to construct sets of variables input to the SML models. The first and lowest dimensional sequence representation is that of sequences' cluster membership. Clusters categorize sequences based on similarities in the states and transitions over life sequences. The second set decodes life history information based on four specific sequence characteristics. Finally, the third set describes life trajectories in unstructured form where each binary variable represents a combination of age and life trajectory status.

Sequences' clusters Typologies or clusters are the most common sequence configuration employed in social science applications. This involves grouping similar sequences in a small number of clusters. To construct clusters, we follow a standardized procedure. We start by creating individual sequences for each life course variable. Next, we assess the dissimilarities between sequences for each life dimension and create a distance matrix. Several measures exist to estimate dissimilarities among sequences (Gabadinho et al., 2011). In our empirical application, we use the Dynamic Hamming Distance (DHD) proposed by Lesnard, 2010. Once we get a matrix of sequences' dissimilarities, we perform cluster analysis (Ward methods) to regroup the more similar sequences into clusters. To select the number of clusters, we used the nbClust package in R (Charrad et al., 2014) and retained the solution that maximizes the silhouette index (Kaufman & Rousseeuw, 2009). The optimal number of clusters differs among the analyzed dimensions and across sexes. In the family trajectories, the algorithm detected four clusters for both females and males; in the work trajectories, the algorithm detected two clusters for males and four clusters for females; in health trajectories, the algorithm identified three groups for both sexes; in the residence location trajectory, it detected five clusters for both sexes; lastly, in the general life trajectories, the algorithm finds two clusters for males and three clusters for females. This predictor set contains around 180 variables. (Appendix 1.D provides detailed information on the adopted clusters.)

This measure belongs to the class of "edit" distances, which equates distance to the minimal cost of transforming one sequence into another. What determines this cost are the number of operations required to transform one sequence into another and the cost of each operation. There are two primary operations: substitution and insertion-deletion (indel). The distinct characteristics of the DHD distance are its state-dependent and time-variant substitution costs (e.g., the cost of changing from the state of "in education" to the condition of "employment" differs whether we are at the beginning of or at the end of the working career)

Sequences' features The second data configuration we employ is that of sequence features. We extract four well-established features: ordering, duration, timing, and entropy (Billari et al., 2006; Bolano & Studer, 2020; Studer & Ritschard, 2016).

Ordering refers to the order in which the states appear along the sequence. The social norms attached to this sequential aspect are well-documented. For example, the social consequences of having a child before marriage differ from when the first childbirth occurs after marriage.

To capture indicators of the ordering, we employ "frequent sub-sequence mining." A sub-sequence is frequent if it occurs in more than 10% of sequences. Given a sequence s, e.g., A-B-C, a sub-sequence z is any subset of s that respects the ordering of s, e.g., A-B, B-C, A-C, A, B, C is all sub-sequences of s. Frequent sub-sequence are not mutually exclusive since the pattern A-B-C does not exclude the pattern A-B. We extract a list of frequent sub-sequences for each life course variable. We generate indicator variables for each extracted sub-sequence indicating the presence or absence of the sub-sequence in each trajectory.

The second extracted sequences feature is the spell duration. The duration represents an individual's overall time in a specific sequence's state, for example, how long it has been married. This sequencing feature mirrors the concept of exposure to a given event. It has a crucial role in life course studies. For example, Mossakowski (2009) estimate a negative effect of unemployment spell duration on mental health and well-being.

The concept of *timing* refers to the age at which a transition from one state to another occurred. The timing of events plays a relevant social role, given the presence of age-related social norms. For example, the critical period model emphasized the differential impact on the mental health of experiencing unemployment at the beginning or middle of a working career. The same applies to childbirth or marriage age. In our analysis, we included a timing indicator that refers to the time of each transition to different states over five years. For example, if an individual gets married at age 25, we created an indicator variable "20-25.married" that captures the transition to married and the time of its occurrence.

Finally, we included a measurement of within-sequence *entropy*. The within-sequence entropy measures the stability of the states along the trajectory. This measure does not account for states' order or distinguish between positive and

adverse conditions. The life course literature has largely overlooked the dimension of entropy when predicting future life outcomes. However, our definition of entropy is consistent with the concept of life changes. The entropy is equal to zero when the individual has experienced no life changes throughout the trajectory and one when the same amount of time has elapsed in each possible variable's states. Life changes were discussed in various areas of research (see Haslam et al., 2021, Lin and Ensel, 1989 and Rahe, 1975). By increasing uncertainty in life, life changes call into question the sense of autonomy and self-continuity, possibly impairing wellness and mental health. Moreover, adverse effects of life changes are more likely to arise when individuals lack substantial social support (Lin & Ensel, 1989).

Following the literature on sequence analysis, we measure within sequence entropy (normalized) by the Shannon entropy formula:

$$h(p_1, \ldots, p_a) = \frac{-\sum_{i=1}^{a} p_i log_2(p_i)}{log_2 a}$$

where a is the size of the sequence alphabet and p_i is the proportion of occurrence's of the ith state in the considered sequence. The sequence features' configuration counts around 360 predictors, combining sub-sequencing, duration, timing, and entropy.

Unstructured sequence Our third input data configuration involves creating binary columns that represent a combination of age and life trajectory state for each life dimension. For example, the housing sequence has six potential states (e.g., owner, tenant, non-private, abroad, parent house, missing) combined with the 34 years; it results in 6*34 = 204 binary columns for the combination of each year of age and housing modality. This procedure generates a high-dimensional predictor set of over a thousand predictors. The resulting configuration is highly unstructured and sparse but has the potential, in principle, to outperform more structured configurations when combined with SML algorithms.

Pre-processing

We pre-processed all input data before feeding them into the algorithms. Around 20% of respondents have at least one missing value in the selected childhood, demographic or life-course variables. We imputed missing values separately by country

to preserve differences in mean and covariance structures and encode missing values patterns. As a sensitivity analysis (available upon request), we repeated our exercise, dropping observations with missing variables; the results were unchanged.

Pre-processing also involved excluding predictors with excessive collinearity. For more information on pre-processing, see Appendix 1.E.

1.3 Methods: Machine learning predictions

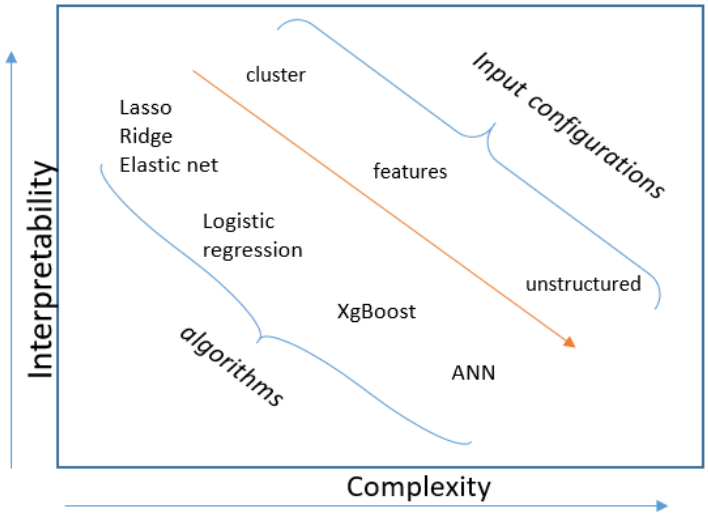
1.3.1 Models

Machine learning algorithms are data-driven methods that help discover patterns otherwise neglected by traditional models. For an extensive illustration of these methods, see Hastie et al. (2009). Supervised Machine learning (SML) algorithms automatically build a predictive function $\mathcal{F}$ that maps $X \in \mathcal{X}$, the predictor set, to a prediction $\hat{y} \in \mathcal{Y}$. The predictive function $\mathcal{F}$ is what we estimate from the data.

As Athey (2019) pointed out, estimating a range of machine learning models is advisable as the predictive performance of different models differs with alternative input data configurations. Accordingly, we considered four types of standard, largely off-the-shelf, models. We started from the simplest approach based on logistic regression models fit by maximum likelihood. We then proceeded along the trade-off between model complexity and interpretability by applying regularization methods (logistic regression with lasso, ridge and elastic-net penalties), a tree-based method (Extreme Gradient Boosting XGBoost), and an artificial neural network. (Appendix 1.F describes each of these four models in more details.) Figure 1.3 illustrates the input configurations and machine learning models explored. Moving along the diagonal, from the top left to the bottom right, learning models and input configurations increase their complexity.

We imputed missing data with the R package missForest (Stekhoven & Bühlmann, 2012), which relies on an iterative method based on the Random Forest algorithm. This non-parametric algorithm has the advantage that it can handle mixed types of variables

Figure 1.3: Models-Inputs framework



1.3.2 Hyperparameter selection and assessment of predictive performance

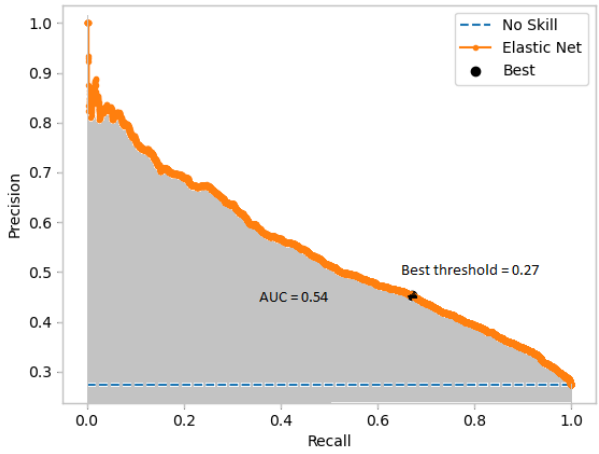
All our models were trained to maximize the Area Under the Precision-Recall curve (PR-AUC). When predicting health outcomes, the recall (sensitivity or true positive rate) and the Area Under the Precision-Recall curve (PR-AUC) are standard model selection and evaluation criteria (Steyerberg et al., 2010). These metrics overcome the accuracy paradox and train the models to maximize their depression detection.

The recall represents the proportion of depressed people that the model correctly identifies. It is an essential metric for disease diagnostic tools, as it measures the reliability of the diagnostic tool in detecting the disease. We optimised this metric as predicting an individual is not at risk of depression when developing depression symptoms is more costly than the opposite mistake. The precision represents the proportion of genuinely depressed among those predicted as depressed.

The PR curve – the orange line in Figure 1.4 – relates, at all possible threshold probability values, the recall and the precision. The baseline of the PR curve is the horizontal line with the y-value equal to the depression prevalence in the sample. The AUC measures the area under the PR curve. When the AUC is near one, the classifier separates classes perfectly. An AUC near zero indicates the worst separability measure. Larger values of this metric indicate better model performance. Optimizing this metric allows achieving the optimal trade-off between

precision and recall (Saito and Rehmsmeier, 2015).

Figure 1.4: Precision-Recall curve



Prior to model estimation, we divided the sample into a training subsample (80% of the total sample) and a separate test subsample (20%). The PR-AUC was then used both to determine model hyperparameters (in the training data) and assess the final models' predictive performance (in the test data).

Generally, an SML model requires optimizing two parameters in a training dataset: structural parameters (e.g., the parameters of the logistic regression) and model hyper-parameters (e.g., the shrinkage factor in regularisation methods or the tree depth in tree-based approaches). Optimizing the structural parameters is embedded in the model estimators and typically involves minimizing a loss function (aggregate prediction error). To determine the model's hyper-parameters, we used stratified ten-fold cross-validation with random or grid search (or both) in the hyper-parameter space.

Predictive performance measures reported in the next section are the PR-AUC measures achieved in the test data with the model (hyper-)parameters obtained in the training data.

We divided the training data set into ten folds of equal size, preserving the percentage of samples for each target class in each fold. We repeated the same procedure ten times for each fold and hyper-parameter configuration: keeping out one fold (validation set) and training the model on the remaining nine folds. We chose the hyper-parameters combination that maximizes the predictive score across folds. See Appendix 1.G for a description of the hyperparameters obtained for each model

1.4 Results

1.4.1 Predictive Performance

Figure 1.5 and 1.6 illustrate the PR-AUC across models, input structures, and sex. The box plots illustrate the distribution of predictive performances in the training sample. The red dots indicate the predictive score in the test sample. We benchmark our models' performance against two different baselines. The first is a minimal baseline where we only use demographic information (age, interview year, interview season, country and macro-region of residence, education, cohort, children, and migrant status). The second is a model from clinical studies that use health records and medical screenings.

The first input structure we explore is that of the sequences' clusters (two-dashed box plot). The predictor set counts around 180 predictors, remaining relatively small not to create multicollinearity issues. The best-performing models are the Gradient Boosting (XGBoost) and regularized regressions, which reach around 0.687 PR-AUC in the females' sample and 0.467 PR-AUC in the males' sample.

We then change the life sequences' configuration structure from clusters to sequence features (long-dashed box plot). The predictor set now counts around 360 predictors. The PR-AUC in the training and test sets increases in all classifiers. The best model is the Gradient Boosting, which settles at a PR-AUC of 0.768 for females and 0.647 for males.

Finally, we try the unstructured sequence configuration (dot-dashed box plot). This predictor set counts around 390 predictors. The PR-AUC is slightly higher than the cluster configuration but smaller than the sequences' features configuration. In this highly multicollinear setting, the noise in the input structure increases substantially. The models find less relevant patterns in the data to improve predictive performance.

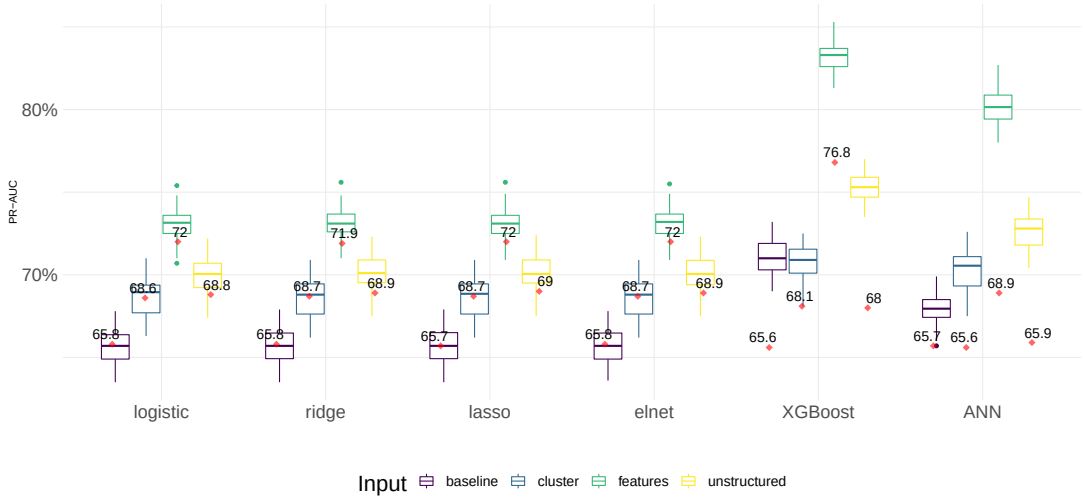
The similarity in predictive performance across different algorithms indicates that complex algorithms are comparable to the traditional logistic regression model for the data at hand. Similar conclusions are reached in other clinical studies (for a systematic review, see Christodoulou et al., 2019).

Differences in predictive performance estimates between the training and test samples reflect overfitting when the algorithms perform well on the training data but poorly out-of-sample.

Compared with the baseline predictor set (solid-line box plot), models trained with life course predictors achieve better predictive performance. The PR-AUC improves by around ten to twenty percentage points for all algorithms: life course information does increase the ability for depression risk detection.

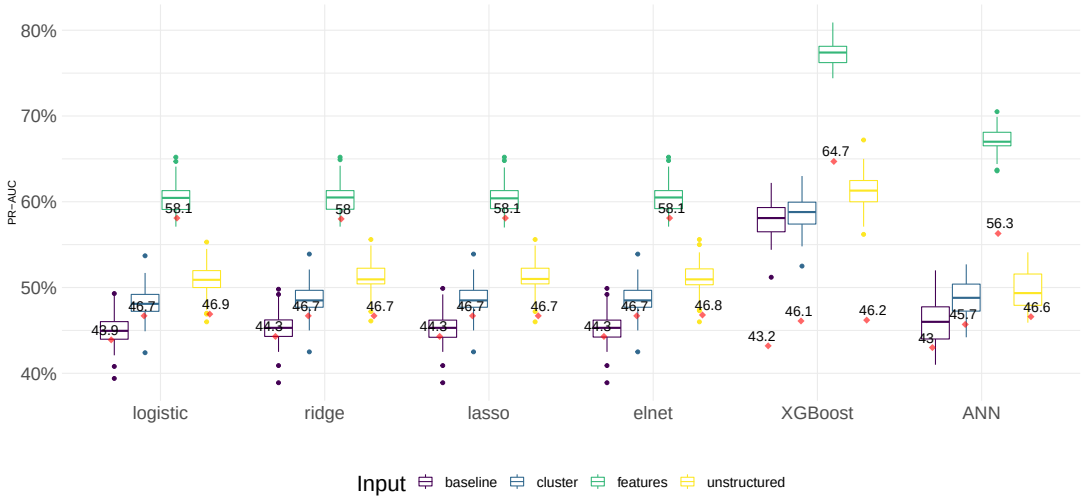
Significant differences in predictive performance across sexes emerge. Using the same type of life course information, the models achieve a PR-AUC of around 10 percentage points higher in the females' sample than in the males' sample. This result highlights a need to differentiate depression diagnosis procedures by sex, as it would be highly insufficient to look only at socio-demographic factors to detect male depression.

Figure 1.5: Area-Under-Precision-Recall curve across models and input configurations, female sample



Note: The dots indicate TEST predictive score, box plots are the training score. Line types represent different predictor sets

Figure 1.6: Area-Under-Precision-Recall curve across models and input configurations, male sample



Note: See Figure 1.5

The comparison of our life-course approach with current medical studies reveals that biographical data have a similar predicting ability than concurrent medical screening and health records (see Librenza-Garcia et al., 2021 and Garriga et al., 2022). These clinical studies reach a ROC-AUC of around 0.71-0.75. We reach a maximum ROC-AUC of around 0.757 for females and 0.772 for males (see Appendix 1.H for other predictive performance metrics).

We illustrate two potential explanations for these prediction results. The first explanation targets the inner nature of the target variable we analyzed and the predictors we included. The expected out-of-sample test error, for a given value x_0 and a given learning algorithm $f(x)$, can always be written as the sum of three fundamental quantities:

$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{variance}} + \underbrace{[\text{Bias}(\hat{f}(x_0))]^2}_{\text{bias}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error variance}} \tag{1.1}$$

The variance of the models is given by the changes in the model's parameter estimates when changing the training set x_0. The bias refers to the error in fitting a real-life problem with an oversimplified function.

In simple terms, the learning procedure aims to create a model that can make accurate predictions by reducing the number of errors it makes. However, there is a limit to how accurate predictions can be, given by the element $\text{Var}(\epsilon)$ in equation

1.1. Our model-building procedure, based on repeated cross-validations, ensures we have taken all necessary steps to minimize model errors.

The value of the irreducible component depends on the nature of the data and the amount of information available. This paper targets a self-reported depression indicator that may be related to unobserved respondents' characteristics. For example, respondents can drastically change their depression perception if a dramatic accident occurs a few days before the interview. This situation will remain unobservable no matter how many predictors we include in the predictor set. Another element we do not control for but that correlates with depression is genetic endowment (Levinson, 2006).

As an extension of the main results presented here we report ML models' performances for different EURO-D depression thresholds, i.e., 3, 5, and 6 in Appendix 1.J. For all algorithms, increasing the EURO-D depression thresholds deteriorates the classification performance. The fewer depressed examples in the sample, the less information is available for the algorithms to learn significant depression patterns. This reduction affects the models' detection ability of depressed cases.

1.4.2 SHAP values across sexes

This section sheds light on the complexity behind the ML algorithms' predictions. We sought to understand how variables contributed to generating the final individual predicted probabilities for each sex. We employ the Shapley Additive exPlanations (SHAP) method for this aim. This method has provided reliable and consistent results in previous research (Lundberg & Lee, 2017). SHAP relies on the Shapley values concept, which originates from the collaborative game theory (Shapley, 1953). Contrary to other variable importance metrics, the SHAP framework is the only explanation method that can, in principle, explain any predictive model, i.e., it is a model-agnostic tool (Lundberg et al., 2020; Molnar, 2020).

The general idea underlying the SHAP framework is to estimate, for any given

A second potential explanation addresses the dimensionality of the sample used to train the algorithms. Black-box SML models typically need big data to exploit their predictive ability fully. The empirical sample in this analysis counts around 60,000 observations. This relatively small sample size may limit the model's detection ability. To test this explanation, we trained our models by increasing fractions of the training data, from 10% to 90%. For each training fraction, we computed the test PR-AUC. We observed that with 20% of the training data, the test AUC reaches almost the same score as the whole training set (see Appendix 1.I), suggesting training-size independence and ruling out limited sample sizes as a source of the limited ability of models to predict depression accurately.

model, a simpler explanation model, which corresponds to an interpretable approximation of the initial model. Given a vector $\boldsymbol{x}$ of p predictor variables, $\boldsymbol{x} = [x_1, \ldots, x_p]$, and a trained model f, SHAP approximate the model f with a simple explanation model g that has the following form:

$$g(\boldsymbol{z}) = \phi_0 + \sum_{i=1}^{p} \phi_i z_i. \tag{1.2}$$

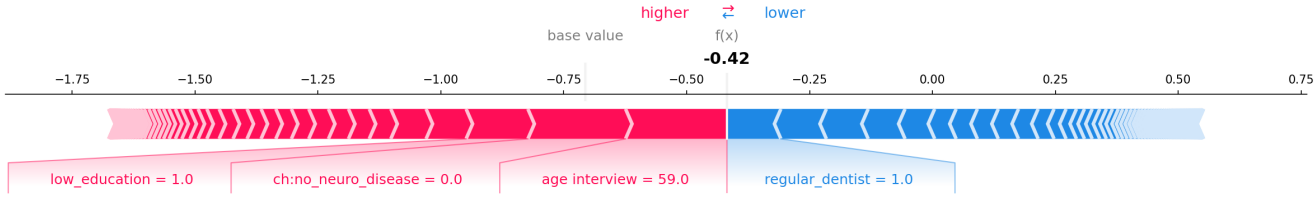
In equation 1.2, $\boldsymbol{z} = [z_1, \ldots, z_p]$ is a coalition vector, where z_i is equal to 1 if the variable x_i is present and 0 if the variable is absent, p is the number of predictors, and $\phi_i \in R$ is the variable i contribution to the model predictions, i.e., the Shapley value. The Shapley value ϕ_i is then estimated through the following equation:

$$\phi_i(f, x) = \sum_{\boldsymbol{z} \subset \boldsymbol{x}} \frac{\mid \boldsymbol{z} \mid !(p - \mid \boldsymbol{z} \mid -1)!}{p!} [f(\boldsymbol{z}) - f(\boldsymbol{z} \backslash i)] \tag{1.3}$$

where $\mid \boldsymbol{z} \mid$ is the number of non-zero entries in $\boldsymbol{z}$.

The exact estimation of each ϕ_i may be computationally infeasible. However, Lundberg and Lee (2017) and Lundberg et al. (2018) introduced efficient algorithms to estimate such values in the case of Gradient Boosting and Neural Networks. To understand the meaning of SHAP values, Figure 1.7 shows which features contributed to the model's prediction for a single randomly selected observation (a not depressed Slovenian female 59 years old). The bold number -0.42 represents the predicted odds of being depressed, which translates to 0.39 in probability terms. We colour the features essential to predicting this observation in lighter and darker shades. A lighter shade represents features that pushed the model probability score higher, and a darker shade indicates features that moved the score lower. Features that had more of an impact on the score locate closer to the dividing boundary between red and blue. The bar represents the impact size. For this random individual, what contributes more to the increase in the depression score is having a low education level, having had a neurological disease in childhood, and her age at the time of the interview. What pushes down the risk of depression is to have regularly used dental care.

Figure 1.7: A SHAP force plot of a single individual



Note: In **bold** is the predicted odd ratio, corresponding to a 0.39 probability of being depressed. Light shades represents features that pushed the model probability score higher, and dark shades represents features that pushed the score lower

We observed a higher predictive power of relying on semi-structured life sequences' features as input data, i.e., timing, duration, ordering, and entropy. In what follows, we illustrate SHAP values of individual input variables for this data configuration only. Aggregating the results for all test predictions, Figure 1.8 illustrates the SHAP summary plot for the top twenty predictors for the Gradient Boosting for males and females. The summary plot combines variable contribution with variable effects. Each point on the summary plot is a Shapley value for a feature and an instance. The feature's position on the y-axis is determined by the absolute average Shapley value and on the x-axis by the Shapley value. The colour represents the variable's value from low (blue) to high (pink). The number on the right of each variable name corresponds to the average SHAP value across all observations.

Comparing SHAP values across sex highlights idiosyncratic and common factors. In line with the literature, we found that, for both females and males, material deprivation in childhood ("childhood: no basic facilities," and "childhood: rooms per capita"), low education and low subjective childhood health predict higher depression likelihood (Clark & Lee, 2021; Layard et al., 2014). These childhood-specific variables appear in all input configurations and all predictive algorithms. No matter the amount and type of adult life course information we provide to train the algorithms, childhood conditions matter most.

We identify low lifetime utilization of dental care services as a predictor of depression for both sexes. As discussed earlier, the interpretation of this predictor is not unequivocal: it may catch the direct effect of dental care across the lifetime (due to a lack of infrastructure or high dental care costs) but, equally, it may proxy

unobserved factors related to broader access to health care or self-care behaviours. Either way, this predictor stands out as a key marker of depression in old-age and can be collected easily in individual health questionnaires.

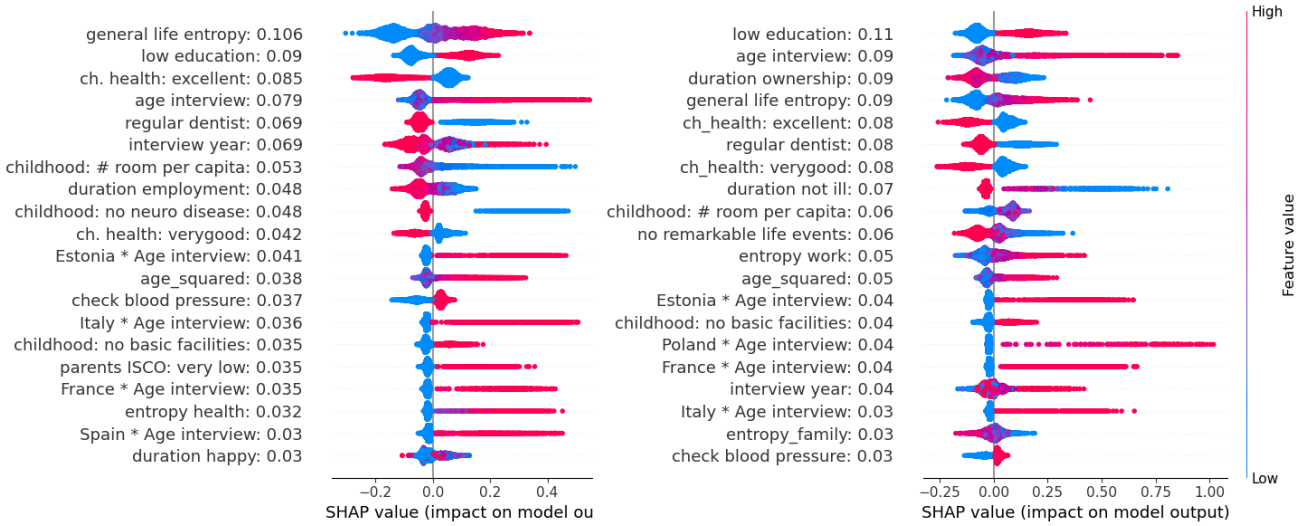
For both sexes, the entropy within the general life sequence ("general life entropy") increased the prediction of depression later in life. Life entropy refers to the number of remarkable emotional periods, such as happiness, stress, and financial stress, adults have undergone across their life courses (see Appendix 1.2.3 for descriptive statistics). This finding highlights the importance of monitoring emotional stressors throughout one's lifetime.

We extracted two distinctive male predictors: the entropy in the work and family sequences. High-frequency changes in work status and low-frequency changes in family status predict higher depression risk for males but not females. As idiosyncratic female predictors, we find that low parents' ISCO and short employment duration increase the likelihood of depression.

Another finding concerns the heterogeneity in the predicting power of age across countries. The SHAP values for the age-countries interaction variables are high in all models and for each sex. However, age's contribution to the likelihood of depression differs across countries. In countries such as Italy, Poland, Hungary, Portugal, and Spain, being older contributes to an increased risk of depression. In countries such as Sweden, Denmark, and Switzerland, the effect is the opposite, with higher ages associated with a lower risk of depression. This result also prompts further investigation of predictors that may vary by country.

To shed some further light on the underlying mechanisms, we examined the SHARELIFE question: *What are the reasons you [have never gone / weren't going] to a dentist regularly for check-ups or dental care?*. We compare the prevalence of individuals selecting each listed reason (1. Not affordable — 2. Not enough information about this type of care — 3. Not usual to get this type of care — 4. No place to receive this type of care close to home — 5. Other reasons). Results suggest that both dental care cost affordability and individual habits equally explain the low utilization of dental care services, and this is consistent across countries (see Appendix 1.K for maps).

Figure 1.8: Shapley values for Gradient Boosting, female (right) and male (left)



Note: Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature rank and on the x-axis by the Shapley value. Colours indicate the relationship between variables and depression probability: dark to light signifies a positive correlation with depression, while light to dark signifies a negative correlation. The number on the right of each variable name corresponds to the average SHAP value across all observations

1.5 Discussion and potential limitations

At the heart of this analysis is the assessment and comparison of supervised machine learning techniques applied to various life course data configurations to predict the risk of depression in people over fifty years old. Three key findings emerged from our analysis.

First, biographical information may foreshadow later-life depression outbreaks, but not perfectly. Second, structuring large-scale life course information is useful to improve prediction tasks. All models achieved the highest predictive performance when fed with life course sequence features, considering the duration, timing of state transitions, the state ordering, and – importantly – entropy within the life course. With this data configuration, the best-performing models predict depression risk with a PR-AUC of 0.77 for females and 0.65 for males in the training and test sample. Compared to the benchmark set of demographic predictors, life course information improves predictive performance by about ten percentage points for both sexes. Life course information yields similar predictive performance than other clinical studies using electronic medical records (Garriga et al., 2022); hence,

our findings suggest that some life-course-based predictors may be integrated into clinical data to improve the diagnosis of depression.

The third relevant aspect of this analysis stems from the sex stratification and the extraction of depression predictors. We merged respondents from nineteen European countries and trained our models independently for females and males. We did so because of the substantial gap in depression prevalence. Females suffer almost twice as much from depression in all the analyzed countries. Stratifying by sex was a straightforward way to shed light on potential differences in depression patterns. SHAP feature extraction combined with this stratification revealed established patterns of depression and new predictor variables. For both females and males and in all models, material deprivation in childhood, poor health in childhood and adulthood, and low education predict a higher probability of depression later in life. The duration of ownership or lease predicts a lower likelihood of depression. As a new predictive feature, we identify life trajectory entropy. Entropy is the frequency of changes in condition over the years. Specifically, males and females who go through multiple periods of happiness, stress, financial stress, or hunger are more likely to experience depression later in life. Higher entropy in the work sequence and lower entropy in the family sequence increase the probability of depression for men only. Similarly to previous well-being studies, our results stress the long-lasting influence of early childhood conditions on later-life well-being outcomes (Clark & Lee, 2021; Zheng et al., 2021). Finally, entropy in the life course also stands as a depression predictor for both sexes– individuals experiencing repeated changes in life domains throughout the life course are at higher risk of mental health problems later in life.

The country-age interactions show significant heterogeneity across countries. In countries like Hungary, Poland, and Italy, increasing age predicts a higher probability of depression. On the other hand, in Denmark, Switzerland, and Sweden, increasing age predicts a lower likelihood of depression. This heterogeneous effect remains unexplained, but it may be related to the differences in social welfare and pension systems across European countries.

Our findings should be considered in light of some data limitations. First, our predictive models are based on retrospective life course data. Therefore, they are subject to potential biases arising from the long-term recall of events and

circumstances long before the time of the survey. While we took steps to limit the extent of these issues (focusing on individuals between 50 and 88 years old and excluding those having difficulties answering the retrospective questionnaire), these biases challenge the fidelity and accuracy of such information, downward estimating the overall predictive ability. We used this data source for lack of comparable sources, covering almost all European countries with such a large set of observations and information range.

Second, our sample may suffer from survival bias. Indeed, we are analyzing old-age people. It may be the case that the chronically depressed or ill individuals died before the time of the survey or refused to answer the questionnaire. Our sample is, therefore, likely limited to a selection of moderately depressed or healthy individuals, and our findings may not generalize to the whole population.

Our study shows a promising path in using such life course trajectories to predict later-life outcomes. Therefore, future research may rely on prospective cohort studies that track individuals over time to obtain more precise life trajectories, with a high potential for improving the accuracy of later-life outcomes prediction.

1.6 Conclusion

Most of the existing literature focuses on treatments for depression and uses electronic health records for predictive tasks. This study set out to look at retrospective data and test their ability to predict later-life depression. Our analysis shows that past life trajectories may foreshadow later life depression outbreaks. These results, which shed light on the relationship between retrospective information and accuracy in depression prediction, call for complementing diagnostic tools and electronic health records with retrospective data on the life course.

We live in a World with high social and economic uncertainties, e.g., the Covid pandemic, the Ukraine war, and the economic downturn. These macro phenomena affect individuals by creating financial and work instabilities, displacements, and insecurities, thus increasing entropy in all life dimensions.

Given our findings on the health risks associated with life instability, future research could explore how welfare systems' interventions could mitigate the impact of sudden and abrupt events.

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual review of sociology*, *21*(1), 93–113.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The" second wave" of sequence analysis bringing the" course" back into the life course. *Sociological Methods & Research*, *38*(3), 420–462.

Arpino, B., Gumà, J., & Julià, A. (2018). Early-life conditions and health at older ages: The mediating role of educational attainment, family and employment trajectories. *PloS one*, *13*(4), e0195320.

Athey, S. (2019). The impact of machine learning on economics. In *The economics of artificial intelligence* (pp. 507–552). University of Chicago Press.

Atkins, R., Turner, A. J., Chandola, T., & Sutton, M. (2020). Going beyond the mean in examining relationships of adolescent non-cognitive skills with health-related quality of life and biomarkers in later-life. *Economics & Human Biology*, *39*, 100923.

Beck, A., Crain, A. L., Solberg, L. I., Unützer, J., Glasgow, R. E., Maciosek, M. V., & Whitebird, R. (2011). Severity of depression and magnitude of productivity loss. *The Annals of Family Medicine*, *9*(4), 305–311.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population/Revue Européenne de Démographie*, *22*(1), 37–65.

Bincoletto, G. (2020). Data protection issues in cross-border interoperability of electronic health record systems within the european union. *Data & Policy*, *2*, e3.

Blazer, D., George, L. K., Landerman, R., Pennybacker, M., Melville, M. L., Woodbury, M., Manton, K. G., & Jordan, K. (1985). Psychiatric disorders: A rural/urban comparison. *Archives of general psychiatry*, *42*(7), 651–656.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., & Pentland, A. (2014). Daily stress recognition from mobile phone data, weather conditions and individual traits. *Proceedings of the 22nd ACM international conference on Multimedia*, 477–486.

Bolano, D., & Studer, M. (2020). *The link between previous life trajectories and a later life outcome: A feature selection approach* (Working Paper 82). Swiss National Competence Center in Research.

Bornstein, M. H. (1989). Sensitive periods in development: Structural characteristics and causal interpretations. *Psychological bulletin*, *105*(2), 179.

Börsch-Supan, A. (2019). Survey of Health, Ageing and Retirement in Europe (SHARE), wave 7. *Release version 7.1.0*, *7*(0).

Brunori, P., & Neidhöfer, G. (2021). The evolution of inequality of opportunity in germany: A machine learning approach. *Review of Income and Wealth*, *67*(4), 900–927.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*, 1–36.

Choudhury, A. A., Khan, M. R. H., Nahim, N. Z., Tulon, S. R., Islam, S., & Chakrabarty, A. (2019). Predicting depression in bangladeshi undergraduates using machine learning. *2019 IEEE Region 10 Symposium (TENSYMP)*, 789–794.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22.

Clark, A. E., & Lee, T. (2021). Early-life correlates of later-life well-being: Evidence from the Wisconsin longitudinal study. *Journal of Economic Behavior & Organization*, *181*, 360–368.

Colman, I., & Ataullahjan, A. (2010). Life course perspectives on the epidemiology of depression. *The Canadian Journal of Psychiatry*, *55*(10), 622–632.

Currie, J., & Almond, D. (2011). Human capital development before age five. In D. Card & O. Ashenfelter (Eds.), *Handbook of labor economics* (pp. 1315–1486, Vol. 4). Elsevier.

Dillon, D. G., & Pizzagalli, D. A. (2018). Mechanisms of memory disruption in depression. *Trends in Neurosciences*, *41*(3), 137–149.

Engstrom, R. N., Hersh, J., & Newhouse, D. (2016). Poverty in HD: What does high resolution satellite imagery reveal about economic welfare ? *Available online: Pubdocs. worldbank*.

Falkingham, J., Evandrou, M., Qin, M., & Vlachantoni, A. (2020). Accumulated lifecourse adversities and depressive symptoms in later life among older men and women in England: A longitudinal study. *Ageing and Society*, *40*(10), 2079–2105.

Flèche, S., Lekfuangfu, W. N., & Clark, A. E. (2021). The long-lasting effects of family and childhood on adult wellbeing: Evidence from British cohort data. *Journal of Economic Behavior & Organization*, *181*, 290–311.

Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of statistical software*, *40*(4), 1–37.

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*, 1240–1248.

Haslam, C., Haslam, S. A., Jetten, J., Cruwys, T., & Steffens, N. K. (2021). Life change, social identity, and health. *Annual Review of Psychology*, *72*, 635–661.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Havari, E., & Mazzonna, F. (2015). Can we trust older people's statements on their childhood circumstances? Evidence from SHARELIFE. *European Journal of Population*, *31*(3), 233–257.

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2015). Multi-task, multi-kernel learning for estimating individual wellbeing. *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kennedy, G. J. (2001). *Geriatric mental health care: A treatment guide for health professionals*. Guilford press.

Kisely, S. (2016). No mental health without oral health. *The Canadian Journal of Psychiatry*, *61*(5), 277–282.

Layard, R., Clark, A. E., Cornaglia, F., Powdthavee, N., & Vernoit, J. (2014). What predicts a successful life? A life-course model of well-being. *The Economic Journal*, *124*(580), 720–738.

Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S. P., Muniz-Terrera, G., & Wade, S. (2022). Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances*, *8*(42).

LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical Psychology Review*, *69*, 51–66.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, *38*(3), 389–419.

Levinson, D. F. (2006). The genetics of depression: A review. *Biological Psychiatry*, *60*(2), 84–92.

Liao, T. F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A. E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., Struffolino, E., & Studer, M. (2022). Sequence analysis: Its past, present, and future. *Social Science Research*, *107*, 102772.

Librenza-Garcia, D., Passos, I. C., Feiten, J. G., Lotufo, P. A., Goulart, A. C., de Souza Santos, I., Viana, M. C., Benseñor, I. M., & Brunoni, A. R. (2021). Prediction of depression

cases, incidence, and chronicity in a large occupational cohort using machine learning techniques: An analysis of the ELSA-Brasil study. *Psychological Medicine*, *51*(16), 2895–2903.

Lin, N., & Ensel, W. M. (1989). Life stress and health: Stressors and resources. *American Sociological Review*, *54*(3), 382–399.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

McBride, L., & Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, *32*(3), 531–550.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mossakowski, K. N. (2009). The influence of past unemployment duration on symptoms of depression among young women and men in the United States. *American Journal of Public Health*, *99*(10), 1826–1832.

Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports*, *11*(1), 1–9.

OECD and European Union. (2018). *Health at a glance: Europe 2018*.

Oparina, E., Kaiser, C., Gentile, N., Tkatchenko, A., Clark, A. E., Neve, J.-E. D., & D'Ambrosio, C. (2022). Human wellbeing and machine learning.

Pakpahan, E., Hoffmann, R., & Kröger, H. (2017). The long arm of childhood circumstances on health in old age: Evidence from SHARELIFE. *Advances in Life Course Research*, *31*, 1–10.

Prince, M. J., Reischies, F., Beekman, A. T., Fuhrer, R., Jonker, C., Kivela, S.-L., Lawlor, B. A., Lobo, A., Magnusson, H., Fichter, M., et al. (1999). Development of the EURO–D scale–a European Union initiative to compare symptoms of depression in 14 European centres. *The British Journal of Psychiatry*, *174*(4), 330–338.

Program on Global Aging, Health, and Policy. (2021). Gateway to Global Aging Data [University of Southern California with funding from the National Institute on Aging (R01 AG030153). https://g2aging.org/ Accessed: 2021-11-11].

Rahe, R. H. (1975). Epidemiological studies of life change and illness. *The International Journal of Psychiatry in Medicine*, *6*(1-2), 133–146.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, *10*(3), e0118432.

Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., et al. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, *51*(16), 2742–2751.

Sansone, D. (2019). Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, *81*(2), 456–485.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2:307–317.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*(1), 128–138.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(2), 481–511.

The Lancet Global Health. (2020). Mental health matters. *The Lancet. Global Health*, *8*(11), e1352.

United Nations, Department of Economic and Social Affairs. (2019). *World population ageing 2019* (tech. rep.). United Nations.

Van de Velde, S., Bracke, P., & Levecque, K. (2010). Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Social science & medicine*, *71*(2), 305–313.

Wahrendorf, M., Blane, D., Bartley, M., Dragano, N., & Siegrist, J. (2013). Working conditions in mid-life and mental health in older ages. *Advances in Life Course research*, *18*(1), 16–25.

Walker, S. S., & Schimmack, U. (2008). Validity of a happiness implicit association test as a measure of subjective well-being. *Journal of Research in Personality*, *42*(2), 490–497.

WHO. (2017). Mental health of older adults [Accessed on March 17, 2023].

Williams, J. M. G., Barnhofer, T., Crane, C., Herman, D., Raes, F., Watkins, E., & Dalgleish, T. (2007). Autobiographical memory specificity and emotional disorder. *Psychological bulletin*, *133*(1), 122.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131–134.

Wood, N., Bann, D., Hardy, R., Gale, C., Goodman, A., Crawford, C., & Stafford, M. (2017). Childhood socioeconomic position and adult mental wellbeing: Evidence from four British birth cohort studies. *PloS one*, *12*(10).

Zheng, X., Shangguan, S., Fang, Z., & Fang, X. (2021). Early-life exposure to parental mental distress and adulthood depression among middle-aged and elderly Chinese. *Economics & Human Biology*, *41*, 100994.

Appendix

1.A Depression

Table 1.A.1: Depression prevalence among people aged 50+ within countries and across sexes

	Female		Male	
Country	N	%	N	%
Austria	1861	42	1275	24
Germany	2339	43	2092	26
Sweden	797	29	738	17
Netherlands	1101	38	918	22
Spain	2390	58	1870	33
Italy	2463	62	2088	37
France	2266	63	1696	39
Denmark	1998	37	1752	22
Greece	1678	46	1314	25
Switzerland	1436	41	1203	24
Belgium	3058	53	2546	33
Czech Republic	2696	46	1781	28
Poland	1093	71	840	53
Luxembourg	584	42	491	24
Hungary	843	42	536	22
Portugal	605	63	448	34
Slovenia	1904	39	1312	23
Estonia	2830	59	1609	39
Croatia	1042	36	830	19
Total	32944	49	25339	29

Note: The depression prevalence (%) indicates the share of respondents with at least one depression measurement in the observation period. We define Depression as a binary indicator that takes the value of one when the respondent has more than three EURO-D symptoms. EURO-D symptoms are sadness, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness

1.B Descriptive statistics

In this section, we describe the construction of demographics and childhood conditions variables and provide descriptive statistics.

Specifically, we include five demographics (in addition to sex and country and macro-region of residence): age, birth cohort, migration status, highest educational attainment and age at first childbirth. We determine migration status by comparing individuals' country of birth to their country of residence at the time of the interview. The highest educational level is classified according to the International Standard of Education (ISCED) and is categorized as low, medium, and high.

The childhood variables come from the childhood retrospective module of SHARELIFE. We include childhood conditions that, according to the literature, influence later-life well-being. These are childhood socioeconomic position, material deprivation, and family composition (see Arpino et al., 2018, Clark and Lee, 2021, Flèche et al., 2021, Arpino et al., 2018 and Wood et al., 2017). We capture childhood socioeconomic position with parental occupational status, the number of books at home, and the number of rooms per person. We code material deprivation with a binary variable indicating whether or not the respondent had the main basic facilities (i.e., hot and cold water, fixed bathroom, indoor bathroom, and central heating). We encode family composition as an indicator of living with a biological father, with siblings, or with grandparents. Regarding health during childhood, we follow Pakpahan et al. (2017) and encode three childhood health-related variables: self-rated health, ever in the hospital and ever missed school because of health. In addition, we include binary variables related to specific diseases (infections, cardiovascular and neurological diseases). We code the parent's occupational position according to the ten main occupational groups of ISCO (see International Labour Office, 1949). Thus, we reduced the groups to four skill levels following standard procedure (see Wahrendorf et al., 2013). To fill the missing values in the parents' occupational position, we retrieve the answers longitudinal respondents provide in the demographic modules of regular SHARE questionnaires in waves 2, 4, and 5.

In addition to the abovementioned variables, we incorporate three predictors pertaining to adulthood that we could not directly mine from life sequences. These entail two indicators of health– whether the respondent has ever measured blood

pressure during their lifetime and whether they regularly visited the dentist, and an indicator of an individual's socioeconomic status between 20 and 30 (Wahrendorf et al., 2013).

Table 1.B.1: Summary statistics demographic variables, male sample

	Depressed			Not depressed		
Variable	N	Mean	SD	N	Mean	SD
Age at interview	8249	66	9.9	19757	63	8.5
Season interview:	8249			19757		
Autumn	1372	17%		4580	23%	
Spring	3316	40%		5948	30%	
Summer	2243	27%		6874	35%	
Winter	1318	16%		2355	12%	
Birth cohort:	8249			19757		
$<$ 1930	608	7%		803	4%	
1930-1939	2000	24%		3607	18%	
1940-1949	2636	32%		7083	36%	
1950-1959	2528	31%		6735	34%	
1960-1969	477	6%		1529	8%	
interview year	17608	2011	3.7	18338	2010	3.8
Migrant:	8249			19757		
No	7494	91%		18183	92%	
Yes	755	9%		1574	8%	
Education	8249			19757		
High	1562	19%		5192	26%	
Low	3216	39%		5919	30%	
Medium	3003	36%		7990	40%	
No education	468	6%		656	3%	
When first child	8249			19757		
older 30 years old	1994	24%		5062	26%	
before 25 years old	2044	25%		4441	22%	
between25-30 years old	3051	37%		7697	39%	
no children	1160	14%		2557	13%	

Table 1.B.2: Summary Statistics demographic variables, female sample

	Depressed			Not depressed		
Variable	N	Mean	SD	N	Mean	SD
Age at interview	17608	65	9.7	18338	62	8.7
Season interview	17608			18338		
Autumn	3059	17%		4361	24%	
Spring	6758	38%		5549	30%	
Summer	4972	28%		6364	35%	
Winter	2819	16%		2064	11%	
Birth cohort	17608			18338		
$<$ 1930	1231	7%		693	4%	
1930-1939	3921	22%		2975	16%	
1940-1949	5605	32%		6072	33%	
1950-1959	5659	32%		6547	36%	
1960-1969	1192	7%		2051	11%	
Migrant	17608			18338		
No	15921	90%		16844	92%	
Yes	1687	10%		1494	8%	
Education	17608			18338		
High	2735	16%		4390	24%	
Low	7816	44%		6209	34%	
Medium	5860	33%		7030	38%	
No education	1197	7%		709	4%	
When first child	17608			18338		
older 30 years old	1921	11%		2241	12%	
before 25 years old	9390	53%		8977	49%	
between 25-30 years old	4513	26%		5180	28%	
no children	1784	10%		1940	11%	

Table 1.B.3: Childhood and adulthood predictors data set. Descriptive statistics by sex.

	Female			Male		
Variables and categories	N. imputed	Mean/Freq	SD or %	N.imputed	Mean/Freq	SD or %
Childhood conditions						
Occupation parents (ISCO) :	6313			5190		
Very High		12%			12%	
High		12%			12%	
Low		22869	0.64		17986	0.64
Very Low		4354	0.12		3182	0.11
Rooms per capita	3347	0.73	0.58	2871	0.76	0.66
Number of books:	2684			2374		
1 (0-10 books)		14030	0.39		11307	0.40
2 (11-25 books)		8337	0.23		6461	0.23
3 (26-100 books)		8248	0.23		6407	0.23
4 (101-200 books)		2702	0.08		1846	0.07
5 (more than 200 books)		2640	0.07		1995	0.07
Basic facilities	0	25814	0.72	0	20458	0.73
Ever in hospital	63	2252	0.06		1843	0.07
Ever missed school	344	4014	0.11	192	2925	0.10
No infectious disease	0	5530	0.15		5563	0.20
No neoplastic disease	0	33418	0.93	0	26124	0.93
No neuro disease		33347	0.93		26720	0.95
Childhood self-rated health:	32			23		
1 (”Fair, poor, spontaneous”)		4128	0.11		2601	0.09
2 (”Good”)		9484	0.26		6786	0.24
3 (”Very good”)		11431	0.32		8847	0.32
4 (”Excellent”)		10914	0.30		9782	0.35
Live with biological father	2610	31996	0.89	2357	25217	0.90
Live with biological brother	2610	30311	0.84	2357	23518	0.84
Adulthood predictors						
Ever blood pressure		24205	0.67		19128	0.68
Regular dentist	41	27596	0.77	25	19547	0.70
General life entropy	0	0.2	0.18	0	0.17	0.17
Housing entropy	0	0.39	0.16	0	0.42	0.15
Work-life entropy	0	0.22	0.17	0	0.18	0.15
Family-life entropy	0	0.39	0.14	0	0.4	0.14
Migration-life entropy	0	0.19	0.2	0	0.2	0.21
SES between 20 and 30:	0					
very high		1052	0.03		755	0.03
high		2276	0.06		2358	0.08
low		15806	0.44		11055	0.39
very low		4858	0.14		4759	0.17
not employed		11965	0.33		9089	0.32

1.C Construction of sequences

We construct life sequences for six variables: work status, housing arrangement, family, health, residence location, and general life events. We draw three sequences (work status, health, and housing) from the Gateway to Global Aging portal (Program on Global Aging, Health, and Policy, 2021). All the variables come from the SHARELIFE questionnaire.

To construct the work history, SHARELIFE asked respondents to report when they finished full-time education and question specific job spells. We used details on the start and end of respective job spells, and we determined if the gap was because of being unemployed (both searching and not searching for a job), home or family work, retirement, or a remaining group of others. The other category includes being sick or disabled, voluntary work, military services, and traveling.

To construct health history variables, SHARELIFE asked respondents how many periods of poor health or disability (lasting more than a year) they had in their life from age 16 onwards. If the number of periods of poor health or disability was more than three, people were automatically classified as "Ill most of their life" throughout their history. In contrast, if the respondent answered three or fewer periods, respondents were additionally asked to report when the respective periods were. SHARELIFE respondents reported the precise years when each period started and ended.

The housing arrangement histories combine details regarding the respondent's housing spells, including the reported year they left their parent's home and reported the year they established their household, if applicable. In SHARELIFE, respondents could report up to 28 housing spells that lasted six months or longer. We classify as non-private those types of residences that are hard to classify, such as rent-free or non-private residences (e.g., boarding schools, hospitals, or prisons). To classify residences as "abroad," we used information on whether the residence was in the country. We did not distinguish between types of residences abroad because the number of people who lived abroad was too small.

The family histories combine the children's, cohabitation, and partner's histories. The children's histories contain information on the age at which the respondent had or adopted their child and the number of children at each respondent' age. We

include information on death in the case that a child dies. The cohabitation history contains information on cohabitation spells. SHARELIFE asked respondents about when they started living with a partner (beginning of spell) and, if they stopped living with the same partner, the age at which they stopped living with them (end of a spell). The end of cohabitation could be because a partner died, a relationship broke up, a partner moved into a nursing or care home, or other reasons. The partner history distinguishes between married or non-married partnership and the alone status.

The residence location histories inform about the location where respondent report they had their accommodations. SHARELIFE asked respondents about housing spells and whether it was in a big city, rural area, large town, or small village for each period.

Finally, the general life history combines the period of stress, financial stress, happiness, and hunger. SHARELIFE asked respondents to reflect on their past life and report whether there was a distinct period during which they were happier, under more stress, with financial hardship, or suffering from hunger. In an affirmative answer, the respondents must report the starting and stopping years or whether the period was still ongoing. In a negative response, we classified the history as ”no events.”

Table 1.C.1: Sequences alphabet

Work status	Health	Location	Family	Event	Housing
Employed (E)	Not ill	Big city (BC)	Alone (A)	Financial stress (FS)	Owner (O)
Self-Employed (SE)	Ill	Large town (LT)	Alone with children (AC)	Happy (H)	Tenant (T)
Unemployed (U)	Ill mostly	Rural area (RA)	Married (M)	Hunger (Hu)	Non-private (NP)
Family work (FW)		Small town (ST)	Married with children (MC)	No events (NE)	Abroad (Ab)
Retired (R)		Suburbs (Sub)	With partner (P)	Stress (S)	Parental home (Par)
In education (FE)		Missing (NA)	With partner and children (PC)	Happy and Stress (H+S)	
Other jobs (Oj)				Happy and fin. stress (H+FS)	
Missing (NA)				Other events (Ot)	
				Stress and fin.stress (S+FS)	

1.D Composition of sequence clustering

Table 1.D.1: Prevalence of cluster solutions and measures of homogeneity for six the variables analyzed. Females sample.

	Work			**House**			**Family**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	26760	67	0.503	7923	20	0.67	2032	5	0.64
2	8017	20	0.634	27880	70	0.45	32178	80	0.62
3	2663	7	0.443	3986	10	0.47	2262	6	0.64
4	2349	6	0.185				3317	8	0.17
Pseudo R^2	0.52			0.57			0.58		
	General Life			**Health**			**Location**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	27026	68	0.62	37669	95	0.98	7196	18	0.7
2	5978	15	-0.17	1187	3	0.40	13339	34	0.7
3	6785	17	0.61	933	2	0.99	8527	21	0.60
4							7367	19	0.53
5							13360	8	0.51
Pseudo R^2	0.50			0.93			0.81		

Note: We measure homogeneity within clusters through the "Average Silhouette Width." Comparing the average distance of an observation from the other members of its cluster and its average weighted distance from the closest group. Low values indicate low cluster homogeneity. The pseudo R^2 informs to what extent the cluster solution allows explaining sequences' variability

Table 1.D.2: Prevalence of cluster solutions and measures of homogeneity for six the variables analyzed. Males sample.

	Work			**House**			**Family**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	26575	86	0.71	8790	28	0.38	24316	79	0.59
2	4394	14	0.55	17473	57	0.57	2289	7	0.69
3				2511	8	0.29	2213	7	0.46
4				2195	7	0.60	2151	7	0.01
Pseudo R^2	0.58			0.62			0.58		
	General Life			**Health**			**Location**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	27026	72	0.73	29602	95	0.98	5500	18	0.63
2	5978	27	0.17	861	3	0.40	10923	34	0.66
3				933	2	0.99	6753	21	0.58
4							5182	19	0.59
5							2611	8	0.48
Pseudo R^2	0.44			0.92			0.79		

Note: see Table 1.D.1

1.E Data pre-processing

We perform five main pre-processing steps on the three: imputation of missing data, one-hot encoding of categorical variables, removal of zero-variance and near-zero variance (sd ¡ 0.015) items, drop perfectly collinear variable and highly collinear variable (correlation ¿ 0.8) and normalizing predictors through the min-max normalization.

Although we could have performed more data pre-processing operations, e.g., eliminating variance inflation factor and variables with very low (but not zero) variance, we decided to perform only these five basic operations for this first empirical exploration. Indeed, automatic, uncontrolled elimination of variables could deprive the model of helpful information for learning that would affect the resulting predictions.

1.F Description of machine learning algorithms

The analyses uses four standard ML predictive algorithms: (i) logistic regression, (ii) regularized logistic regression, (iii) random forest (XGBoost) and (iv) artificial neural networks. These have become off-the-shelf approaches.

1.F.1 Logistic regression

When dealing with classification problems with a binary outcome, logistic regression models have been largely applied and have been proven to achieve high predictive performance, especially when compared to other simple probabilistic classification methods such as linear and quadratic discriminant analysis.

The logistic regression function is

$$\hat{Y} = \frac{exp^{(\hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}}{1 + exp^{(\hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}} \tag{1.4}$$

where $\hat{Y}$ is the probability that the target y is positive.

In the logistic regression model, the optimization criterion (loss function) is the log likelihood:

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \left\{ \sum_{i=1}^{N} y_i \ln(\hat{y}_i) + (1 - y_i)(1 - \ln(\hat{y}_i) \right\} \tag{1.5}$$

The logistic regression model has several advantages. Firstly, the logistic model allows an interpretation of the regression coefficients in terms of increasing the probability of a positive outcome. Furthermore, it is very efficient from a computational point of view. However, when the number of predictors increases, multicollinearity issues and the course of dimensionality limit the possibility of using logistic regressions. In this case, shrinkage methods such as Ridge, Lasso, and Elastic net can come to the aid.

1.F.2 Regularized logistic regression – shrinkage

Shrinkage methods act similarly to subset selection methods because they reduce the number of initial predictors to a subset that has the highest predictive power while shrinking or setting all the other coefficients to zero.

Shrinkage methods control for over-fitting by adding a penalization term $E_\beta(\beta)$ the loss function $\mathcal{L}$. The optimization criteria take the form of:

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \{\mathcal{L}(\beta) + \lambda E_\beta(\beta)\} \tag{1.6}$$

where λ is the regularization coefficient that controls the importance of regularization; this parameter must be estimated through cross-validation (J. Friedman et al., 2001).

The form of the regularization term $E_\beta(\beta)$ determines the regularized models. The Ridge regression imposes the l_2-penalty to the coefficients such that $E_\beta(\beta) =|| \beta ||_2= \sum_{j=1}^{p} \beta_j^2$; the Lasso regression imposes the l_1 norm such that $E_\beta(\beta) =|| \beta ||_1= \sum_{j=1}^{p} \mid \beta_j \mid$. These two methods have been intensively used, and they differ essentially in the shrinkage effects they have on the parameters: the ridge regression shrink less important parameter towards zero while never setting them exactly to zero, and the lasso method, instead, allows the parameter to be exactly equal to zero thus implementing real variable selection.

The Elastic net penalty is a convex combination of the lasso and ridge penalties and takes the following form:

$$(1-\alpha) \mid\mid \beta \mid\mid_1 + \alpha \mid\mid \beta \mid\mid_2 .$$

The elastic net solves the inner problem of Lasso and Ridge parameter dimensionality but requires higher computational power. When $\alpha = 0$, the elastic net is equal to the ridge regression. When $\alpha = 1$, the net Elastic net penalty reduces to the lasso penalty.

1.F.3 Regression trees and random forests

Decision tree algorithms are non-parametric models that recursively segment the prediction space X into non-overlapping regions. This procedure gives rise to the "tree" structure that gives the algorithm its name. The partitioning approach divides the data into smaller subsets until the algorithm determines that the data within the subsets are sufficiently homogeneous.

The model equation for decision trees is

$$\hat{y} = \sum_{i=1}^{M} c_m \cdot 1_{(X \in R_m)};$$

where $R_1, \ldots, R_M$ are disjoint partitions of the predictor space as resulting in the terminal nodes and c_m is a constant. The optimization criterion, in this case, is the average log-likelihood.

Various methods can calculate homogeneity within subsets. The Gini entropy for classification problems is the most widely used for classification problems and the sum of square errors for regression problems. Moreover, the algorithm requires tuning other hyper-parameters, namely the maximum depth, the minimum number of samples at the leaf nodes, and the maximum number of features to consider for splitting.

Tree-based methods often yield good predictions on the training set but are likely to overfit the data, resulting in dire out-of-sample predictions. One way to solve this problem is by relying on ensemble learning. Ensemble learning is a learning paradigm that, instead of trying to learn one super-accurate model, focuses on training a large number of low-accuracy models and then combining the predictions given by those weak models to obtain a high-accuracy meta-model.

The Random Forest is an ensemble method built upon decision trees (Breiman, 2001). The Random Forest optimization procedures first require randomly selecting an independent subsample (bootstrap) of the training sample. For each random sub-sample $b = 1, \ldots, B$, it builds a depth tree and estimates a prediction of the test sample. Finally, it averages over B to obtain a low-variance statistical learning model. The functional form looks as follows:

$$\hat{y}_{avg} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}^b$$

So far, the Random Forest procedure follows the ”bagging” procedure. In addition to bagging, the Random Forest algorithm not only selects a random sub-sample of the training set but also performs a random selection within the predictor matrix, choosing at each iteration a subset of the predictor set, $\bar{X} \subseteq X$, of size m. In the presence of a strong predictor, this procedure allows other less significant predictors to be selected, reducing the correlation among trees and the variance of the learning

algorithm. The parameter m is a hyper-parameter that we tuned with combined cross-validation and grid search.

Like the Random Forest, Gradient Boosting is an ensemble of decision trees (for an in-depth explanation, see J. H. Friedman, 2001). However, contrarily to Random Forest, which builds each tree independently, the Gradient Boosting procedure builds the tree sequentially. Each new tree helps correct the error from the previous by modifying the weight of the misclassified observations. The AdaBoost algorithm gives the simple version of boosting algorithm. The AdaBoost starts by training a simple decision tree where each observation has an equal weight. After evaluating the first prediction error, the algorithm increases the problematic observations' importance and lowers the easy ones' weights. Thus, the second tree is grown on the weighted data. The idea is then to learn and improve from the predictions of the previous tree. This procedure runs for a specified number of iterations. The final prediction is then a weighted average of all the predictions of these successive iterations. Extreme Gradient Boosting modifies this procedure by calculating gradients in the loss function. Thus it can handle any differentiable loss function.

1.F.4 Artificial neural networks

Artificial neural networks (ANN) belong to the algorithmic class of the so-called "black box" methods. An ANN models the relationship between the set of predictors and the output in a way that mirrors the process of reaction of the biological brain to external sensory input. Like the human brain, the ANN structure involves a network of interconnected artificial neurons that transform the initial input signal into an output signal. ANNs have shown outstanding performance in image recognition and detection tasks. They can adapt to classification or numeric prediction problems. Their flexible structure allows for modelling more complex patterns than nearly any algorithm. Despite these significant advantages, ANN applications are scarce in social sciences, mainly due to the impossibility of interpreting the parameters of the optimized structure. In addition, they require substantial computational and data requirements.

The typical optimization procedure of ANNs is that of backpropagation. In its more general form, the backpropagation algorithm iterates several times in two sequential processes. The completion of this cycle is called an epoch. Each

epoch consists of a forward phase and a backward phase. In the forward phase, the input features transmit through the network, transforming themselves through the combination of activation functions and weights until they reach the output layer, where a prediction or output signal is produced. All predictions are compared to the true target value in the training data to estimate a cost given a loss function. The backward phase consists of adjusting the connection weights by taking the loss derivative with respect to each connection weight; this technique is called gradient descent.

Over time, the complex training procedure of an ANN will reduce the total error of the network, but it is likely to overfit the data. Resulting in bad out-of-sample performance. Various methods have been proposed to control overfitting in ANN. In this analysis, we use the skip connections residual connection. The skip connection allows the construction of regularized deep networks by skipping one layer in the network and feeding the output of one layer as the input to the next layers.

1.G Optimal hyper-parameters

As explained in the main text, optimal hyper-parameters were obtained by cross-validation in the training datasets (one for each sex). The resulting parameters are shown in Tables 1.G.1 and 1.G.2.

Table 1.G.1: Optimal hyper-parameters selected through stratified 10-folds cross-validation. Female sample

panel A: Ridge				
Hyper-parameter	Baseline	Cluster	Features	Unstructured
λ	0.54	0.096	0.06	0.002
α	0	0	0	0
panel B: Lasso				
λ	0.37	0.083	0.183	0.147
α	1	1	1	1
panel C: Elastic Net				
λ	0.7	0.118	0.431	0.025
α	0.5	0.46	0.46	1
panel D: Gradient Boosting				
Max depth	11	11	7	9
Min child weight	5	18	22	24
Max delta step	7	7	6	1
N estimators	50	55	67	74
Learning Rate	0.1	0.7	0.1	0.1
panel E: Neural Network				
N. Epochs	50	3	3	4
Learning rate	0.01	0.01	0.01	0.01
N. Neurons	20	20-30	100-250	100-200
Batch size	1024	1024	1024	1024
Activ.function	”sigmoid”	”sigmoid”	”sigmoid”	”sigmoid”
N. Hidden Layers	3	4	5	5
N. Skip connection	1	2	2	2

Table 1.G.2: Optimal hyper-parameters selected through stratified 10-fold cross-validation. Male sample

panel A: Ridge				
Hyper-parameter	Baseline	Cluster	Features	Unstructured
λ	0.014	0.013	0.019	0.026
α	0	0	0	0
threshold	0.27	0.249	0.259	0.268
panel B: Lasso				
λ	0.025	0.023	0.023	0.683
α	1	1	1	1
panel C: Elastic Net				
λ	0.012	0.135	0.008	0.05
α	0.55	0.37	0.1	0.19
panel D: Gradient Boosting				
Max depth	17	8	8	10
Min child weight	9	18	14	14
Max delta step	7	1	7	5
N estimators	60	94	55	65
Learning Rate	0.1	0.2	0.2	0.1
panel E: Neural Network				
N. Epochs	50	5	3	5
Learning rate	0.01	0.01	0.01	0.01
N. Neurons	20	20-30	200-250	150-300
Batch size	1024	1024	1024	2048
Activ.function	”sigmoid”	”sigmoid”	”sigmoid”	”sigmoid”
N. Hidden Layers	3	5	4	5
N. Skip connection	1	2	2	2

1.H Predictive performance

We report here additional predictive performance metrics for the sequence features' predictor set.

Accuracy is a metric that measures the overall correctness of the model's predictions. It is the ratio of the correctly predicted cases (both true positives and true negatives) to the total number of cases. Accuracy provides an overall assessment of the model's performance but can be misleading when the dataset is imbalanced.

ROC-AUC, i.e. Receiver Operating Characteristic Area Under the Curve (ROC-AUC), is a performance metric for binary classification models. It measures the ability of a model to distinguish between positive and negative classes across various classification thresholds. The ROC curve plots the true positive rate (recall) against the false positive rate (1 - specificity) at different threshold values. The AUC represents the area under the ROC curve and provides a single-value summary of the model's performance.

Table 1.H.1: Predictive performance metrics for the sequence features predictor set. Female sample

model	recall	accuracy	roc-auc	pr-auc	precision
logistic	0.606	0.663	0.725	0.720	0.675
ridge	0.610	0.663	0.724	0.719	0.675
lasso	0.611	0.664	0.725	0.720	0.676
elnet	0.609	0.662	0.725	0.720	0.674
XGBoost	0.655	0.690	0.757	0.768	0.697
ANN	0.565	0.651	0.704	0.689	0.665

Table 1.H.2: Predictive performance metrics for the sequence features predictor set. Male sample

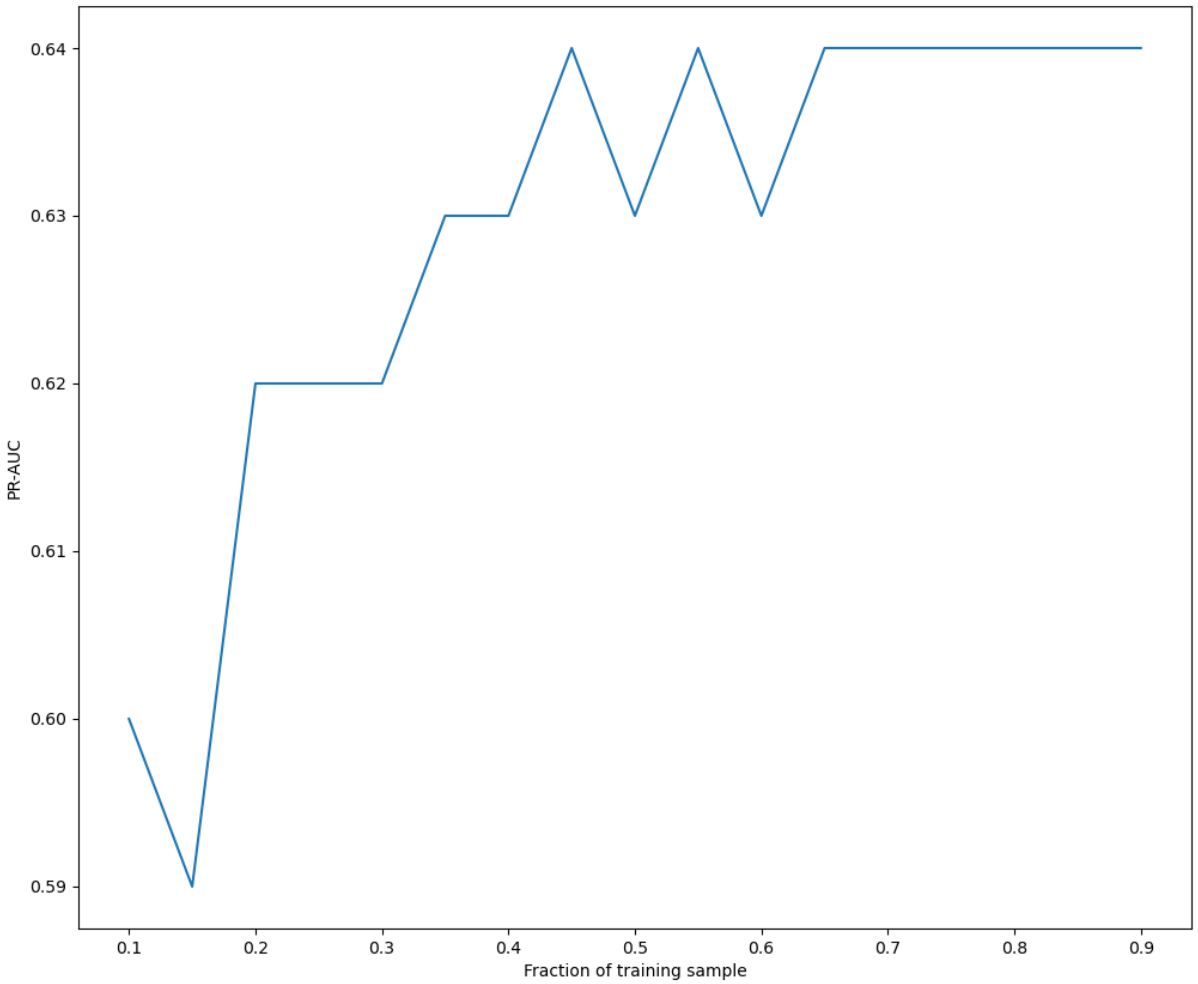
model	recall	accuracy	roc-auc	pr-auc	precision
Logistic	0.363	0.753	0.739	0.581	0.650
Ridge	0.361	0.752	0.739	0.580	0.648
Lasso	0.377	0.751	0.740	0.581	0.634
Elnet	0.369	0.750	0.740	0.581	0.637
XGBoost	0.408	0.778	0.772	0.647	0.721
ANN	0.389	0.748	0.732	0.563	0.596

1.I Sample size independence test

This section presents the sensitivity of the performance of the best-performing model (Gradient Boosting) to the training set's sample size.

First, as the performance of any model may vary when applied to different datasets or populations, this test helps assess the generalizability of the model's performance across different sample sizes. Second, it provides insights into the stability and robustness of the model's performance. Results shown in Figure 1.I.1 suggest that, with 20% of the training sample already, the metric remains stable at around 0.62-0.64, indicating good reliability.

Figure 1.I.1: Test PR-AUC and training data dimensionality. Gradient Boosting model.



1.J Robustness to alternative depression measurement

Figure 1.J.1 shows the sensitivity of models' prediction capacity metrics to changing the depression threshold, i.e., the number of symptoms an individual must have to classify depression to 3, 4, 5, and 6. Increasing the depression threshold deteriorates the predictive accuracy. Indeed, higher thresholds imply a smaller number of depressed individuals. With fewer depressed individuals, the models have limited exposure to the patterns and characteristics of depression. Models do not have enough instances to learn the distinguishing features and nuances of depression. Consequently, the models have difficulty generalizing well to new, unseen instances of depression. However, the Gradient Boosting model remains the best-performing algorithm, achieving the best performance metrics across all threshold values.

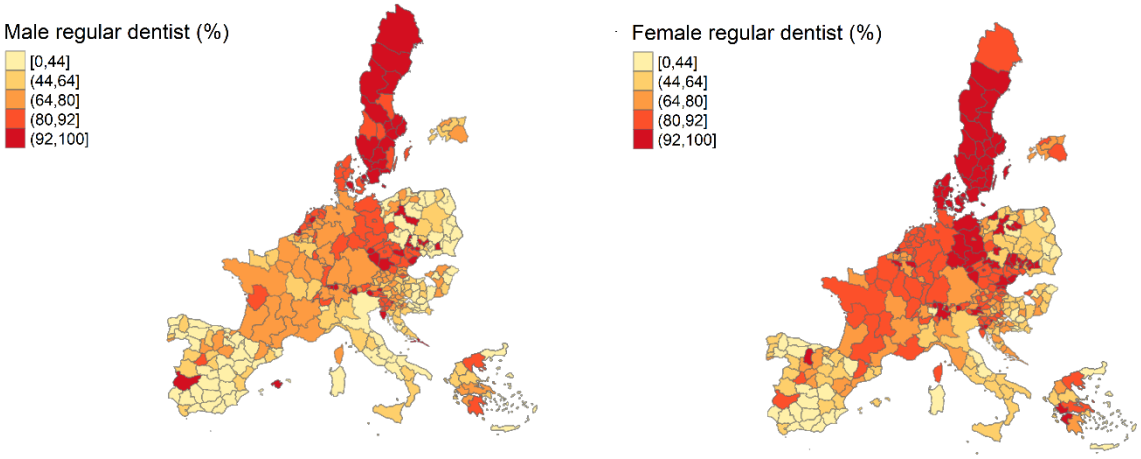
Figure 1.J.1: PR-AUC in the test sample for increasing EURO-D depression discrimination thresholds

1.K Mapping key predictors

This section presents descriptive maps of our two key predictors: regular dentist attendance and general life entropy. For the regular dentist, we also map the prevalence of individuals selecting each of the following reasons for not attending regular dentist: 1. Not affordable, 2. Not enough information about this type of care, 3. Not usual to get this type of care, 4. No place to receive this type of care

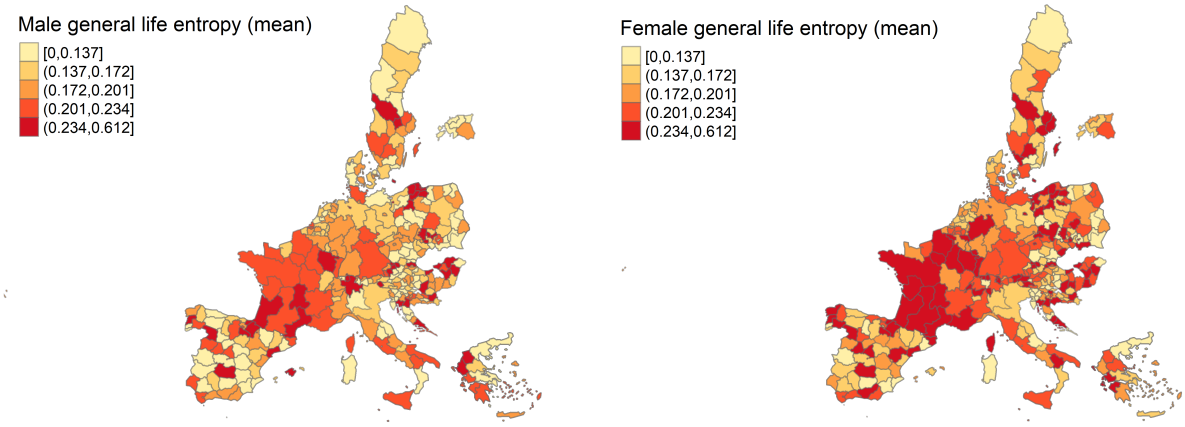
close to home, 5. Other reasons

Figure 1.K.1: Prevalence of individuals reporting regular dental visits, male (left) and female (right).



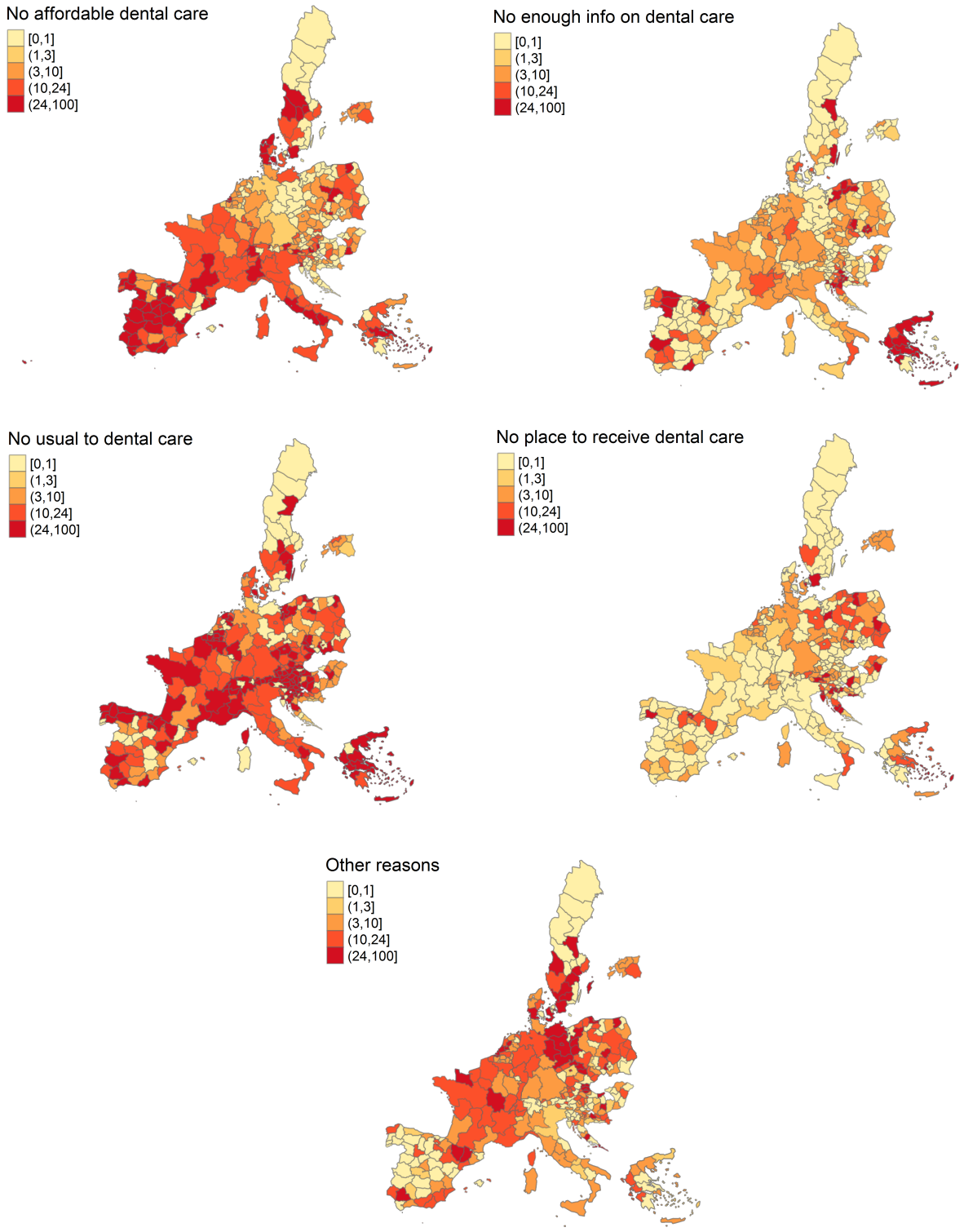
Note: each colour represents ventiles of the pooled distribution. Lighter colours are bottom ventiles.

Figure 1.K.2: Mean entropy in the general life sequence, male (left) and female (right).



Note: see Figure 1.K.1

Figure 1.K.3: Prevalence of individuals selecting a given reason for not attending dental care regularly



Note: Each respondent can select multiple responses. In the maps, each colour represents a ventiles of the pooled distribution. Numbers in square brackets represent percentages (%). Lighter colours are bottom ventiles.

Chapter 2

The Old Folks at Home: Parental Retirement and Adult Children Well-being

2.1 Introduction

Population aging is a major challenge faced by many OECD countries, including the United Kingdom. With life expectancy on the rise, UK projections suggest that 24% of the population will be aged 65 and over by 2043 (Lewis, 2021). While greater longevity is one of the benefits of development, it puts significant pressure on modern welfare states, and in particular on pension systems that rely on current contributions to fund the pension benefits of those who are currently retired (see Lewis et al., 2021). In response to this fiscal imbalance, governments worldwide have started to implement reforms increasing the age at which workers are eligible for State Pensions.

The postponement of statutory retirement is an effective tool to improve the sustainability of pension systems amidst population aging. However, this intervention raises a number of concerns about the effect of later retirement both on the individuals concerned (Clark and Zhu, 2024; Zhu and He, 2015) and their family members (Atalay & Zhu, 2018). We here examine potential inter-generational spillover effects between older parents and their adult children, as parental retirement will likely affect the transfer of both resources and non-pecuniary support between parents and children. In this spirit, we here ask first whether parental retirement affects the well-being of adult children, and if so why?

Our research draws upon and contributes to several strands of the existing literature. The first relates to the informal exchanges between parents and adult children, and we consider both time and financial transfers (OECD, 2012). Both of these transfers vary over the life course, and therefore may well change at the time of parental retirement (see Cox, 1987 and Coe and Zamarro, 2011). There is, however, only little research on the impact of retirement on adult children's well-being, and the mediating role of these transfers. If parental retirement reduces financial transfers but increases time transfers, the net effect on adult children's well-being will be ambiguous.

First, as parents age, they may experience health issues, disabilities, and reduced financial resources. Retirement may exacerbate these physical (Filomena & Picchio, 2023) and financial challenges (see Mazzonna and Peracchi, 2012, Cribb et al., 2016 and Gorry et al., 2018). The greater demands of older parents on their adult

children, either through financial transfers or informal caregiving (Van Houtven et al., 2013, Van den Berg et al., 2014), may in turn negatively affect adult children's well-being.

At the same time, grandparents play a crucial caring role in many OECD countries. With retirement, they likely have more time to engage with their grandchildren and provide more-substantial care. In the UK, around 40% of grandparents regularly provide childcare for their grandchildren, helping working parents save an estimated £7 billion in childcare costs (see Buchanan and Rotkirch, 2018). Grandmothers provide most of this care, from occasional babysitting to formal arrangements of regularly caring for their grandchildren.

A second recent literature has explored the impact of parental retirement on adult children's outcomes, focusing primarily on fertility and labour-force participation. For instance, Eibich and Siedler, 2020 examines adult daughters' fertility around the time of their parents' retirement in Germany, using the early-retirement age threshold as an exogenous cutoff. Similarly, Ilciukas, 2023 analyse the exogenous delay in older mothers' retirement in the Netherlands, uncovering a substantial negative effect on adult daughters' fertility. Focusing on labour-market outcomes, Kaufmann et al., 2023 find that an increase in grandmothers' working hours produces lower working hours for adult daughters with young children. However, in China Wu and Gao, 2020 shown that adult children's annual labour supply falls following parental retirement. To date, this literature has not considered adult children's well-being outcomes following parental retirement, and we here aim to fill this critical gap.

We will appeal to two causal identification strategies applied to panel data from the United Kingdom: the British Household Panel Survey (BHPS) and its successor, Understanding Society (UKHLS). We construct child-parent dyads, linking socio-economic information on adult children to their older parents' retirement transition.

The first identification strategy exploits the discontinuous increase in the probability of retiring at the State Pension Age in a Fuzzy Regression Discontinuity design to identify the direct and spillover effects on parents and adult children's well-being. In the second identification strategy, we leverage two UK Pension Acts, from 1995 and 2011, in a difference-in-differences design (DiD) to estimate the effect of an unexpected increase in the parental State Pension Age on their children's

well-being.

The Fuzzy RDD estimates reveal a positive and significant impact of maternal retirement on adult children's life and income satisfaction but no effect on mental health. There is no effect of paternal retirement.

Heterogeneity analyses help shed light on potential mechanisms. It also reveals adult children and father sub-groups where the causal impact of paternal retirement turns statistically significant.

We consider moderation by first assessing the presence of grandchildren at the time of retirement and then determining the age of the grandchildren. For maternal retirement, the well-being benefits for adult children are highest when grandchildren are in the 5-11 age range. Further stratification reveals larger rises in satisfaction for adult children with lower incomes and who lived near their mothers in the years pre-retirement. On the contrary, paternal retirement affects more negatively low-income adult children. Lastly, we consider the retired mother's and father's marital status and health. Retirement-positive spillovers are larger for elder mothers who are not married (i.e. separated, divorced, or widowed) and have never been hospitalised in the years pre-retirement. On the contrary, we observe larger negative retirement spillovers for not-married elder fathers.

This battery of moderation results is consistent with maternal retirement causally affecting their adult children's well-being via time transfers, with grandmothers having more time available to provide child care to their grandchildren, reducing their adult children's child-care costs and increasing their well-being. It also reveals opposite retirement spillover effects retirement between elder mothers and elder fathers on their adult child mental health.

Regarding the second identification strategy, the reform delayed the retirement of the directly-affected older parents. There was no reform effect via mothers on their adult children's well-being, but significant positive effects of the reform via fathers on their sons' life and income satisfaction. The heterogeneity analyses again show that the effect is concentrated among adult children with lower incomes and (to a lesser extent) adult sons still living with their fathers in the years around retirement. This second set of results suggests that infra-family financial transfers play a role, with later paternal retirement increasing adult children's additional financial resources.

The remainder of this article is organised as follows. Section 2.2 describes the institutional setting and the potential relationships between parental retirement and adult child well-being. Section 2.3 presents the data and the key variables of interest. Section 2.4 outlines the empirical models, and Section 2.5 describes the estimation results. Last, Section 2.6 concludes.

2.2 Background

2.2.1 The UK Pensions System and Pension Reform

The UK State Pension Age (SPA) is the earliest age at which workers can claim the public pension. In 1948, this was set at 60 for women and 65 for men (having previously been 65 for both sexes), figures which remained unchanged until April 2010. Faced with an ageing population and increased life expectancy, concerns were raised about the sustainability of the pension system. As a result, the UK government implemented significant pension reforms in 1995, including introducing a single-tier flat-rate state pension and a programmed rise in the SPA to be started in 2010.

The central point of this 1995 reform was the phased introduction over ten years of equal State Pension Ages for men and women. The SPA for women born after March 1950 increased gradually starting from April 2010. The 2011 Pensions Act then modified this initial timetable, legislating a more rapid increase in women's State Pension age to 65 between April 2016 and November 2018 instead of the initially planned April 2020. The same act also established that from December 2018, the State Pension age for men and women born after November 1953 would be increased to 66 by October 2020. Figure 1 illustrates the planned date of reaching the State Pension Age as a function of women's and men's cohorts under these two pension Acts.

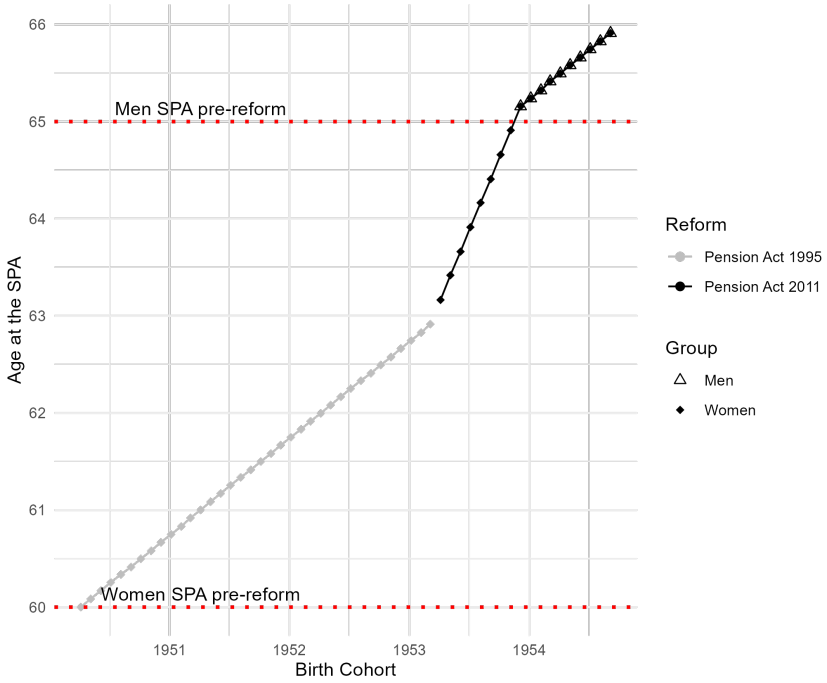
In addition to the State Pension, many UK workers have occupational and private pension funds, which provide additional income after retirement. However, the State Pension remains a significant source of income for many retirees, particularly those without other pension arrangements. To receive the full basic State Pension,

The basic State Pension was designed to provide a minimum level of income for all retirees, while the earnings-related state pension, known as the State Second Pension (SERPS), provided additional income for those with moderate to high earnings.

individuals must have 30 qualifying years of National Insurance contributions or credits. The level of the basic State Pension depends on the contributions that the individual made throughout their working life, with a minimum level of £141.85 per week for those who meet the eligibility criteria.

Deferring receipt of the State Pension allows individuals to receive an increased entitlement, which depends on the number of weeks deferred. For every five weeks of deferral, the level of the State Pension rises by 1% up to a maximum of 10.4% after one year of deferral (Cribb et al., 2016). These higher entitlements are designed to encourage deferred receipt and continued employment, thereby contributing to the economy and reducing the burden on the State Pension system. Even though the deferral rate seems generous, in practice, only relatively few individuals put off their State Pension receipt: in 2010 English Longitudinal Study of Ageing data, only 5% of those aged between the SPA and 75 in 2008-09 had chosen to defer their State Pension (Crawford and Tetlow, 2008 and Cribb et al., 2016).

Figure 2.2.1: Women's and Men State Pension Age under the 1995 and 2011 Pension Acts



Note: The Y-axis lists the State Pension Ages as legislated by the 1995 and 2011 Pension Acts. The X-axis shows the birth cohorts of women and men affected by the reforms. The men's and women's lines overlap after State Pension Age of 65, as the 2011 Pension Act affected both sexes equally. The horizontal dotted lines indicate the pre-reform SPA for women and men. *Source*: Data from Gov.UK State Pension Age timetable.

2.2.2 Theoretical Mechanisms

The impact of parental retirement on adult children's well-being is theoretically ambiguous. On the one hand, parental retirement can benefit adult children, as it relaxes time constraints and can increase both leisure time and hours of work. On the other hand, it may also come with adverse effects via an increased demand for informal care and lower net financial transfers from working parents. This section illustrates four potential channels between parental retirement and adult children's well-being.

First, retirement may well directly affect the intensive margin of time transfers between parents and adult children. Based on the literature on retirement's physical- and mental-health consequences (see *e.g.* Dave et al., 2008 and Charles, 2004), the first kind of time transfer may run from adult children to their parents via informal care. Evidence in this field is somewhat mixed, with some UK results finding adverse effects of retirement on health and mental well-being (Carrino et al., 2020, and Fé and Hollingsworth, 2016). Any rise in informal care and support from adult children following parental retirement may reduce the satisfaction of the former (see Lacey et al., 2019).

Time transfers may also flow in the opposite direction, with grandparents' retirement increasing their availability to provide childcare for their grandchildren, as discussed in Eibich and Siedler, 2020. This grand-parental childcare will likely positively impact adult children's well-being, especially for adult daughters who often face a "child penalty" regarding their career prospects and earnings. There is recent evidence in Kaufmann et al., 2023 that greater childcare by grandmothers reduces this child penalty and increases the labour supply of adult daughters, and in addition produces better educational outcomes for the grandchildren. The availability and quality of grandparental childcare may well vary, however, according to the geographical distance between the households, grandparents' health, and family structure.

The third channel is direct financial transfers. Retirement almost certainly has financial consequences (Cribb et al., 2022) and may lead to greater financial support from adult children to their parents (or less support from newly-retired parents to their adult children), reducing the adult children's well-being.

Last, financial transfers may also be indirect. Adult children who are parents

may save money on childcare costs by receiving grandparental care, again increasing their well-being.

The net effect of these four channels on adult children's well-being is ambiguous and likely varies between different types of adult children. From a policy perspective, it seems important to understand how the changes in the State Pension system will affect the outcomes of not only retirees but also their families.

2.3 Data

Our analysis uses panel data from the British Household Panel Survey (BHPS waves 6-18) and the UK Household Longitudinal Study (UKHLS, also known as Understanding Society, waves 1-12), covering the period from 1996 to 2022 (University of Essex, Institute for Social and Economic Research, 2023). The BHPS began in 1991 with a sample of 5,000 households and was later expanded to include additional households from Scotland, Wales, and Northern Ireland. The ongoing Understanding Society survey started in 2008 with approximately 40,000 households. These two surveys include many of the same questions, allowing harmonised samples to be constructed.

Both surveys interview all adult members (16+) in participating households. Survey respondents who leave the initial household, for instance children who move out of their parent's home or parents who separate, are followed and their new household becomes part of the panel. This survey design allows us to link data on adult children and their parents over time, even when they live in different households.

Our sample is constructed by linking each child in a household to their biological mother and father. If the child lives with a stepfather/mother or a father/mother-in-law, we include them in the sample. When adult children in the original sample start cohabitating with a partner, the new partner inherits this information regarding the biological mother and father. This produces an unbalanced panel dataset. Appendix 2.B contains more information on the initial sample composition and the attrition analysis.

This sample-selection procedure may result in co-residence bias if the characteristics of the adult children retained in the sample differ significantly from those

we never observe living with their parents. To address this issue, we run a simple descriptive analysis to show how different our samples are (from the BHPS for the RDD analysis, and from the UKHLS for the Difference-in-Differences) from the full sample of respondents in the same age range: the results appear in Appendix 2.B. The two samples are statistically different from the full sample of respondents in terms of some demographic characteristics. However, we note that the effects of retirement on a battery of parental outcomes that we estimate match in sign and size those found in other research carried out on the full sample of parents (see, e.g., Della Giusta and Longhi, 2021)

We use different samples in the two distinct causal identification settings. The first uses the State Pension Age as an exogenous cutoff point in a Fuzzy Regression Discontinuity Design. Here, we analyse data from the BHPS, as the method requires a fixed SPA cutoff. The second exploits the pension reform that took place in April 2010. This gradually raised the SPA from 60 to 66 over a period of ten years for women born after April 1950, and starting in 2018 raised the SPA for men born after 1953. This second approach applies Difference-in-Differences to UKHLS data to evaluate the impact of delayed parental retirement on the well-being of adult children.

2.3.1 Adult children's outcomes

This paper evaluates how parental retirement affects adult children's well-being. Well-being is a multi-dimensional concept, and we here consider three different types of outcome.

The first is a measure of psychological distress. This is derived from 12 questions in the General Health Questionnaire (GHQ), in which respondents indicate the extent of their agreement on a four-point scale (Appendix 2.A shows the full questionnaire). Some of the questions are negatively couched while others are positively so. After recoding the negative questions, we add up the individual's 12 responses to produce a 0-36 scale, where higher numbers refer to better outcomes. The GHQ appears in all BHPS and UKHLS waves. However, we will only use BHPS Waves 6-10 and 12-18, as these also include the two satisfaction measures described below.

The second and third well-being variables refer to self-reported satisfaction, and appear in BHPS Wave 6 (1996) onwards. All of the satisfaction questions are

answered on a 1-7 Likert scale, where one means completely dissatisfied, seven is completely satisfied, and four is neutral.

The first variable is overall life satisfaction, which has been very widely analysed across the Social Sciences. The second is satisfaction with income. This variable is important because, as suggested above, parental retirement can have opposing effects on the adult child's financial status: increasing transfers of money from adult children to their now-retired parents but with time transfers reducing the child-care-related costs of the adult children via grandparental childcare.

2.3.2 Retirement and pension eligibility

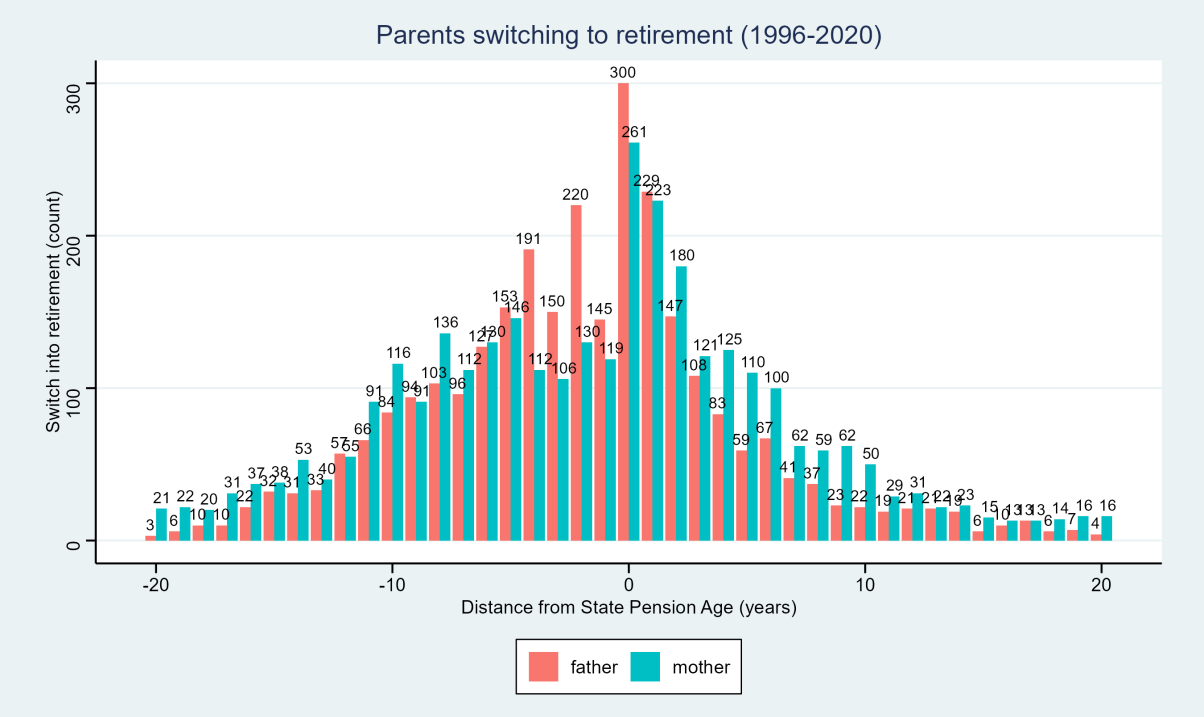
The treatment variable in this analysis is parental retirement. In the main analysis, we consider older parents who self-reported being retired at the date of the interview as treated. We assume that retirement is an absorbing state, so that once individuals retire they will remain so. In the sensitivity analysis, we will change the definition of retirement in two ways. First, we also consider parents to be retired if they did not self-report this status but were unemployed and not actively looking for work in the month prior to the interview. Second, we consider as retired those who receive pension income. The results of these analyses appear in Appendix 2.D.

Figure 2.3.1 plots the number of parents who move into retirement as a function of their distance from their State Pension Age. Over the entire analysis period, from 1996 to 2019, we have information on 1812 mothers and 1190 fathers who enter retirement. For both sexes, there is a notable spike around the mandatory SPA. However, a non-negligible proportion of parents retire before the SPA.

The UK State Pension eligibility Age changed significantly in the period covered by our data. Up to April 2010, the SPA for men was 65 and 60 for women. These figures increased for women born after April 1950 starting in April 2010, and for men born after December 1953 starting in December 2018. The effect of this reform on retirement can be seen in Figure 2.3.2. Compared to the untreated cohort (with retirement ages at 65 and 60 for men and women), the treated cohorts (with higher SPAs) have a significantly lower retirement probability at the ages of 65 and 60.

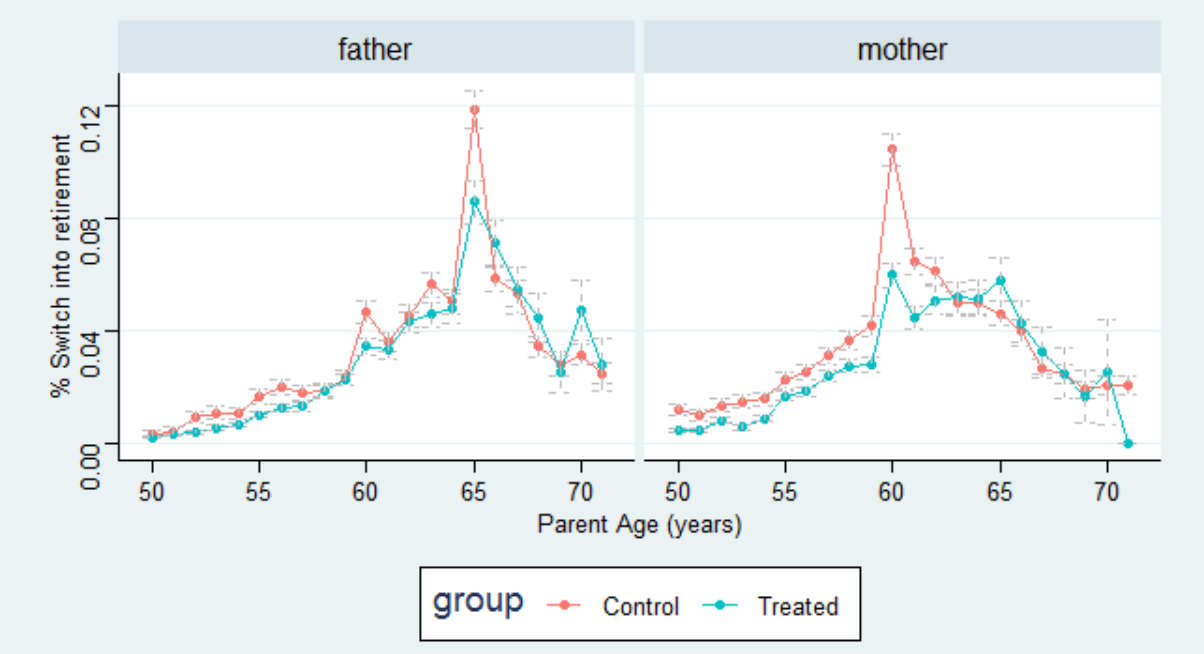
Finally, Figure 2.3.3 shows the shares of fathers and mothers who are above the SPA thresholds by their adult children's ages. As was found by Eibich and Siedler, 2020 in German data, under 20% of parents attain the State Pension Age

Figure 2.3.1: The number of parents retiring as a function of the distance to the SPA in years and cohort.



Source: Pooled BHPS and UKHLS sample (1996-2020).

Figure 2.3.2: Percentage of parents retiring as a function of the distance to the SPA in years and treatment group.

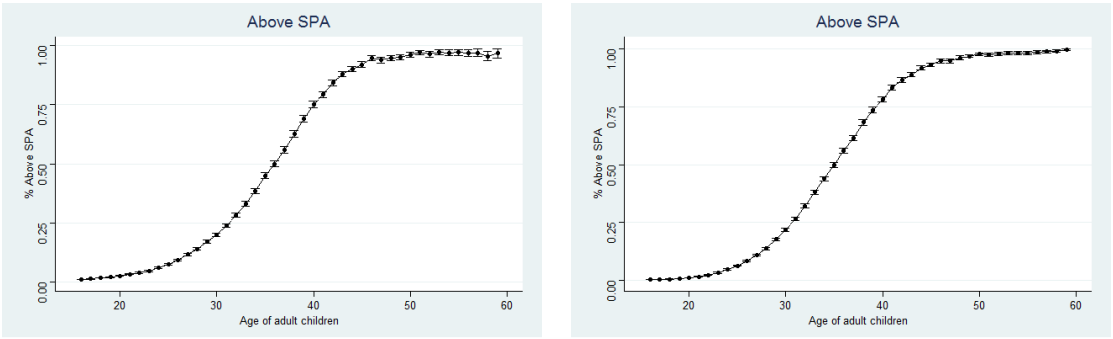


Source: Pooled BHPS and UKHLS sample (1996-2020).

threshold before their child's 25^{th} birthday, while almost all parents have attained this threshold by the time their adult child turns 45. We consequently apply the same sample restriction as in Eibich and Siedler, 2020, and only consider adult children aged 20-45 (as parental retirement is only rare outside of this range). We will test whether our results are sensitive to this restriction: the results and associated discussion appear in Appendix 2.E.

Figure 2.3.3: The proportion of fathers (left) and mothers (right) above the State Pension Age (at ages 60 and 65, respectively) as a function of their adult child's age

(a) (b)



Source: our elaboration on UKHLS and BHPS pooled sample (1996-2020)

2.4 Empirical Approach

Parental retirement is a choice, and is related to both parental and adult child characteristics, including their well-being. Older parents may choose to retire in order to help their children if the latter are unwell, or to help with childcare and household chores, and provide support in general. We tackle this endogeneity via two identification approaches. Both of these rely on the individual's eligibility for the State Pension, which in the UK is likely to represent a major comeponent of their retirement income (see Cribb et al., 2022).

As in other contributions (Coe and Zamarro, 2011, Gorry et al., 2018, Eibich and Siedler, 2020), the first of these exploits the age threshold for pension eligibility (up to 2010, at age 60 for mothers and 65 for fathers) as an exogenous cutoff in a fuzzy regression discontinuity design. As State Pension eligibility is conditional on attaining these ages, moving from being under to over the age threshold should be associated with a considerable discontinuity in the probability of retiring.

The second identification strategy exploits the 1995 and 2011 UK Pension Acts

that, starting in April 2010, gradually increased the State Pension Age from 60 to 66 over a ten-year period, initially only for women and then, starting in 2018, for both sexes. These reforms affected women born after March 1950 and men born after March 1953. We will carry out a difference-in-differences analysis to compare the well-being of adult children of parents subject to different State Pension Ages.

2.4.1 The Fuzzy Regression Discontinuity Design

There are two main requirements for the causal interpretation of the coefficients in a fuzzy RD design. First, being above or below the State Pension age should not directly affect the well-being of adult children. While parental age in general may well be related to adult children's well-being, it does not seem likely that there should be a discontinuity in this relationship exactly at the SPA. This assumption is then likely to hold conditional on a continuous trend in parental age. The second requirement is that parents cannot manipulate whether they are above or below the threshold. With age in months being the running variable for the threshold, this assumption should be held by construction. To check, we run the density continuity tests of the assignment variable proposed by Cattaneo et al., 2019: see Section 2.C in the Appendix.

Assuming that these two requirements are met, we can estimate the causal effect of parental retirement on the three adult child well-being outcomes. The regression model is:

$$r_{it} = \alpha + g_1(\text{age}_{it}) + h_1(\text{page}_{it}) + \pi D_{it} + \omega_i + \tau_t + \nu_{it} \quad \text{first stage} \tag{2.1}$$

$$y_{it} = \beta + g_2(\text{age}_{it}) + h_2(\text{page}_{it}) + \lambda r_{it} + \xi_i + \kappa_t + \epsilon_{it} \quad \text{second stage} \tag{2.2}$$

In these equations, y_{it} represents the well-being outcome of adult child i at interview time t. The variable age_{it} denotes the age of the adult child in months, while page_{it} represents the parent's age, centered at the cutoff, also in months. The variable r_i indicates the retirement status of the parent at time t.

In the model, α_i and ω_i are fixed effects for each adult child, and τ_t and κ_t are fixed effects for year and month, respectively, to account for secular and seasonal trends. The terms ϵ_{it} and ν_{it} represent the idiosyncratic errors in the second and

first stages, respectively. The functions $g(.)$ and $h(.)$ are parametric functions of the child's age (age_{it}) and the parent's age (page_{it}), respectively.

The dummy variable D_{it} indicates whether the parent of adult child i is above the state pension age at time t. In the first stage, the parameter π quantifies the impact of the parent crossing the SPA cutoff on their retirement probability. In the second stage, the parameter λ rflects the treatment effect of parental retirement on the well-being of the adult child.

This model is estimated using two-stage least squares (2SLS). We apply a bandwidth of 10 years for both mothers and fathers (*i.e.* we only include observations with mothers aged 50 to 70 and fathers aged 55 to 75), and consider a quadratic trend for both parental and adult child age in our main specification. Heteroskedastic robust standard errors are clustered at the adult-child level. We will check whether the results hold using different parental and child age bandwidths and different functional form specifications (see Appendix 2.E).

Descriptive statistics Fuzzy RDD

The main Fuzzy RDD analysis is carried out on the sample of adult children and their spouses who are aged 20–45 years matched to their parents who are within a band of ± 10 years around the State Pension age. The sample here is restricted to the years before the UK pension reform so that the SPA is fixed. We drop children whose parents never worked (385 mothers and 203 fathers) or died within the age bandwidth around the State Pension cutoff (58 mothers and 96 fathers).

This selection procedure yields a sample of 16984 observations in the mother sample and 13450 observations in the father sample. These cover 3518 adult children and 1622 of their mothers and 1232 of their fathers. Table 2.4.1 presents the descriptive statistics, divided into the adult-child-father and adult-child-mother samples. The difference between the two samples reflects observations on mothers only and fathers only, as opposed to both at the same time. As such, some adult children appear in one of the two samples but not the other.

Table 2.4.1: RDD Sample Descriptive Statistics

	Mother sample				Father sample			
Variable	N	Individuals	Mean or %	SD	N	Individuals	Mean or %	SD
A. Adult child outcomes								
GHQ (0-36)	16984	3518	25.3	5.3	13032	2635	25.3	5.3
Life Satisfaction (1-7)	16984	3518	5.2	1.1	13450	2635	5.2	1.1
Income satisfaction (1-7)	16984	3518	4.5	1.4	13450	2635	4.5	1.4
A. Adult child characteristics								
Age	16984	3518	29.0	5.6	13450	2635	30.3	5.6
Year of birth	16984	3518	1974.2	6.0	13450	2635	1973.2	5.9
Married	16984	3518	32		13450	2635	39	
Age left school	16984	3518	17.3	2.2	13372	2614	17.3	2.2
Female	16984	3518	49		13450	2635	49	
Number of children	16984	3518	0.6	0.9	13450	2635	0.7	0.9
Real monthly individual income	16984	3518	1384.0	1116.7	13113	2588	1496.5	1170.8
Lives with father	16984	3518	23		13450	2635	22	
Lives with mother	16984	3518	30		13450	2635	23	
White	16984	3518	83		13367	2598	84	
B. Older Parent								
Retired	16984	1622	26		13450	1232	27	
Above SPA	16984	1622	29		13450	1232	24	
Age	16984	1622	56.8	4.9	13450	1232	61.2	4.9
Real monthly income	16984	1622	868.0	887.2	13450	1232	1614.4	1506.0

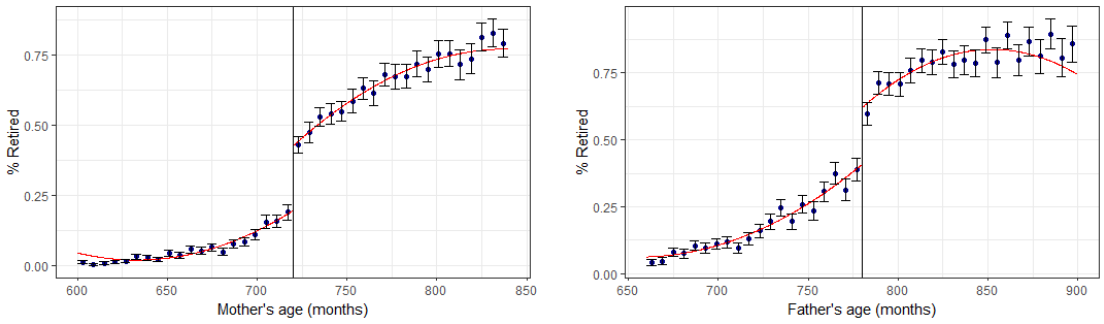
Note: Real income is derived by deflating nominal gross incomes to 2015 GBP using the CPI All Items (D7BT).

Graphical evidence

The legislated State Pension Age provides an exogenous cutoff for retirement decisions. Figure 2.4.1 reveals a sharp jump in the retirement rate around the cutoff, suggesting that many individuals react to State Pension eligibility. As such, our RDD estimates can be interpreted as valid intent-to-treat effects of parental retirement on adult children's well-being, as long as any of the other factors affecting parental retirement do not change discontinuously around this cut-off.

Figure 2.4.1: Parents' Propensity to Retire by Age: BHPS

(a) Mother (b) Father



Note: These figures plot the retirement rate in the sample of parents in a window of ten years before and after the State Pension Age. The points refer to fuzzy regression discontinuity estimates from a flexible quadratic specification using a 10-year bandwidth. Standard errors are clustered at the parent level. The sample consists of parents whose adult children or in-laws are in the main sample.

2.4.2 Difference-in-Differences

The second identification strategy exploits the change in the SPA from the UK Pension Acts of 1995 and 2011, which gradually raised the State Pension Age from 60 to 66 over the April 2010 to October 2020 period for women, and from 65 to 66 over the December 2018 to October 2020 period for men.

We can estimate the impact of the resulting delayed retirement on adult children's well-being as we have data on the well-being of otherwise similar adult children whose parents face different State Pension eligibility ages. We thus carry out a difference-in-differences analysis, as in Cribb et al., 2016, Della Giusta and Longhi, 2021 and Cribb et al., 2022, who estimated the impact of these same reforms on the retirees' own labour-market outcomes and well being. The model is as follows:

$$y_{it} = \alpha T_{it} + \lambda_i + \gamma_t + \sum_{\text{pa}=50,55}^{70,75} \delta[\text{page}_{it} == \text{pa})] + X_{it}\theta + \epsilon_{it} \tag{2.3}$$

Here the outcome of interest y_{it} for adult children i observed in period t is regressed on a dummy variable T_{it} for whether his/her parent is above or below the State Pension age and a set of parental and adult-children controls. These are the adult child's age in months and marital status, the elderly parents' and adult children's home ownership, and a dummy variable for the adult child and parent living in the same household. The dummy T_{it} is constructed by comparing the adult

child's survey interview date to their parents' State Pension eligibility. Given the nature of the reform, this is determined by both the parent's birth cohort and their age at the time of the interview.

Descriptive statistics in the DiD sample

The sample in the difference-in-differences is of mothers born between 1935 and 1965 (who are aged 50 to 70 at the time of the interview) and fathers born between 1938 and 1968 (aged 55 to 75). This allows for the comparison of adult-child outcomes across 15 parental cohorts who were unaffected by the reform (born 1935 to March 1950) to the subsequent 15 cohorts (born April 1950 to 1965) who were exposed to the gradual increases in the SPA.

The data used in this analysis come from the harmonized British Household Panel Survey (BHPS) and Understanding Society (UKHLS) surveys. The resulting panel from this combination is unbalanced. The attrition analysis for this sample appears in Appendix 2.B. As above, we exclude older mothers and fathers who never worked (2,815 mothers and 683 fathers) and those who passed away within the specified age range (115 mothers and 185 fathers). The final estimation sample encompasses 11036 adult children, 5196 older mothers, and 3768 older fathers. The descriptive statistics for this sample are listed in Table 2.4.2.

Table 2.4.2: Difference-in-differences Sample Descriptive Statistics

	Mother sample				Father sample			
Variable	N	Individuals	Mean or %	SD	N	Individuals	Mean or %	SD
A. Adult child outcomes								
GHQ (0-36)	59796	11036	24.8	5.6	35344	7447	24.9	5.5
Life satisfaction (1-7)	59796	11036	5.1	1.6	35942	7525	5.2	1.5
Income satisfaction (1-7)	59796	11036	4.5	1.8	35942	7525	4.6	1.7
A. Adult child characteristics								
Age	59796	11036	29.4	6.1	35942	7525	29.7	6.1
Year of birth	59796	11036	1981.5	8.5	35942	7525	1981.9	8.4
Married	59796	11036	29		35942	7525	32	
Female	59795	11035	51		35942	7525	51	
Number of children	59796	11036	0.6	0.9	35942	7525	0.6	0.9
Real Monthly Individual Income	59281	10987	1631.5	2805	35565	7475	1698.8	2355.8
Live with father	59796	11036	26		35942	7525	35	
Live with mother	59796	11036	36		35942	7525	34	
White	59685	11016	82		35869	7506	79	
B. Older Parent								
Retired	54216	5196	26		35942	3768	34	
Above SPA	59796	5666	26		35942	3768	27	
Age	59796	5666	58.0	5.2	35942	3768	61.5	5.1
Real Monthly Individual Income	49429	5033	1264.5	2251.16	31040	3457	2246	4187

Note: See Table 2.4.1.

2.5 Results

In the following sections, we investigate the effect of reaching the State Pension Age and its legislated postponement on the well-being and labour-market outcomes of the directly-affected parents. These results will help us to evaluate the proposed theoretical mechanisms. Retirement will be unlikely to affect the adult-child outcomes if it has no effect on parents' well-being or labour-market outcomes: these are weekly working hours, leisure-time satisfaction, subjective financial situation, subjective physical and mental health, and life satisfaction.

We then estimate the spillover effect of parental retirement on the adult children's well-being, stratifying the sample in four ways to investigate the underlying mechanisms. We first consider adult children who are responsible for one or more children under the age of 12. Second, we split the sample by the door-to-door travel distance between adult children and their parents. The third stratification refers to the adult children's income bandwidth. Last, we consider the marital status of the

older parents. All of the variables used for this stratification are measured before the parents reaching their State Pension Age.

2.5.1 Retirement Effects on Parental Labor Supply and Well-Being: BHPS Data

This section reports the effect of reaching the State Pension Age on the labour supply and well-being of elderly parents. The first-stage results in Table 2.5.1 and 2.5.2 confirm that reaching the State Pension Age does predict the probability of retirement at age 60 (for mothers) and age 65 (for fathers), with an increase of around 29% and 23% for mothers and fathers, respectively. This eligibility is a strong instrument for parental retirement status, with F-statistics that are much higher than the rule-of-thumb F-statistic of 10–12 (Staiger and Stock, 1997).

The first parental outcome is the weekly number of work hours. We expect these to fall after retirement, and this is indeed the case for both mothers and fathers: see column (1) of Tables 2 and 3. A one standard-deviation rise in the probability of retirement reduces weekly working hours by by 1.1 and 1.2 standard deviations for the mother and father, respectively. As a consequence, leisure satisfaction is expected to rise: in column (2) this increases significantly by 0.81 and 0.76 standard deviations for retired mothers and fathers. The third outcome is the subjective financial situation. Pensions in the UK are relatively low compared to labour income, and financial satisfaction is expected to drop after retirement. This is what is found in column (3), with a fall of -0.34 and -0.52 standard deviations for mothers and fathers, respectively. Last, columns (4), (5), and (6) refer to the estimated effect of retirement on retirees' mental health, subjective health, and overall life satisfaction. Mental health rises significantly post-retirement (by 0.30 and 0.27 standard deviations for mothers and fathers). The analogous figure for subjective health is 0.23 standard deviations for mothers, while there is no significant effect for fathers. Lastly, life satisfaction increases, but significantly only for fathers, by 0.47 standard deviations.

Comparing the second-stage IV results with the OLS results shed light on the magnitude and direction of reverse causality. Apart from financial satisfaction (column 3), all the coefficients have the same sign, but the OLS coefficients are smaller (as is often the case). However, despite being insignificant, the OLS co-

efficient for parental retirement is oppositely signed in the financial satisfaction regressions. This may reflect reverse causality, whereby financially-satisfied older parents retire earlier, irrespective of the State Pension Age, than do those who are more financially-pressed who continue to work longer.

These estimated retirement effects on parental outcomes do not provide support for one of the theoretical mechanisms in Section 2.2.2: the potential rise in informal care from adult children to their parents. This channel is at odds with the mainly positive effect of retirement on parental mental and subjective health, and overall life satisfaction. However, the positive effects on leisure satisfaction and negative effects on financial situation are consistent with the other proposed channels.

Table 2.5.1: The Effect of Mother's Retirement on her Labor Supply and Well-being

Dependent Variables:	Weekly working hours	Leisure Satisfaction	Financial Satisfaction	GHQ	Subjective health	Life Satisfaction
	(1)	(2)	(3)	(4)	(5)	(6)
Second-stage IV results						
Mother retired	-1.10***	0.81***	-0.34**	0.30*	0.14	0.19
	(0.13)	(0.15)	(0.15)	(0.17)	(0.12)	(0.15)
R^2	0.80	0.60	0.65	0.59	0.68	0.63
First-stage IV results						
Mother above SPA	0.29***	0.29***	0.29***	0.28***	0.29***	0.29***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
F-test	1586.00	1487.47	1510.51	1429.79	1534.23	1497.58
Reduced Form						
Mother above SPA	-0.32***	0.24***	-0.11***	0.09*	0.04	0.06
	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
R^2	0.79	0.61	0.66	0.59	0.68	0.64
OLS						
Mother retired	-0.80***	0.31***	0.008	0.10**	0.06*	0.05
	(0.06)	(0.06)	(0.04)	(0.04)	(0.04)	(0.04)
R^2	0.81	0.61	0.66	0.59	0.68	0.64
Individuals	1,610	1,586	1,594	1,564	1,605	1,585
Observations	16,897	16,370	16,560	16,090	16,677	16,368

Clustered (mother) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: All models include individual and year and month of interview-fixed effects. The models include a quadratic trend for the child and mother's age. Bandwidth of 10 years. The coefficients are standardized.

Table 2.5.2: The Effect of Father's Retirement on his Labor Supply and Well-being

Dependent Variables:	Weekly working hours	Leisure Satisfaction	Financial Satisfaction	GHQ	Subjective health	Life Satisfaction
	(1)	(2)	(3)	(4)	(5)	(6)
Second-stage IV results						
Father retired	-1.22***	0.76***	-0.52***	0.27**	-0.04	0.47**
	(0.14)	(0.20)	(0.17)	(0.14)	(0.15)	(0.23)
R^2	0.82	0.64	0.62	0.60	0.67	0.68
First-stage IV results						
Father above SPA	0.22***	0.23***	0.23***	0.23***	0.23***	0.23***
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
F-test	426.92	398.27	405.85	393.90	425.20	401.07
Reduced form						
Father above SPA	-0.35***	0.22***	-0.16***	0.08*	-0.01	0.11**
	(0.05)	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)
R^2	0.77	0.61	0.64	0.58	0.67	0.68
OLS						
Father retired	-1.03***	0.55***	0.06	0.11**	-0.003	0.12***
	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.04)
R^2	0.82	0.65	0.64	0.60	0.67	0.68
Individuals	1,013	908	910	895	940	907
Observations	10,517	9,506	9,604	9,314	9,885	9,500

Clustered (father) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: All models include individual and year and month-fixed effects. The models include a quadratic trend for the child and mother's age. Bandwidth of 10 years. The coefficients are standardized.

2.5.2 The Spillover Effects of Parental Retirement on Adult Children's Well-Being

Tables 2.5.3 and 2.5.4 show the main estimates of the effect of mother's and father's retirement on a battery of adult child well-being outcomes. The regressions control for a quadratic age trend and apply a bandwidth of ten years before and after the pension-eligibility cutoff. The results are shown first for all adult children and then separately for daughters and sons.

In Table 2.5.3, mothers' retirement significantly increases their adult children's life and income satisfaction by about 0.2 standard deviations. In the third panel of this table, the estimated effect on adult-child GHQ is also positive, but insignificant. In Table 2.5.4, the paternal retirement estimates (including those from the reduced-form regression) are all smaller, insignificant and less precisely estimated than those for maternal retirement.

Table 2.5.3: Mother's Retirement and Adult Children's Well-being.

Dependent Variables:	Life satisfaction			Income satisfaction			GHQ		
	All	Daughters	Sons	All	Daughters	Sons	All	Daughters	Sons
Second-stage IV results									
Mother retirement	0.20**	0.24*	0.18	0.21**	0.18	0.24*	0.11	0.18	0.06
	(0.10)	(0.14)	(0.14)	(0.10)	(0.14)	(0.13)	(0.11)	(0.16)	(0.14)
F-test	1577.75	705.16	871.57	1577.75	705.16	871.57	1840.1	920.2	912.3
R^2	0.57	0.56	0.58	0.57	0.57	0.57	0.50	0.48	0.52
Reduced form									
Mother above SPA	0.06**	0.07*	0.05	0.06**	0.05	0.07*	0.03	0.04	0.02
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.04)
R^2	0.58	0.56	0.59	0.57	0.57	0.58	0.50	0.49	0.52
OLS									
Mother retired	0.02	0.001	0.03	0.01	-0.003	0.03	-0.02	-0.02	-0.01
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.0)	(0.03)	(0.04)	(0.04)
R^2	0.58	0.56	0.59	0.57	0.57	0.58	0.50	0.49	0.52
Individuals	3,513	1,721	1,797	3,513	1,721	1,797	3,513	1,721	1,797
Observations	16,984	8,292	8,692	16,984	8,292	8,692	16,984	8,292	8,692

Clustered (individual) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: The regressions include a quadratic trend for the adult child and mother age, and individual, month and year-fixed effects but no other control variables. The age bandwidth is ten years. The coefficients are standardized. The F-test variable refers to the Cragg-Donald F-statistic from the first stage.

Table 2.5.4: Father's Retirement and Adult Children's Well-being.

Dependent Variables:	Life satisfaction			Income satisfaction			GHQ		
	All	Daughter	Son	All	Daughter	Son	All	Daughter	Son
Second-stage IV results									
Father retired	0.07	0.04	0.09	-0.10	-0.29	0.07	-0.24	-0.20	-0.28
	(0.21)	(0.31)	(0.29)	(0.22)	(0.34)	(0.29)	(0.23)	(0.37)	(0.29)
F-test	345.83	151.12	198.84	345.83	151.12	198.84	345.83	151.12	198.84
R^2	0.57	0.57	0.58	0.58	0.58	0.58	0.50	0.48	0.52
Reduced form									
Father above SPA	0.01	0.02	0.004	-0.03	-0.06	0.006	-0.04	-0.03	-0.05
	(0.03)	(0.04)	(0.05)	(0.03)	(0.05)	(0.05)	(0.04)	(0.06)	(0.05)
R^2	0.57	0.57	0.58	0.58	0.58	0.58	0.50	0.48	0.52
OLS									
Father retired	0.04	-0.07	-0.03	0.04	0.02	0.06	0.006	0.03	-0.010
	(0.04)	(0.05)	(0.05)	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.05)
R^2	0.58	0.57	0.58	0.58	0.58	0.58	0.50	0.48	0.52
Individuals	2,635	1,299	1,336	2,635	1,299	1,336	2,635	1,299	1,336
Observations	13,457	6,632	6,825	13,457	6,632	6,825	13,457	6,632	6,825

Clustered (individual) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: see Table 2.5.3.

2.5.3 Heterogeneity

This sub-section asks whether the effect of parental retirement on adult child well-being depends on family characteristics. The first of these is whether the adult children are themselves parents, and if they are the age of the grandchildren. If the positive effect of maternal retirement reflects time transfers via grandchild care, this should only appear for adult children who are parents. In addition, Kaufmann et al., 2023 highlights that the effect of maternal retirement on daughters' labour supply depends on the age of the grandchildren, with an increase in adult daughters' working hours only when the children are aged between 4 and 7.

Table 2.5.5 presents the estimates for maternal retirement using the same specification as in Table 5. First, separating the adult children by parenthood and the ages of their children in columns (1)-(5) (an adult child who has children of different ages may well appear in more than one of these columns). The results are consistent with grandparental childcare, in that maternal retirement has no

significant effect on any dimension of well-being for childless adult children. On the contrary, there are positive significant effects for adult children who are parents. With respect to the grandchild age, the smallest effects are found for grandchildren under the age of three, although all of the estimated coefficients are statistically equal to each other.

We interpret these findings as revealing the importance of grand-maternal childcare for younger children. This is an important transfer, as private childcare in the UK is very expensive. An alternative approach to the extensive margin of retired or not is to consider the intensive margin of older mothers' work hours: the results remain statistically significant and in the same direction (see Appendix 2.E).

Second, inter-generational support may well vary by the geographical distance between adult children and their retired mothers. In Chan and Ermisch, 2011, exchanges between households fall with the travel distance between them in the United Kingdom. A similar result with respect to grand-parental childcare provision is found across 10 European countries in Zanasi et al., 2023. Last, Eibich and Siedler, 2020 find a significant effect of paternal retirement on adult children's fertility but only for travel distances of under one hour between the parents and the adult children. We should however note that the well-being of adult children could influence their parents' location and retirement choices. These heterogeneity results regarding travel distances should, therefore, be interpreted with some caution.

Columns (6)-(8) reveal significant heterogeneity. When the travel time between retired mothers and their adult children is less than one hour, the effect on the latter's life satisfaction, income satisfaction, and mental health is substantial (with rises of 0.41, 0.24, and 0.31 standard deviations, respectively). The estimates for longer travel times are all smaller in absolute value and insignificant.

Third, given that retired grandparents can provide free childcare to their grandchildren, maternal retirement may matter more for poorer adult children (as childcare is less affordable for them). In columns (9) and (10), we carry out separate estimations for adult children whose gross monthly income in the years before their mothers reached the State Pension age was in the bottom quartile or the top quartile in that year. For both life and income satisfaction, the positive effect of maternal retirement is driven by adult children in the bottom income quartile.

We then turn to the marital status of the retired mother. In particular, elderly

non-partnered mothers — whether widowed, divorced or separated — are more likely to support their adult children, including via childcare, due to their greater availability after retirement. Columns (11) and (12) support this hypothesis, with larger effects for all dimensions of adult children from not-married elderly mothers after retirement.

Last, we stratify the sample by the pre-retirement health status of elderly mothers, as their ability to provide childcare and support to their adult children likely depends on their health. This latter is measured all BHPS waves via the question:"In the last 12 months, have you been in a hospital or clinic as an in-patient overnight or longer?". We divide our sample according to whether the elderly mothers were hospitalized at least once in the pre-retirement period. In columns (13) and (14), the life and income satisfaction of adult children with better health are significantly and positively affected by maternal retirement. By way of contrast, the coefficient for adult children with hospitalized mothers is negative (although not statistically significant).

Overall, these results suggest that a substantial part of the main estimates in Table 2.5.3 reflects the time that retired mothers transfer to their adult children. It also sheds light on the gendered nature of childcare responsibilities and the significant role that retired mothers can play in supporting their adult children and grandchildren. Ultimately, these findings may have important implications for policymakers and households, emphasizing the value of intergenerational support and the importance of recognizing and addressing the challenges faced by women in the workforce.

Next, Table 2.5.6 presents the analogous heterogeneity results for paternal retirement. In general, there is little consistent evidence of heterogeneity here with respect to adult-child or parent characteristics. However, there is some evidence (in columns (11)-(12)) that the retirement of unmarried fathers reduces the well-being of their adult children more than the retirement of married fathers.

Table 2.5.5: Mother's Retirement and Adult Child Well-being– Heterogeneity Results

Strata:	No child	Child	Age 5-11	Age 3-4	Age 0-2	Live together	≤ 1 hrs	> 1 hrs	$\leq 25^{th}$ pct	$\geq 75^{th}$	Married	Not married	Not hospital	Hospital
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent Variable														
Life satisfaction	0.17	0.28**	0.38**	0.35**	0.23*	-0.03	0.41***	0.35	0.57***	0.17	0.12	0.44**	0.27**	-0.31
	(0.17)	(0.13)	(0.16)	(0.15)	(0.14)	(0.25)	(0.14)	(0.30)	(0.20)	(0.18)	(0.13)	(0.21)	(0.11)	(0.55)
R^2	0.62	0.53	0.52	0.51	0.52	0.63	0.53	0.48	0.68	0.58	0.58	0.61	0.58	0.79
Income satisfaction	0.0002	0.39***	0.41***	0.49***	0.36***	0.19	0.24*	0.11	0.54***	0.11	0.21*	0.47**	0.26**	-0.16
	(0.16)	(0.13)	(0.15)	(0.15)	(0.14)	(0.24)	(0.14)	(0.27)	(0.19)	(0.16)	(0.12)	(0.21)	(0.11)	(0.38)
R^2	0.61	0.53	0.51	0.51	0.52	0.62	0.55	0.53	0.66	0.55	0.57	0.60	0.58	0.84
GHQ	0.20	0.11	0.35**	0.28*	0.20	0.04	0.31**	-0.20	0.31	0.38*	0.07	0.09	0.12	0.47
	(0.19)	(0.14)	(0.17)	(0.16)	(0.15)	(0.26)	(0.15)	(0.36)	(0.21)	(0.20)	(0.14)	(0.20)	(0.12)	(0.66)
R^2	0.53	0.47	0.46	0.46	0.46	0.56	0.47	0.39	0.61	0.54	0.52	0.51	0.51	0.76
Individuals	2,219	1,299	778	844	1,022	1,654	784	299	1,331	1,914	2,755	914	3,447	846
Observations	8,693	8,291	5,419	6,149	7,096	6,191	5,444	1,900	3,805	7,597	12,871	3,885	15,605	1,379

Clustered (pidp) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: *Not married* includes divorced, separated and widowed older mothers. *Hospital* refers to older mothers who have spent at least one night in hospital in the 12 months before the interview date. For all stratification levels, the coefficients refer to the second-stage IV estimates of the regression of adult child well-being on the residual of the first stage in equation 2.1. All models include a quadratic age trend for adult child and parental age and individual, year, and month fixed effects. There are no other control variables.

Table 2.5.6: Father's Retirement and Adult Child Well-being– Heterogeneity Results

Strata:	No child	Child	Age 5-11	Age 3-4	Age 0-2	Live together	≤ 1 hrs	> 1 hrs	$\leq 25^{th}$ pct	$\geq 75^{th}$	Married	Not married	Not hospital	Hospital
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent Variable														
Life satisfaction	-0.21	0.33	0.16	0.22	0.53	0.10	0.55	-1.31	-0.24	-0.22	0.07	-0.41	0.18	-1.28
	(0.29)	(0.32)	(0.36)	(0.36)	(0.34)	(0.36)	(0.55)	(1.1)	(0.46)	(0.27)	(0.23)	(0.37)	(0.24)	(0.97)
R^2	0.62	0.52	0.53	0.53	0.50	0.59	0.48	0.53	0.68	0.59	0.59	0.59	0.58	0.80
Income satisfaction	-0.18	-0.04	-0.23	-0.02	0.00	-0.18	0.25	-0.36	-0.24	-0.28	-0.10	-0.29	-0.01	-1.36*
	(0.31)	(0.31)	(0.37)	(0.36)	(0.31)	(0.37)	(0.52)	(0.95)	(0.45)	(0.30)	(0.24)	(0.40)	(0.25)	(0.74)
R^2	0.61	0.53	0.51	0.51	0.52	0.62	0.55	0.53	0.66	0.55	0.57	0.60	0.59	0.82
GHQ	-0.46	-0.06	-0.11	-0.45	-0.13	0.04	-0.59	0.85	-1.06**	-0.34	0.03	-1.22***	-0.27	0.69
	(0.34)	(0.33)	(0.41)	(0.39)	(0.34)	(0.37)	(0.76)	(0.99)	(0.50)	(0.31)	(0.25)	(0.46)	(0.27)	(0.68)
R^2	0.53	0.47	0.46	0.46	0.46	0.56	0.47	0.39	0.61	0.54	0.52	0.51	0.51	0.82
Individuals	1,484	1,151	709	745	875	972	136	81	1,027	1,487	2,194	209	2,606	546
Observations	6,016	7,441	5,010	5,528	6,139	4,735	987	501	3,098	6,181	10,276	864	12,616	841

Clustered (pidp) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: *Not married* includes divorced, separated and widowed older fathers.

Note: See Table 2.5.5.

2.5.4 Evidence from Pension Reforms: UKHLS

One potential limitation of the RDD approach is potential anticipation effects, whereby older parents and their adult children may adjust their behavior in anticipation of changes in their well-being. For example, adult children might make choices regarding fertility or employment that affect their parents' propensity to retire and/or their parents' overall well-being around the time of the parent's eligibility for the State Pension. Disentangling these alternative explanations using the RDD design is challenging. However, using changes in policy or other exogenous events that affect parents' eligibility for pension as an alternative source of exogenous variation can help alleviate some of these concerns and provide more robust evidence on the relationship between adult children's well-being and retirement decisions.

As for the RDD analysis above, in the following sections we will first illustrate the direct effect of the UK pension reform on the labour market and well-being outcomes of the parents who were exposed to it; we then evaluate the policy reforms' indirect effects on the parents' adult children.

The direct effect on parents appears in Table 2.5.7. In column (1), in line with our results above and other contributions (see Cribb et al., 2016, Della Giusta and Longhi, 2021), being above the SPA decreases parental weekly working hours by 0.20 and 0.27 standard deviations for mothers and fathers respectively. The analogous effects on parents' leisure satisfaction (Column 2) are positive and significant at 0.11 and 0.17, and those on financial satisfaction are significantly negative at 0.10 and 0.16. The results for mental and subjective health (Columns 4 and 5) are mixed, with only a significant positive mental health impact of being above the SPA for mothers. Lastly, overall life satisfaction (Column 6) is positively affected by being above the SPA at 0.06 and 0.13 standard deviations for mothers and fathers, respectively.

Table 2.5.7: The Rise in the State Pension Age and Older Parents' Labour-market and Well-being Outcomes

	Weekly working hours	Leisure Satisfaction	Financial Satisfaction	GHQ	Subjective health	Life Satisfaction
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Mother above SPA	-0.20***	0.11***	-0.10***	0.10***	0.04*	0.06*
	(0.04)	(0.03)	(0.03)	(0.04)	(0.02)	(0.03)
R^2	0.76	0.52	0.67	0.58	0.74	0.53
Individuals	5,048	4,913	4,981	4,886	5,045	4,913
Observations	49,516	47,785	49,048	47,335	49,275	47,785
Panel B						
Father above SPA	-0.27***	0.17**	-0.16***	-0.06	0.05	0.13**
	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)
R^2	0.75	0.53	0.68	0.63	0.75	0.53
Individuals	3,518	3,265	3,324	3,247	3,464	3,265
Observations	31,576	29,069	30,066	28,816	16,040	29,066

Clustered (birth year) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: All coefficients are standardized. The control variables are being married, having a degree, living with their adult child, the adult child's age in months and individual, interview year and month fixed effects.

Table 2.5.8 then turns to the estimated effects of the pension reform on the adult children's well-being. In opposition to the RDD results, there are here no significant effects of maternal retirement on adult-child well-being, with all estimated coefficients being close to zero. This may reflect the anticipation of reforms by adult children, who adjust their expectations and behaviour accordingly. This anticipation will reduce the estimated effect of the reform (as some of the observations in the control group will actually be treated). A second potential explanation is the concurrent expansion of publicly-provided and free childcare in the UK during the period over which the pension reforms were implemented (UK Government, 2023).

However, in Panel B of Table 2.5.8 paternal retirement has significant negative effects on adult sons' life and income satisfaction. Paternal retirement thus increases the leisure and life satisfaction of the parents who are concerned but is detrimental to the well-being outcomes of adult sons.

Table 2.5.8: The Rise in the State Pension Age and Adult Child Well-being.

Dependent Variables:	Life satisfaction			Income satisfaction			GHQ		
	All	Daughters	Sons	All	Daughters	Sons	All	Daughters	Sons
Panel A									
Mother above SPA	0.003	0.01	-0.008	-0.007	-0.02	0.004	-0.008	-0.03	0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
R^2	0.51	0.49	0.54	0.53	0.51	0.55	0.51	0.50	0.52
Individuals	11,033	5,476	5,566	11,033	5,476	5,566	10,957	5,435	5,530
Observations	59,778	30,464	29,313	59,778	30,464	29,313	58,957	30,009	28,947
Panel B									
Father above SPA	-0.04	0.05	-0.14**	-0.02	0.06	-0.12**	-0.005	0.02	-0.04
	(0.05)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.07)	(0.07)
R^2	0.52	0.49	0.55	0.54	0.51	0.57	0.53	0.52	0.54
Individuals	7,523	3,770	3,754	7,523	3,770	3,754	7,445	3,735	3,713
Observations	35,923	18,222	17,665	35,923	18,222	17,665	35,328	17,930	17,397

Clustered (pidp) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Note: All coefficients are standardized. Standard errors in parentheses are clustered at the adult child level. All regressions control for adult children and older parent variables: age, marital status, housing tenure, labour-market activity, and individual, interview year and month fixed effects.

2.5.5 Heterogeneity

The above finding that adult sons' life and income satisfaction are negatively affected by parental retirement could reflect that adult children support their fathers financially in retirement. Equally, not working older fathers may stop financially supporting their adult children.

One implication of these channels is that the effect of parental retirement should differ by the adult child's income. With transfers from the adult child to the parent, delayed retirement of the latter should have a greater effect on higher-income sons (as they are more likely to support their retired fathers financially); conversely, lower-income sons will likely be more affected by transfers in the opposite direction. To investigate heterogeneity by adult-child income, we estimate Equation 2.3 with an interaction term between the treatment dummy and the income quartile dummies of the adult child.

Table 2.5.9 lists the results. The effect for adult children in the first income

quartile is given by the estimated treatment coefficient (father above the SPA). At the same time, the interaction term shows the differential impact on adult children in higher-income quartiles. There are no significant estimates for the whole sample of adult children in columns (1), (3) and (5). The results for sons in columns (2), (4) and (6) mostly reveal smaller point estimates for richer adult sons regarding life and income satisfaction (although none of the interaction terms is statistically significant).

Another piece of evidence comes from the travel-distance stratification, where we expect to find larger effects for adult children who still live with their fathers or are nearby. As well as the time-transfer channel, it may also be the case that financial support from older parents to adult children increases with proximity (Berry, 2008).

Table 2.5.10 shows the estimated coefficients on the interaction terms between travel distance and the treatment dummy. The main treatment coefficient (father above SPA) reveals the effect for adult children who live with their fathers. The effect of father's retirement on adult-child well-being is almost always larger for children (and especially sons) who live close by, as shown by the estimated coefficient on Father above SPA $\times < 1$.

Table 2.5.9: The Rise in Father's State Pension Age and Adult Child Well-being: Heterogeneity by Adult Child Income

Dependent Variables:	Life satisfaction		Income satisfaction		GHQ	
	All	Sons	All	Sons	All	Sons
	(1)	(2)	(3)	(4)	(5)	(6)
Father above SPA	-0.05	-0.22***	-0.04	-0.16*	-0.0008	0.03
	(0.06)	(0.08)	(0.06)	(0.08)	(0.06)	(0.07)
Father above SPA × 2nd quartile	-0.005	0.05	0.02	-0.04	-0.01	-0.11
	(0.04)	(0.07)	(0.04)	(0.07)	(0.05)	(0.07)
Father above SPA × 3rd quartile	0.008	0.10	0.010	0.04	-0.03	-0.04
	(0.04)	(0.06)	(0.04)	(0.06)	(0.05)	(0.07)
Father above SPA × 4rt quartile	0.02	0.09	0.03	0.07	-0.008	-0.03
	(0.04)	(0.06)	(0.04)	(0.06)	(0.05)	(0.07)
R^2	0.52	0.55	0.54	0.57	0.54	0.55
Individuals	7,473	3,727	7,473	3,727	7,444	3,712
Observations	35,547	17,501	35,547	17,501	35,325	17,395

Clustered (pidp) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

Table 2.5.10: The Rise in Father's State Pension Age and Adult Child Well-being: Heterogeneity by Travel Distance

Dependent Variables:	Life satisfaction		Income satisfaction		GHQ	
	All	Sons	All	Sons	All	Sons
Father above SPA	0.07	0.02	-0.007	-0.11	0.08	0.11
	(0.06)	(0.08)	(0.06)	(0.08)	(0.07)	(0.10)
Father above SPA × ¡1hr	-0.14***	-0.17***	-0.05	-0.03	-0.09*	-0.18***
	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)
Father above SPA × ¿1hr	-0.07	-0.10	0.04	0.02	-0.06	-0.14*
	(0.05)	(0.07)	(0.06)	(0.07)	(0.06)	(0.08)
R^2	0.46	0.49	0.50	0.52	0.51	0.51
Individuals	4,076	1,906	4,076	1,906	4,052	1,895
Observations	24,579	11,517	24,579	11,517	24,252	11,381

Clustered (pidp) standard-errors in parentheses

*Significance: *** = 10%; ** = 5%; * = 10%*

2.6 Conclusion

We have here used linked parent-child information from two UK household panel datasets to establish the spillover effects of parental retirement on the well-being of their adult children. This effect was identified first in a Regression Discontinuity Design analysis using eligibility age for the State Pension as a tool for identification within a (RDD) framework. Here, only the mothers' retirement increased their adult children's life satisfaction and income satisfaction. These effects were larger for adult children who live close to their parents, who have children themselves, and who have lower incomes. These findings are consistent with inter-generational time transfers from retired mothers to their adult children, highlighting the importance of childcare provisions and affordability. Delayed retirement will then have a potentially large spillover effect on adult children's well-being and labour-market outcomes, especially those from lower-income households.

The second analysis considered the rise of the UK's state pension age for women and men. The difference-in-difference analysis here shows that fathers' retirement

reduces their adult children's life and income satisfaction, with the results being driven by adult sons; there was no significant effect of mothers' retirement on adult-child well-being. This is consistent with inter-generational financial transfers from fathers to adult children.

Our most general finding is that public policies can have inter-generational spillover effects with significant distributional consequences, underscoring the importance of both financial and time transfers. These spillovers should enter into any evaluation of policies that aim to change retirement behaviour.

References

Atalay, K., & Zhu, R. (2018). The effect of a wife's retirement on her husband's mental health. *Applied Economics*, *50*(43), 4606–4616.

Berry, B. (2008). Financial transfers from living parents to adult children: Who is helped and why? *American Journal of Economics and Sociology*, *67*(2), 207–239.

Buchanan, A., & Rotkirch, A. (2018). Twenty-first century grandparents: Global perspectives on changing roles and consequences.

Carrino, L., Glaser, K., & Avendano, M. (2020). Later retirement, job strain, and health: Evidence from the new State Pension age in the United Kingdom. *Health Economics*, *29*(8), 891–912.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Chan, T. W., & Ermisch, J. (2011). Intergenerational exchange of instrumental support: dynamic evidence from the British Household Panel survey.

Charles, K. K. (2004). Is retirement depressing?: Labor force inactivity and psychological well-being in later life. *Research in Labor Economics*, 269–299.

Clark, A. E., & Zhu, R. (2024). Taking back control? Quasi-experimental evidence on the impact of retirement on locus of control. *The Economic Journal*, *134*(660), 1465–1493.

Coe, N. B., & Zamarro, G. (2011). Retirement effects on health in Europe. *Journal of Health Economics*, *30*(1), 77–86.

Cox, D. (1987). Motives for private income transfers. *Journal of Political Economy*, *95*(3), 508–546.

Crawford, R., & Tetlow, G. (2008). Employment, retirement and pensions.

Cribb, J., Emmerson, C., & O'Brien, L. (2022). *The effect of increasing the state pension age to 66 on labour market activity* (tech. rep.). IFS Working paper.

Cribb, J., Emmerson, C., & Tetlow, G. (2016). Signals matter? Large retirement responses to limited financial incentives. *Labour Economics*, *42*, 203–212.

Dave, D., Rashad, I., & Spasojevic, J. (2008). The effects of retirement on physical and mental health outcomes. *Southern Economic Journal*, *75*(2), 497–523.

Della Giusta, M., & Longhi, S. (2021). Stung by pension reforms: The unequal impact of changes in state pension age on uk women and their partners. *Labour Economics*, *72*, 102049.

Eibich, P., & Siedler, T. (2020). Retirement, intergenerational time transfers, and fertility. *European Economic Review*, *124*, 103392.

Fé, E., & Hollingsworth, B. (2016). Short-and long-run estimates of the local effects of retirement on health. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *179*(4), 1051–1067.

Filomena, M., & Picchio, M. (2023). Retirement and health outcomes in a meta-analytical framework. *Journal of Economic Surveys*, *37*(4), 1120–1155.

Gorry, A., Gorry, D., & Slavov, S. N. (2018). Does retirement improve health and life satisfaction? *Health Economics*, *27*(12), 2067–2086.

Ilciukas, J. (2023). Fertility and parental retirement. *Journal of Public Economics*, *226*, 104928.

Kaufmann, K., Özdemir, Y., & Ye, H. (2023). *Spillover effects of old-age pension across generations: Family labor supply and child outcomes* (tech. rep. No. crctr224_2023_403). University of Bonn and University of Mannheim, Germany.

Lacey, R. E., McMunn, A., & Webb, E. (2019). Informal caregiving patterns and trajectories of psychological distress in the UK Household Longitudinal Study. *Psychological Medicine*, *49*(10), 1652–1660.

Lewis, A., Barton, C., & Cromarty, H. (2021). Housing an ageing population: A reading list. *London: Commons Library Briefing*.

Lewis, A. (2021). Housing an Ageing Population: A reading list.

Mazzonna, F., & Peracchi, F. (2012). Ageing, cognitive abilities and retirement. *European Economic Review*, *56*(4), 691–710.

OECD. (2012). OECD family database pf1.7: Intergenerational solidarity 2012.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with Weak Instruments. *Econometrica*, *65*(3), 557–586.

UK Government. (2023). £755 million to double free childcare offer for 2-year-olds [Accessed: 21.03.2024].

University of Essex, Institute for Social and Economic Research. (2023). Understanding Society: Waves 1-13, 2009-2022 and Harmonised BHPS: Waves 1-18, 1991-2009.

Special Licence Access [data collection] (18th Edition) [http://doi.org/10.5255/UKDA-SN-6614-19].

Van den Berg, B., Fiebig, D. G., & Hall, J. (2014). Well-being losses due to care-giving. *Journal of Health Economics*, *35*, 123–131.

Van Houtven, C. H., Coe, N. B., & Skira, M. M. (2013). The effect of informal care on work and wages. *Journal of Health Economics*, *32*(1), 240–252.

Wu, Q., & Gao, X. (2020). The effects of parental retirement on adult children's labor supply: Evidence from China. *Available at SSRN 3718085*.

Zanasi, F., Arpino, B., Bordone, V., & Hank, K. (2023). The prevalence of grandparental childcare in europe: A research update. *European Journal of Ageing*, *20*(1), 37.

Zhu, R., & He, X. (2015). How does women's life satisfaction respond to retirement? A two-stage analysis. *Economics Letters*, *137*, 118–122.

Appendix

2.A General Health Questionnaire

Table 2.A.1: GHQ questions/responses

GHQ questions / responses	**1**	**2**	**3**	**4**
Been able to concentrate on whatever you are doing?	Better than usual	Same as usual	Less than usual	Much less than usual
Lost much sleep over worry?	Not at all	No more than usual	Rather more than usual	Much more than usual
Felt that you are playing a useful part in things?	More so than usual	Same as usual	Less so than usual	Much less capable
Felt capable of making decisions about things?	More so than usual	Same as usual	Less so than usual	Much less capable
Felt constantly under strain?	Not at all	No more than usual	Rather more than usual	Much more than usual
Felt you could not overcome your difficulties?	Not at all	No more than usual	Rather more than usual	Much more than usual
Been able to enjoy your normal day-to-day activities?	Much more than usual	Same as usual	Less so than usual	Much less than usual
Been able to face up to your problems?	More so than usual	Same as usual	Less able than usual	Much less able
Been feeling unhappy and depressed?	Not at all	No more than usual	Rather more than usual	Much more than usual
Been losing confidence in yourself?	Not at all	No more than usual	Rather more than usual	Much more than usual
Been thinking of yourself as a worthless person?	Not at all	No more than usual	Rather more than usual	Much more than usual
Been feeling reasonably happy all things considered?	More so than usual	About same as usual	Less so than usual	Much less than usual

2.B Sample biases

2.B.1 Attrition bias

The bar chart in Figure 2.B.1 represents the sample composition of adult children (16+) in the British Household Panel Survey (BHPS) across various waves, from wave 1 in 1991-92 to wave 18 in 2008-09. Each bar is color-coded to indicate the wave in which respondents first participated, providing insight into adding new respondents over time and retaining participants across consecutive waves.

The initial wave (1991-92), represented by the dark purple segment at the base of each bar, has the highest number of adult children respondents. Over time, additional cohorts of adult children and their spouses or partners entered the survey, which is evident in the different colours appearing in subsequent waves. The introduction of households from Scotland and Wales in wave 9 (1999-2000) and Northern Ireland in wave 11 (2001-02) also increased the number of adult children in these respective waves.

Furthermore, the chart visually indicates attrition over time; the diminishing height of the colored segments corresponding to the first wave suggests decreased participation from the original cohort. This is consistent with the reported attrition rates, with 52% of the initial adult children participants remaining after 18 years. The varied height of the bars also reflects the survey's dynamic nature, with different numbers of respondents in each wave due to attrition and the addition of new households.

Figure 2.B.1: BHPS Adult children waves composition

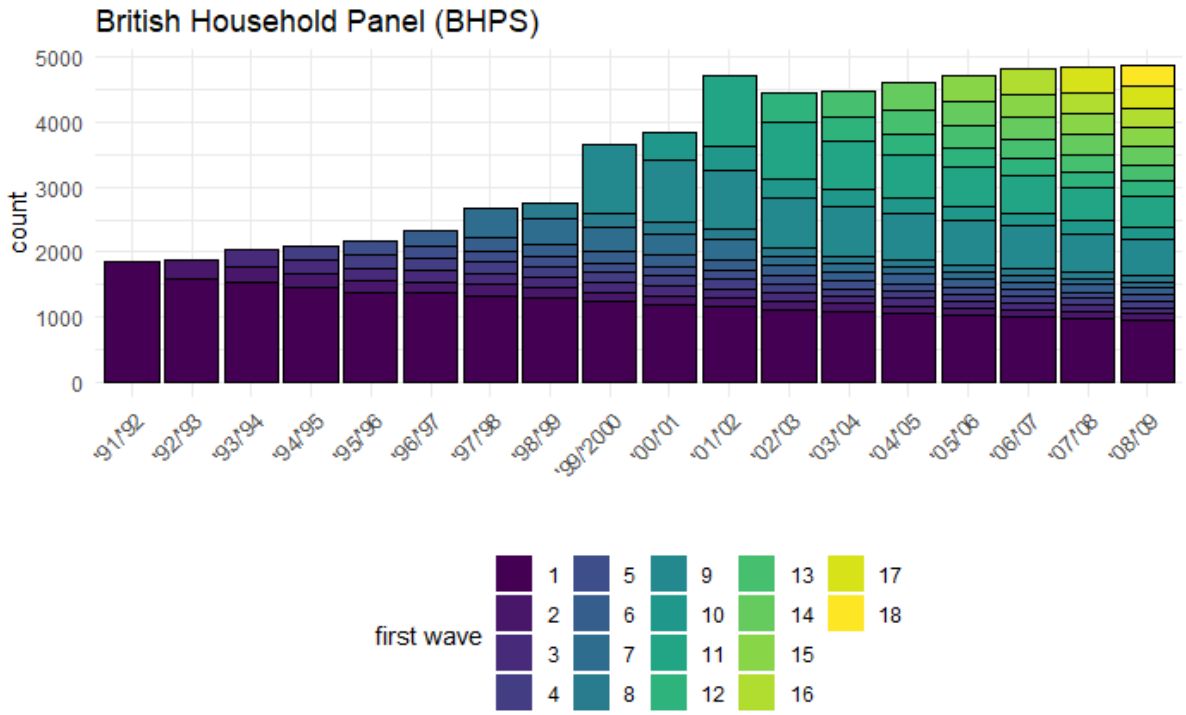


Table 2.B.1 illustrates results from a regression of individual attrition on a set of demographic and outcome variables in BHPS for adult children (16+), elder mothers, and elder fathers. The main demographic predictors of attrition are male, white, and older ages for adult children and elderly parents. Retirement status predicts a higher likelihood of dropping from the sample for only fathers, which may indicate a potential reason for the insignificant effect we found in this sample.

Table 2.B.1: Characteristics of Attritors in BHPS sample

	Attritors Adult Children	Attritors Elder mothers	Attritors Elder Fathers
Variables			
Life Satisfaction	-0.006**	0.006	-3.7×10^{-5}
	(0.003)	(0.005)	(0.006)
Income satisfaction	-0.009***	-0.01***	-0.01**
	(0.002)	(0.004)	(0.005)
GHQ	0.001	0.0002	-0.003*
	(0.0006)	(0.001)	(0.002)
White	0.05***	0.10***	0.08***
	(0.009)	(0.01)	(0.02)
Age	-0.005***	-0.0002	-0.001
	(0.0004)	(0.0008)	(0.001)
Female	-0.02**		
	(0.009)		
Active	0.002		
	(0.007)		
Retired		0.02	0.08***
		(0.02)	(0.02)
Fit statistics			
Observations	45,998	19,763	14,365
Pseudo R^2	0.01983	0.00983	0.01073
BIC	47,220.6	22,688.0	16,853.5

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The degree of attrition from the UKHLS survey is also high: of the initial sample of adult children, 33% still participate after 12 years, figure 2.B.2. Table 2.B.2 illustrates results from a regression of individual attrition on the demographic and outcome variables set in UKHLS. Compared to the BHPS, in the UKHLS sample, the same predictors are statistically significant, but being white decreases the

probability of dropout. Moreover, in UKHLS, retirement predicts drop out for both fathers and mothers.

Figure 2.B.2: UKHLS waves composition

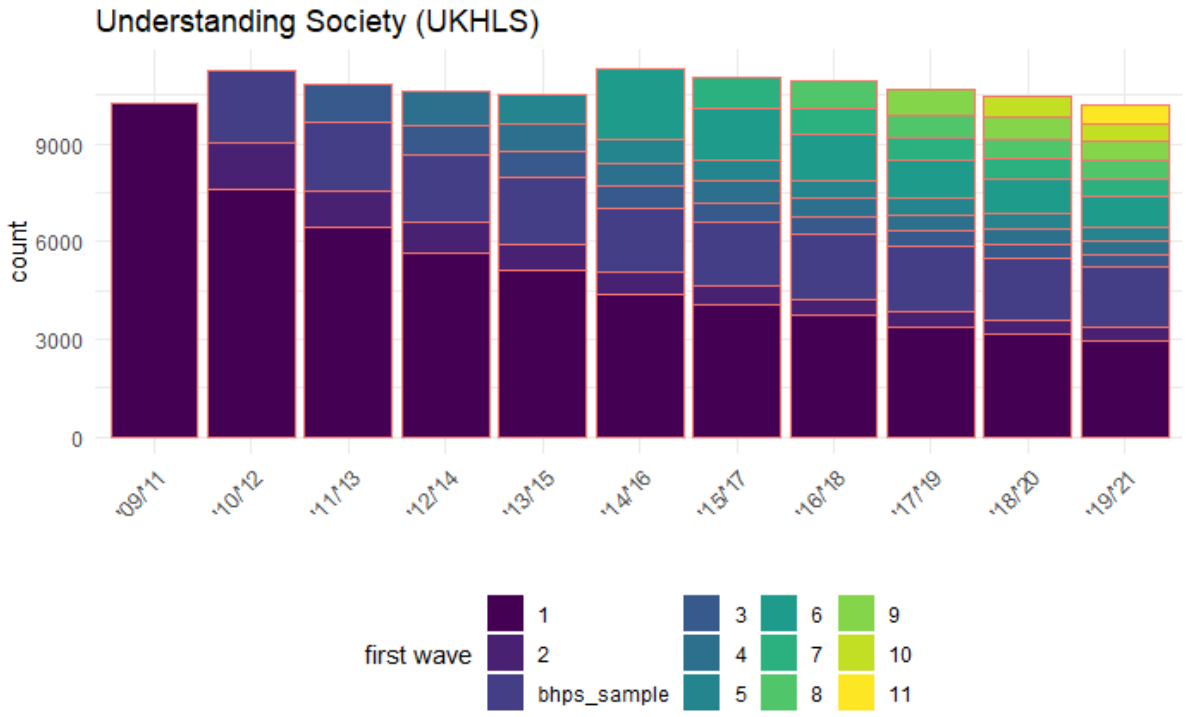


Table 2.B.2: Characteristics of Attritors in UKHLS sample

	Attritors Adult Children	Attritors Elder mothers	Attritors Elder Fathers
Life Satisfaction	0.006***	0.005**	0.006**
	(0.002)	(0.002)	(0.003)
Income Satisfaction	-0.02***	-0.02***	-0.02***
	(0.002)	(0.002)	(0.003)
GHQ	0.004***	-0.0002	-0.002**
	(0.0004)	(0.0006)	(0.0009)
White	-0.08***	-0.08***	-0.14***
	(0.006)	(0.009)	(0.01)
Age	-0.005***	-0.004***	-0.004***
	(0.0003)	(0.0006)	(0.0007)
Female	-0.02***		
	(0.007)		
Active	-0.03***		
	(0.005)		
Retired		0.05***	0.06***
		(0.01)	(0.02)
Fit statistics			
Observations	111,136	58,764	38,275
Pseudo R^2	0.03350	0.01416	0.02445
BIC	143,868.7	81,235.3	51,659.2

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

2.B.2 Co-residence bias

According to Torche, 2019: "If older co-resident children are included in the analysis, this induces the risk of bias insofar as children who continue to live with parents after late adolescence might not be a representative sample of their cohort. Selection bias induced by selecting co-resident children beyond their late adolescence is a

concern, even if the sample is restricted to children who are young adults".

To assess co-residence bias in our sample of adult children, we ran a simple descriptive analysis comparing summary statistics of the demographic variables in our main sample of adult children and the full sample of respondents from BHPS in the same age range and cohort as the one used in the main analysis (25-45; 1963+). Table 2.B.3 reports the mean and standard deviation for the full and adult children samples and the associated p-value for the statistical test of their difference. The two samples differ statistically significantly in all dimensions considered. This is a significant bias we should consider when considering the validity of our main causal estimates.

Table 2.B.3: Co-residence bias in the BHPS sample.

Variables	Adult Children	Full Sample	p.adj.signif
GHQ	25.15	24.64	****
	(5.33)	(5.60)	
Income satisfaction	4.50	4.42	****
	(1.44)	(1.52)	
Life satisfaction	5.20	5.10	****
	(1.14)	(1.23)	
Active	0.87	0.84	****
	(0.34)	(0.37)	
Age	31.46	35.58	****
	(4.94)	(5.79)	
Female	0.46	0.54	****
	(0.50)	((0.50)	
Labour income	1548.48	1495.68	****
	(1176.82)	(1343.01)	
Married	0.39	0.58	****
	(0.49)	(0.49)	
Number children	0.71	1.14	****
	(0.98)	(1.14)	
Year of birth	1972	1966	****
	(5.35)	(6.51)	
Years of education	16.99	16.42	****
	(2.00)	(1.03)	
Number of Observations	18333	57686	

Mean and standard deviation (in parenthesis) of outcome and demographic variables in the BHPS Adult Children sample and the Full sample and associated p-value for t-test for the difference in means.

Table 2.B.4: Co-residence bias in the UKHLS sample.

Variables	Adult Children	Full Sample	p.adj.signif
GHQ	24.46	24.54	
	(5.75)	(5.76)	
Income satisfaction	4.38	4.16	****
	(1.99)	(2.57)	
Life satisfaction	4.95	4.80	****
	(1.90)	(2.52)	
Active	0.88	0.84	****
	(0.33)	(0.37)	
Age	31.37	(35.76	****
	(5.13)	(5.95)	
Female	0.50	0.57	****
	(0.50)	(0.49)	
Labour income	1933.79	2139.41	****
	(1857.12)	(5409.37)	
Married	0.31	0.55	****
	(0.46)	(0.50)	
Number children	0.55	1.14	****
	(0.89)	(1.14)	
Year of birth	1984.06	1978.18	****
	(5.70)	(6.86)	
Years of education	16.65	16.62	***
	(1.10)	(1.19)	

Mean and standard deviation (in parenthesis) of outcome and demographic variables in the UKHLS Adult Children sample and the Full sample and associated p-value for t-test for the difference in means.

2.C Fuzzy RDD Assumption

Smoothness in density: For an RDD design to be valid, individuals must not manipulate the assignment variable, which, in our case, is the parent's age in months. We run continuity density tests around the cutoff for mothers and fathers separately to test the continuity in the parent's age range. The density test consists of a null hypothesis that the density of the running variable is continuous at the cutoff. In other words, the null hypothesis is that there is no "manipulation" of the density at the cutoff. Failing to reject implies no statistical evidence of manipulation at the cutoff (Cattaneo et al., 2019). Figure 2.C.1 illustrates the results from this test and confirms the absence of manipulation around the cutoff.

Choice of bandwidth: One of the most critical decisions in RDD is selecting the appropriate bandwidth around the cutoff. This parameter establishes the maximum age range from the discontinuity. Observations beyond this range are unused.

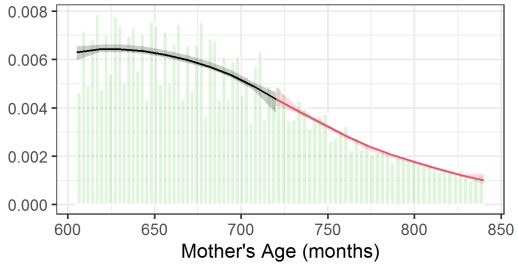
Choosing a narrow bandwidth minimizes bias, but it may increase variance due to a smaller number of observations. On the other hand, selecting a larger bandwidth reduces variance but can potentially increase bias. In the main specification, we use a bandwidth of ten years, covering ages 50 to 70 for mothers and 55 to 75 for fathers. We perform robustness checks with bandwidths of eight, five and three years.

Smoothness in covariates: One fundamental assumption of the RD design is that other predetermined characteristics of the parents and adult children that may affect adult children's well-being should not change discontinuously at the threshold. Parents' predetermined variables include race, college degree, and number of biological children. Adult children's predetermined covariates are race, female/male ratio, years of education, and degree. Figures 2.C.2, 2.C.3, and 2.C.4 illustrate the RD plot for children, mothers, and fathers, overlaid with lines from local linear regressions using data within ± 10 years window. The graphs show no visible discontinuities at the cutoff, indicating that local assignment around the cutoff is random. Overall, the RD validity checks support our empirical strategy and provide no evidence of violations of the key identifying assumptions.

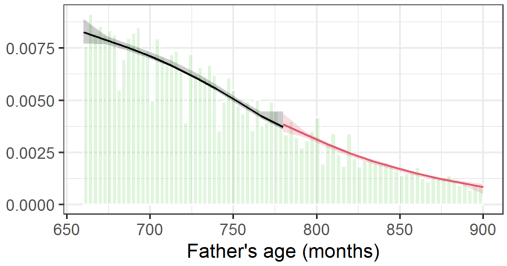
Instrument validity: There are three conditions necessary to interpret the two-stage least squares estimate. First, parents' age is strongly associated with retirement status. We show the validity and magnitude of the first-stage relationship in Section 2.4.1. Second, we need to assume that parents' age only impacts adult children's outcomes through the change in retirement probability. This assumption might be violated if adult children anticipate their parent's eligibility for a state pension and adjust their well-being accordingly.

Figure 2.C.1: Density plots of the running variable. Mothers (left panel) and father (right panel

(a) Mother

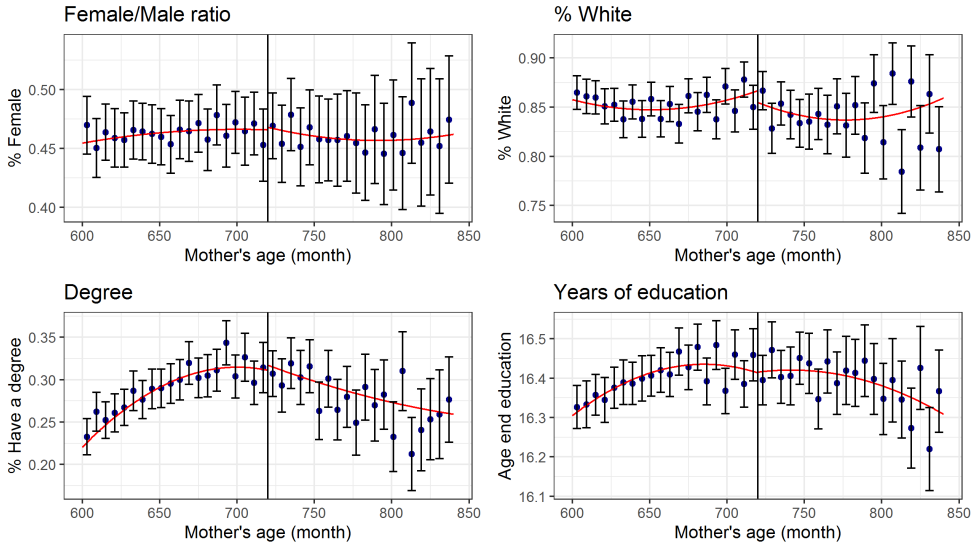


(b) Father



Notes: The plots show the estimated probability density function of the running variable. The plot uses parental age (in months) as the running variable and assumes a threshold at age 720 for mothers and 780 for fathers. The density functions were estimated using the *rddensity* package in R, using a local quadratic polynomial for the estimation, a cubic polynomial for the bias correction, a triangular kernel, and jackknife standard errors.

Figure 2.C.2: Tests for the continuity of the adult child's predetermined variables around the mother SPA.



Source: BHPS, own calculations. The dots show averages by parental age in years. The lines show a quadratic fit, and the shaded areas show a 95% confidence interval.

Figure 2.C.3: Tests for the continuity of the adult child's predetermined variables across the father SPA. BHPS, own calculations. *Note*: see Figure 2.C.2.

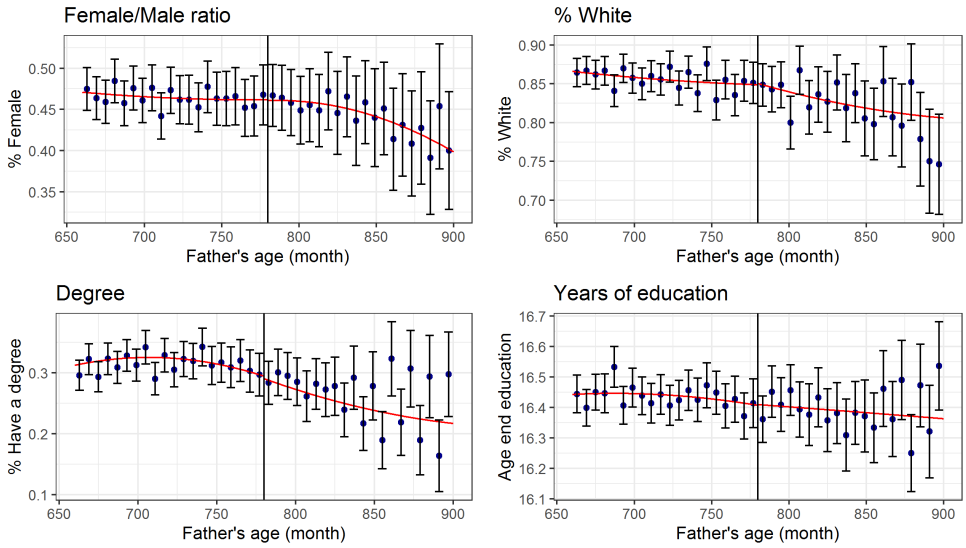
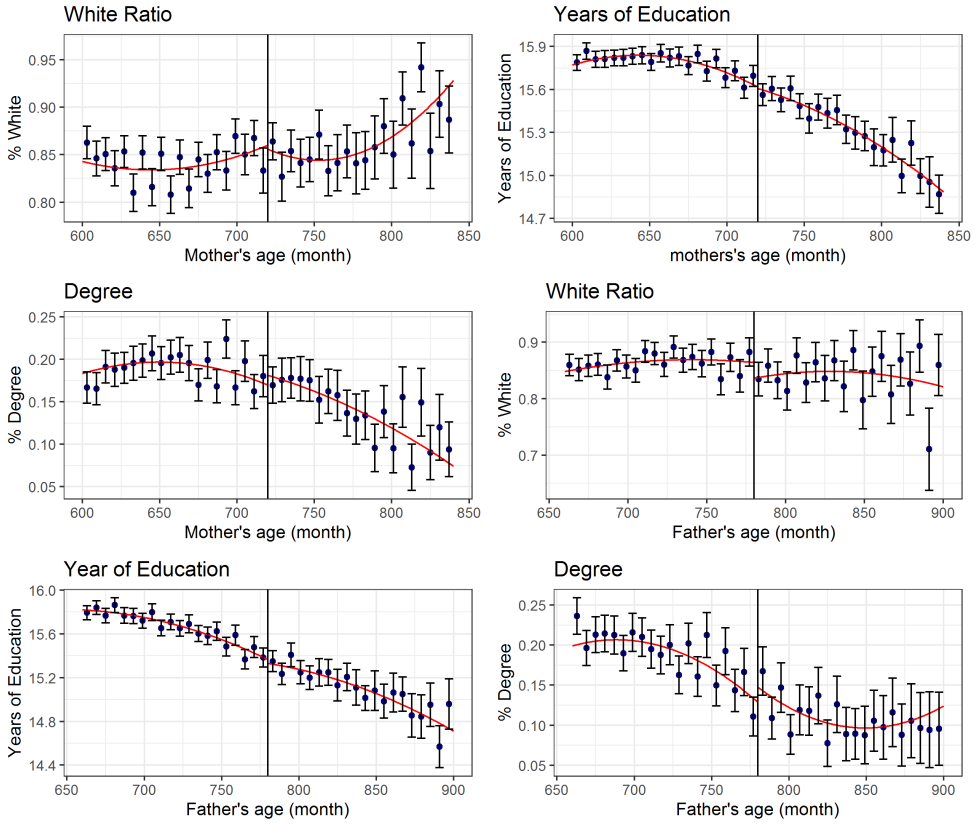


Figure 2.C.4: Tests for the continuity of the parent's predetermined variables across the parent's SPA threshold. BHPS, own calculations.



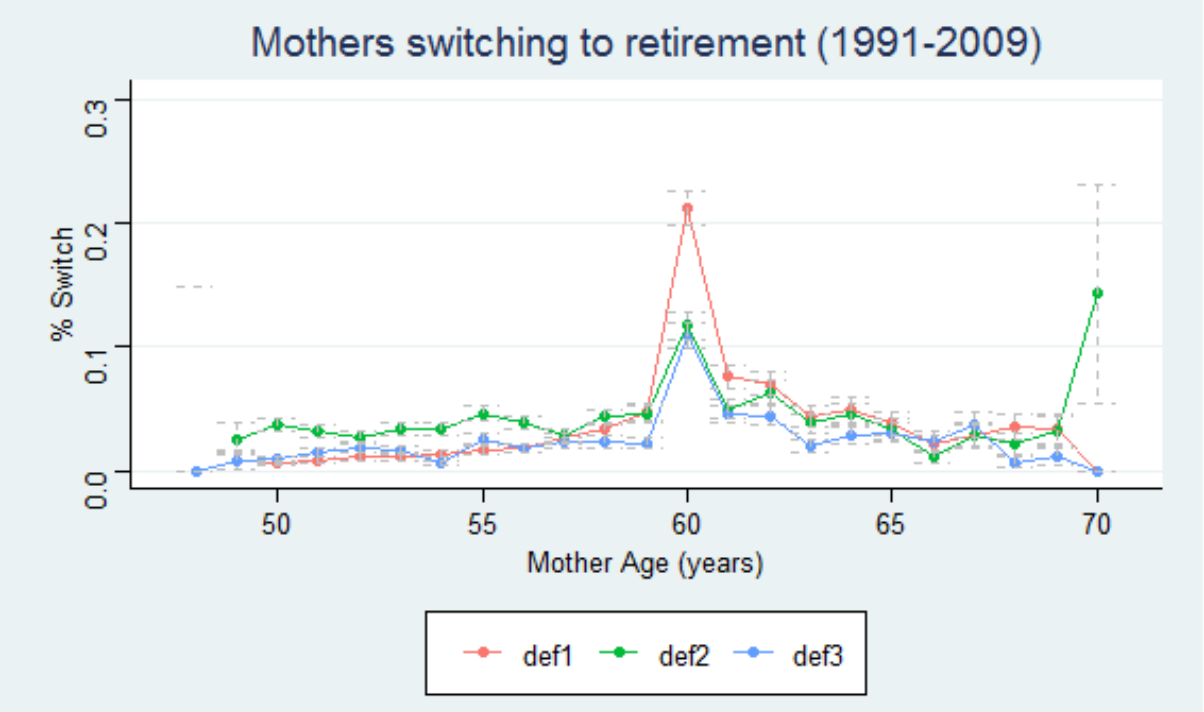
Note: see Figure 2.C.2.

2.D Sensitivity Analysis

The RDD results may be sensitive to the retirement definition used. Indeed, in the main analysis, we defined parents as retired if they self-report being retired. However, the literature presents other definitions of retirement that may still be valid. This section assesses the sensitivity of our results to the retirement definition. We assess the sensitivity of results focusing on maternal retirement only.

We analyze three plausible variants. The first variant considers parents as retired if they are self-declared retired or report being inactive and not looking for a job in the month before the interview date. The second variant considers parents as retired only if they receive a state pension benefit. Figure 2.D.1 illustrates the switching probability according to the three definitions. The third variant uses parental job hours instead of the self-reported retirement definition as the outcome variable in stage 1 of equation 4.1.

Figure 2.D.1: BHPS own elaboration. Mothers switching to retirement according to the three definitions of retirement.



Note: Definition 1 refers to the retirement definition adopted in the main analysis, Definition 2 refers to the variant where we include also inactive and not looking for jobs individual, Definition 3 refers to the variant where we consider the condition of receiving the State Pension£

Table 2.D.1: Mother retirement and adult children's well-being- Sensitivity analysis by retirement definition.

Dependent Variables:	Life satisfaction			Income satisfaction			GHQ		
	First variant	Second variant	Job hours	First variant	Second variant	Job hours	First variant	Second variant	Job hours
Second-stage IV results									
retired (1st var)	0.36**			0.39**			0.20		
	(0.18)			(0.17)			(0.19)		
retired (2nd var)		0.33*			0.39**			0.19	
		(0.18)			(0.17)			(0.19)	
Job hours			-0.20**			-0.21**			-0.12
			(0.09)			(0.09)			(0.10)
Observations	16,775	16,572	16,897	16,775	16,572	16,897	16,398	16,197	16,510
R^2	0.56948	0.57427	0.56902	0.56752	0.57192	0.56457	0.50278	0.50547	0.50067
First-stage IV results									
mother above SPA	0.17***	0.17***	-0.33***	0.17***	0.17***	-0.33***	0.17***	0.17***	-0.33***
	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)
Observations	16,775	16,572	16,897	16,775	16,572	16,897	16,398	16,197	16,510
R^2	0.79270	0.75460	0.79874	0.79270	0.75460	0.79874	0.79438	0.75511	0.79999
F-stat	410.8	640.1	401.6	626.5	403.6	641.9	447.1	628.2	
OLS									
etired (1st var)	-0.01			0.02			-0.009		
	(0.03)			(0.03)			(0.03)		
retired (2nd var)		-0.010			0.04			0.01	
		(0.03)			(0.04)			(0.04)	
Job hours			-0.006			-0.008			9.9×10^{-5}
			(0.01)			(0.01)			(0.01)
Observations	16,775	16,572	16,897	16,775	16,572	16,897	16,398	16,197	16,510
R^2	0.57658	0.57853	0.57655	0.57386	0.57549	0.57275	0.50487	0.50647	0.50361

Clustered (pidp) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The models include a quadratic trend for the child and parents' age. All models include fixed effects for adult children and year and month-fixed effects. Age bandwidth of ten years. The coefficients are all standardized.

2.E Robustness checks

We perform four robustness checks for the model specification. The first assesses the robustness of the RDD-IV results to change in the parental age bandwidth around the cutoff, looking at eight, five, and three years. The second enlarges the age bandwidth of adult children from 20-45 to 16-50, holding the parental age at ± 10 years before and after the State Pension Age. The third check considers a different, binary specification of the outcome variables. The fourth modifies the functional form of the age variable in the main specification from quadratic to linear, cubic or quartic.

Table 2.E.1: Mother retirement and adult children's outcomes- Robustness checks: Age bandwidths.

Dependent Variables:	Life Satisfaction				Income Satisfaction				Leisure Satisfaction				GHQ			
Bandwidth:	8 years	5 years	3 years	16-50	8 years	5 years	3 years	16-50	8 years	5 years	3 years	16-50	8 years	5 years	3 years	16-50
Second-stage IV results																
mother retirement	0.25**	0.29**	0.28	0.20**	0.19*	0.20	0.17	0.19**	0.03	-0.02	-0.07	-0.005	0.09	0.08	0.16	0.13
	(0.11)	(0.13)	(0.20)	(0.10)	(0.10)	(0.12)	(0.19)	(0.09)	(0.11)	(0.13)	(0.20)	(0.10)	(0.10)	(0.13)	(0.18)	(0.09)
Observations	14,423	9,434	6,104	19,417	14,444	9,458	6,121	19,420	14,449	9,459	6,120	19,448	16,785	10,746	6,838	22,943
R^2	0.581	0.618	0.669	0.566	0.585	0.625	0.666	0.571	0.558	0.582	0.626	0.539	0.499	0.529	0.581	0.482

Clustered (pidp) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: The models include a quadratic trend for the child and parents' age. All models include fixed effects for adult children and year and month-fixed effects. Age bandwidth of ten years. The coefficients are all standardized.

Table 2.E.2: Mother retirement and adult children's outcomes- Robustness checks: Age functional form.

Dependent Variables:	Life Satisfaction			Income satisfaction			GHQ		
	Linear	Cubic	Quadratic	Linear	Cubic	Quadratic	Linear	Cubic	Quadratic
mother retirement	0.19*	0.22**	0.22**	0.24***	0.23**	0.23**	0.14	0.14	0.14
	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)	(0.10)	(0.10)	(0.11)	(0.11)
Fit statistics									
Observations	16,984	16,984	16,984	16,984	16,984	16,984	16,597	16,597	16,597
R^2	0.57381	0.57372	0.57379	0.57026	0.57052	0.57050	0.50274	0.50256	0.50296

Clustered (pidp) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: All models include fixed effects for adult children and year and month-fixed effects. Age bandwidth of ten years. The coefficients are all standardized

2.F Placebo regressions

We perform three placebo regressions to support the robustness of our Fuzzy RDD estimates. Specifically, we estimate our main specification using variables as outcomes that maternal retirement should not affect. Specifically, we looked at (i) whether the adult children have a university degree, (ii) whether they vote or support any political party, and (iii) their subjective health. Table 2.F.1 illustrates the results. Effects sizes are small and not statistically significant.

Moreover, we estimate separate regression with varying State Pension Ages as placebo cutoffs. The results of this exercise are in Table 2.F.2.

Table 2.F.1: Mother retirement and adult children's outcomes- Placebo regressions.

Dependent Variables:	Degree			Vote			Subj. health		
Sample:	All	Daughters	Sons	All	Daughters	Sons	All	Daughters	Sons
Second-stage IV results									
mother retirement	-0.02	-0.04	0.005	0.02	-0.003	0.05	-0.01	-0.10	0.07
	(0.02)	(0.02)	(0.02)	(0.10)	(0.18)	(0.11)	(0.08)	(0.12)	(0.10)
Observations	23,037	10,928	12,109	21,262	10,306	10,956	22,245	10,631	11,614
R^2	0.92580	0.92158	0.92971	0.70835	0.72974	0.59814	0.54805	0.52506	0.56944
F-stat	1,855	907.2	942.7	1,786.2	871.5	909.0	1,820.4	888.1	924.1

Clustered (pidp) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: The models include a quadratic trend for the child and parents' age. All models include adult children's fixed and year- and month-fixed effects. Age bandwidth of ten years. The coefficients are all standardized.

Table 2.F.2: Placebo State Pension age for maternal retirement

Distance to actual cutoff:	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
Outcome Variables									
life satisfaction	0.41	0.46	0.36*	0.32*	0.22**	0.24**	0.10	0.12	0.02
	(0.41)	(0.40)	(0.18)	(0.18)	(0.10)	(0.11)	(0.12)	(0.12)	(0.18)
income satisfaction	0.37	0.39	0.20	0.21	0.23**	0.23**	0.25**	0.21*	0.27
	(0.41)	(0.40)	(0.18)	(0.18)	(0.10)	(0.10)	(0.12)	(0.12)	(0.18)
GHQ	0.54	0.44	0.19	0.17	0.14	0.20*	-0.03	-0.03	0.11
	(0.42)	(0.41)	(0.19)	(0.19)	(0.11)	(0.12)	(0.13)	(0.13)	(0.19)
Observations	16,984	16,984	16,984	16,984	16,984	16,984	16,984	16,984	16,984
Clustered (pidp) standard-errors in parentheses									
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*									

Note: The numbers represent years from the actual cutoff

Chapter 3

Small Pictures, Big Biases: The Adverse Effects of an Airbnb Anti-Discrimination Policy

3.1 Introduction

Ethnic disparities persist today across many domains, including the labor market, housing, criminal justice, credit, and education, with evidence pointing to discrimination as a contributing factor (Lang & Spitzer, 2020). At the policy level, designing effective anti-discrimination interventions remains an unsolved challenge (Valfort, 2018). Indeed, the effectiveness of each intervention depends on the underlying motives of discrimination (Bohren et al., 2023). Such motives are context-dependent and may vary according to the discrimination category, i.e., ethnicity, gender, or age.

Given this complexity, the question of whether reducing the prominence of personal information for individual quality assessment can effectively mitigate discrimination outcomes remains open, as the literature provides mixed evidence on this issue (see, e.g., Agan and Starr, 2018). This ongoing debate is particularly relevant in digital platforms, which rely heavily on exchanging personal information to boost users' mutual trust. This paper leverages a design change on the Airbnb platform, a leading online marketplace for short-term rentals, to measure how reducing the prominence of personal information impacts ethnic disparities in digital market outcomes.

Founded in 2008, Airbnb has stood out as a dominant actor in the online short-term rental market. With over 6.6 million active listings in over 220 countries worldwide (Airbnb, 2022a), Airbnb has solidified its leading role, surpassing established hotel giants like Marriott (Edelman et al., 2017). Like many other digital platforms, Airbnb connects virtually suppliers and consumers (Einav et al., 2016) from all over the world. Its design embeds trust-building mechanisms like peer reviews and personal information sharing (Gössling et al., 2021). In particular, names and photos are crucial design components to increase trust and reduce anonymity among users in online interactions (see, e.g., Guttentag, 2013; Bente et al., 2012). At the same time, they provide an avenue for users to enact discriminatory practices based on visible attributes, such as perceived ethnicity, age, and gender (see Edelman and Luca, 2014, Fisman and Luca, 2016, Edelman et al., 2017, Levy and Barocas, 2017).

Focusing on New York City, we empirically investigate the extent of ethnic

disparities between service providers (i.e., Airbnb Hosts), considering Asian, Black, Hispanic, and White ethnic groups, and assess a policy to reduce them. We create a face classification algorithm to predict hosts' ethnicity from profile pictures and merge the ethnicity classifications to our main Airbnb listings panel. Here, we derive an extensive set of traditional and innovative control variables, including apartment location and observable characteristics of apartments and hosts. Using an Ordinary Least Squares (OLS) model, we estimate the impact of host ethnicity on two market outcomes: occupancy rate and prices per night. Subsequently, we investigate the impact of an Airbnb anti-discrimination policy that reduced the displayed size of user profile pictures from 256 to 104 square pixels (see Figure 3.A). To estimate the causal effect, we employ Two-Way Fixed Effects (TWFE) regressions alongside Difference-in-Differences (DiD) estimators and Event-Study (ES) approaches.

The policy strategically reduced the dimension of profile pictures within the Airbnb users' interface. We hypothesize that this design transformation affected the salience of positive cues inferred by guests from the host's profile pictures, such as attractiveness, trustworthiness, friendliness, or whether the person is smiling. Such characteristics are known to be correlated with market outcomes, but the salience of these features might differ according to guest ethnicity and profile picture size (Ert et al., 2016; Jaeger et al., 2019). Notably, the new design did not involve other significant simultaneous design changes.

This study yields three main findings. First, in line with previous research (see, e.g., Edelman and Luca, 2014, Marchenko, 2019, Laouénan and Rathelot, 2017), our analysis shows that ethnic minority hosts have lower occupancy rates than comparable White hosts. Specifically, our main findings indicate that Black hosts have approximately 7.2 percentage points lower occupancy rates than their White counterparts. For Asians and Hispanics, the disparity is around 1.4 percentage points and differs from zero at the 1% statistical level. Notably, in contrast to prior studies, we observe very small ethnic differences in pricing, which are insignificant for Black and Hispanic hosts. After considering a broader and more detailed set of control variables, we find that ethnicity plays only little discernible role in pricing.

Second, our analysis shows that the Airbnb design transformation did not narrow the gap in occupancy rates between ethnic minority hosts and White hosts. Our

Occupancy rate is a proxy for the number of bookings.

results indicate that, within six months of its adoption, the new Airbnb design leads to a 4 percentage point increase in the occupancy rate disparity between White and Black hosts. Price disparity does not change. However, we observe Black minority hosts reacting endogenously, changing a more flexible feature than price. Black hosts notably increased the basic amenities listed in their listing descriptions. We do not observe the same behavior for other ethnic groups.

Third, through heterogeneity analysis, we shed light on the potential motives underlying discrimination and provide suggestions for the adverse effect of the design transformation. First, by stratifying the sample by apartment types, i.e., shared rooms or entire apartments, we find that Black hosts offering private rooms suffer a larger occupancy rate penalty than their White hosts' counterparts. This result suggests that some taste-based discrimination is in place on the platform. Second, by stratifying the sample by the hosts' number of reviews, our findings indicate that listings without reviews (i.e., new listings) or those with lower-than-average reviews experienced greater discrimination and slightly higher impacts of the design transformation. This observation suggests a mixture of screening and statistical discrimination likely lies at the core of the observed residual ethnic disparities. In the absence of more objective evaluations, such as guest reviews, potential guests may rely on the host's profile pictures to assess the quality of the listing. The policy's significant reduction in the displayed size of pictures constrains the information available to guests, limiting their ability to judge the host's trustworthiness and other positive facial clues. This transformation negatively impacts the accuracy of guests' overall listing quality assessments, increasing the penalty against minority hosts.

This paper contributes to several strands of literature. First, our results speak to the broader literature assessing the existence and extent of ethnic discrimination in digital platforms (see, e.g., Doleac and Stein, 2013, Edelman and Luca, 2014, Marchenko, 2019, Laouénan and Rathelot, 2022). We created more comprehensive data than previous research by adding new explanatory variables, mainly host information derived from written text and precise apartment location. Therefore, we constructed a panel data set of Airbnb listings and hosts from May 2016 to December 2019 that we leveraged by including listing fixed effects in our causal identification strategy.

With a more comprehensive set of controls, we reduce the Omitted Variable Bias (OVB) risk. As in these previous studies, we found a significant occupancy rate disparity between majority and minority hosts, which is substantial for Black hosts. However, by including a larger, more comprehensive set of control variables, we found no significant residual price differences across ethnicity.

Second, we contribute to the growing yet inconclusive literature assessing the impact of anti-discrimination policies that limit or reduce the salience of personal information shared among transacting agents. Current research tends to focus on labour market discrimination from employers to prospective candidates (Behaghel et al., 2015, Agan and Starr, 2018). We are the first to evaluate anti-discrimination policies' short-term effects on digital platforms. Our results highlight an adverse effect of the policy on Black hosts. Despite the fact that we cannot fully disentangle the mechanisms underlying this negative effect, our study reveals new insights that could guide the rationale of anti-discrimination intervention on digital platforms more broadly.

Lastly, we contribute methodologically by developing and refining state-of-the-art Vision Transformers models (ViT) (Dosovitskiy et al., 2020) to encode images and text into traditional and novel variables; our new features include guest review sentiment and the encoding of information provided in the "hosts' about" section of the platform. This way, we create an automated and powerful machine that streamlines the manual process of decoding visuals and textual features from the platform, making it more efficient and less burdensome. As a result, our work advances our understanding of discrimination in general and specifically on digital platforms. It offers a valuable resource for future research, enabling the identification of these new features on a broader scale and across various image data sources.

We proceed as follows. Section 2 presents the theoretical and empirical context, and Section 3 illustrates data and descriptive statistics. Section 4 illustrates our empirical strategies, and Section 5 shows the main results. Section 6 presents robustness checks and discusses policy implications. Section 3.8 concludes.

3.2 Background

3.2.1 Discrimination and Anti-Discrimination Policies

Ethnic discrimination on digital platforms can be analyzed through several theoretical lenses: taste-based, statistical, screening, and inaccurate statistical discrimination. Each framework provides insights into potential mechanisms driving discriminatory behaviours in online environments. This section synthesizes the key theoretical and empirical literature on discrimination models, focusing on those types linked to the main results of this paper.

The taste-based discrimination model, grounded in Becker's seminal work (Becker, 1957), posits that discrimination stems from personal biases against specific ethnic groups independent of economic rationality. In digital platforms, this model suggests that users may avoid interactions with individuals from certain ethnic backgrounds, irrespective of their qualifications or the quality of their service. We assess this type of discrimination by stratifying the sample by apartment types, i.e., shared/private rooms and entire apartments, characterized by high and low levels of host-guest interactions, respectively. We expect that if taste-based discrimination is in place, ethnic minorities that offer shared or private rooms will be more penalized compared to White hosts.

Conversely, the statistical discrimination model, advanced by Phelps and Arrow (Phelps, 1972, Arrow, 1973), contends that discrimination may occur even in the absence of explicit prejudice, as rational behavior. The rationale relies on true aggregate differences between groups' underlying characteristics, such as apartment quality or host reliability. In Airbnb, users have incomplete information about an apartment's quality. Therefore, they might default to generalizing based on true ethnic group differences in apartments' characteristics, resulting in discriminatory outcomes. If statistical discrimination exists, we should not find any ethnic differential after controlling for all the objective and observable characteristics that differ across ethnicities.

The screening discrimination model, proposed by Cornell and Welch, 1996, eliminates the notion of irrational prejudices, inherent group differences, or intra-group preferences. It posits that, in settings where a group of evaluators has to screen and choose the best candidate, the overall accuracy in screening will be higher for those

candidates belonging to the same ethnic group as the majority of evaluators. In this case, discrimination is a rational response to incomplete information. In contexts like Airbnb, guests more accurately assess accommodation quality when sharing their ethnicity with the host. For example, the trustworthiness signal derived from a smile in the profile picture of a White host will be more precisely assessed by White guests than by other ethnic minorities. We test for this model by analyzing the matching patterns among hosts and guests.

The concept of inaccurate statistical discrimination, introduced by Bohren et al. (2023), extends the model of statistical discrimination by attributing discrimination to inaccurate/incorrect beliefs about group characteristics, often due to informational deficits. This is the only theoretical model empirically tested in the context of Airbnb by Laouénan and Rathelot (2022), who found substantial evidence of its prevalence in the platform.

Understanding the root causes of discrimination is crucial for devising effective anti-discrimination measures. Despite a lack of empirical evidence on digital platforms, the literature provides evidence of successful and unsuccessful anti-discrimination interventions in other domains. The insights from these examples highlight that each type of discrimination requires tailored interventions.

Goldin et al.'s (2000) study assessing the introduction of blind auditions within symphony orchestras highlights a significant reduction in gender discrimination, promoting a higher rate of female musicians' employment (Goldin & Rouse, 2000) Another landmark study explored an information intervention aimed at reducing prejudice (taste-based) against transgender individuals, with effective results that persisted for three months after the intervention (Broockman & Kalla, 2016).

However, some interventions can backfire. Behaghel et al., 2015 found that anonymizing resumes reduced minority candidates' chances of obtaining interviews, as it prevented counterbalance of other negative resume signals, which in turn may be due to systemic/institutional discrimination (Bohren et al., 2022). Similarly, Agan and Starr's study on "ban the box" policies revealed that such interventions decreased callback rates for minority applicants, as employers generalized criminal backgrounds more frequently to Black applicants in the absence of the "criminal box" (Agan and Starr, 2018).

This paper contributes to this growing literature by showing that, in the digital

context, reducing the dimension of service providers' profile pictures without compensating with another measure that might increase user mutual trust amplifies the residual ethnic disparity in digital market supply outcomes.

These studies' insights guide policymakers' interventions to limit exposure to sensitive characteristics when taste-based discrimination is in place. On the other hand, increasing transparency and information could be more effective when discrimination is based on incorrect statistical beliefs. Finally, reducing informational differences in signals between ethnic groups could be beneficial to mitigate screening discrimination. Without accurately targeting the underlying motives of discrimination, interventions may fail and exacerbate the issues they aim to mitigate.

3.2.2 Airbnb

This section provides a detailed overview of how Airbnb's hosts and guests interact within the platform. Understanding these dynamics is the first step for framing our identification strategy and empirical analysis. Moreover, explaining the standard Airbnb host-guest interaction clarifies the role of users' ethnicity throughout the booking process. We also highlight the design features that Airbnb's anti-discrimination package modified over the years, with a particular focus on the one we assess in this paper.

Becoming a guest or host on Airbnb involves a straightforward process. Prospective hosts have to provide detailed information about their properties and themselves. Hosts can modify the information on their listings, update property details and personal information, or revise prices anytime. Similarly, guests undergo a registration process, sharing personal details that enable hosts to make informed decisions about their visitors.

The usual interaction on the Airbnb platform starts with guests searching for their desired city and period. This initial step opens up a filtering system where properties can be sorted based on various criteria, including maximum and minimum price, number of guests, and room type. After filtering, guests receive a visual list of the available properties with basic information, such as daily price per night, property pictures, some information on the host and its profile picture, and the overall property rating. The guest can also locate the listing on a map. Guests can access detailed information by clicking on the listing, including the host's first name,

a comprehensive property description, a standardized list of amenities, additional photos, and reviews from previous guests. From this page, guests can further click on the host's picture and access the host's page. Here, guests get more information on the hosts, such as the number of reviews, overall ratings, and personal information in the "host about" section. The host's personal page, mainly the prominence of the host picture, has been primarily modified by the anti-discrimination policy we assess in this paper.

Once guests create their final preference, they click the "Book It" button, shifting the decision to the host when the instant booking option is not activated. In this case, the host can accept or reject the guest without justification. The listing will be directly booked when the instant booking option is activated. Guests who get rejected receive an email encouraging them to look for another place (Laouénan & Rathelot, 2017). The rejection is not displayed on the host's profile. If the host accepts the guest, the booking is finalized. Guests can still cancel their booking with penalties varying based on the host's chosen cancellation policy. Hosts can also cancel the booking, incurring no financial penalty but a reputation price, as the cancellation automatically appears on their profile as a review.

Addressing concerns of fairness and inclusivity head-on, in 2016, Airbnb started implementing a robust anti-discrimination package designed to promote equality within its community. This comprehensive set of platform design transformations aims to mitigate potential biases and discrimination for all users Murphy (2016). In 2016, they introduced an Instant Booking option, which allows guests to book immediately without the Host's approval, provided the guest completed a basic information form. In October 2018, Airbnb significantly changed the platform design and reduced the size of profile pictures displayed on the user's pages from 256 to 104 square pixels (Airbnb, 2022b).. Finally, in March 2019, Airbnb eliminated the photo of guests before the booking from the hosts was confirmed.

Guests might still zoom in on the profile picture to make it bigger. However, the picture will be of lower quality, as the original uploaded picture is smaller. Moreover, the Airbnb App did not allow zooming in on pictures in 2018.

We checked the exact starting date of the policy transformation by consulting WeyBack Machine ("Internet Archive", 2023) and checking the appearance of the Airbnb platform through time, see Appendix 3.A for before and after pictures

3.3 Data

3.3.1 Data Sources

We assembled datasets from publicly available sources to examine the relationship between Airbnb hosts' ethnicity, prices, and occupancy rates. We created a monthly panel from May 2016 to December 2019 (44 months) of the universe of Airbnb listings in New York City (NYC). The main data comes from Inside Airbnb, an independent project that scrapes monthly information worldwide from the Airbnb website (Cox, 2017). The Inside Airbnb data stores all public information appearing on the listing pages: listing availability (i.e., if a listing is available or not for rent on a specific day), the price per night in US$, and many other listings' characteristics and hosts' personal information. Lastly, it contains information on the guests' first names and reviews for each analysed listing.

We restrict our analysis to months after May 2016 since there are some months for which the scrape data is unavailable. We include only months before December 2019 since Airbnb activities were strongly affected in 2020 and 2021 due to the COVID-19 pandemic(see, e.g., Hossain, 2021). Within this time frame, we analyze only active listings (i.e., those receiving at least one review in six consecutive months). We impose this restriction as we want to focus on hosts who are seriously committed to renting their accommodation through Airbnb. Moreover, we select listings with valid URLs for the hosts' profile pictures. Thus, we focus only on hosts having one human face in their profile pictures.

To examine whether this last selection criterion affects the representativeness of our sample, we employ the Moran I test for spatial auto-correlation. This test helps us determine if hosts' tendency not to show human faces in their profile pictures—or to include multiple faces—is geographically clustered, such as in neighbourhoods predominantly inhabited by Black residents. If such a pattern were found, excluding

"To scrape" refers to web scraping, a technique used to collect data from websites.

As a reference point, the first available scrape for download, at the time we downloaded the data, was March 2015. Even though listing data was available for most months before Mai 2016, the calendar data was not.

We also remove hosts that have been on the Airbnb platform for at least 6 months but never received a guest and those hosts that have not updated their calendars for one year or longer.

Analyzing multi-face images is technically challenging due to the need for a specialized model to classify demographics in such pictures, and interpretation becomes complex as group images (like families) may send mixed signals. Profiles without human images offer no demographic data, as they lack facial features for analysis.

these hosts could reduce the representativeness of these areas in our sample. Results from the Moran test (Appendix) indicate no significant spatial correlation in the pattern of profile picture characteristics. Suggesting that our last sample selection criterion does not affect the neighbourhood representativeness.

3.3.2 Outcome Variables

We use the listing availability information to create monthly occupancy rates, one of our main dependent variables. The occupancy rate is the number of days a listing is unavailable for booking in the 30 days following the scraping date divided by 30. Two points are worth mentioning. First, the data scraping may not occur on the first day of each month. In these few cases, we could not observe the listing availability information for the days of the month before the scraping. The second important point is that the unavailability of a listing on the Airbnb calendar doesn't necessarily imply an actual booking. Hosts might choose to block their accommodation on certain days or periods, introducing a potential measurement error in this outcome variable.

Therefore, the occupancy rate is a proxy for the monthly bookings. For this proxy to be valid, the critical assumption is that there is no systematic difference in the renting availability patterns between ethnic minority and White hosts. If this assumption holds, the measurement error in occupancy rate will not bias our results.

However, if there is a systematic bias that correlates with hosts' ethnicity – for instance, if minority hosts more frequently block out dates unrelated to bookings – it would artificially inflate ethnic minorities' occupancy rates, resulting in an underestimation of the actual disparity in occupancy rates between ethnic minorities and Whites hosts. Conversely, if White hosts are more prone to such non-booking-related unavailability, it would overestimate the actual disparity in occupancy rates.

According to Marchenko (2019), the first scenario is more likely to occur, as White hosts tend to provide more entire apartments and houses on average. The reasoning suggests that if White hosts are less likely to reside in the properties they offer, possibly because they own additional properties, then the availability of their listings might remain high, irrespective of actual demand. The descriptive statistics in Table 3.F.2 of Appendix 3.F show that, on average, White hosts provide

In most sampled months, the scraping was made within the first three days of the respective month. However, in October 2019, the earliest scrape occurred on the 14th day of the month

more entire apartments or houses than minority hosts. In contrast, minority hosts often offer private or shared rooms, which may not always be available for rent as they reside on the property.

Another way to explore this assumption is by looking at host cancellation rates. If minority hosts block out dates unrelated to bookings more frequently, their cancellation rates are expected to be lower, as you can only cancel a booked date and not a blocked one. On the other hand, if White hosts block out dates unrelated to bookings more frequently, their cancellation rates are expected to be lower. Cancellation rates are almost identical across host ethnicity (see Table 3.F.3 in Appendix 3.F). Therefore, if there are systematic differences in the renting availability patterns between ethnic minorities and White hosts, we are most likely underestimating the ethnic disparity.

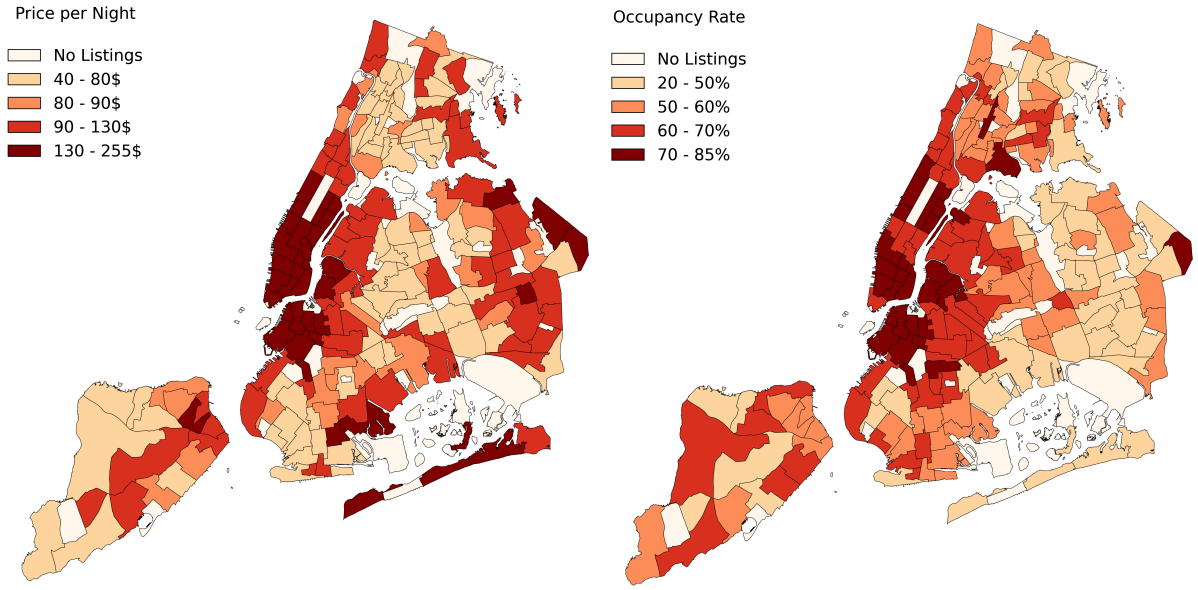
The second dependent variable is the monthly price per night in US$, which is reported in the listing information, excluding additional fees, for example, a cleaning fee or price surcharges (e.g., due to weekends). We report the spatial distribution of average occupancy rates and average prices in Figure 3.3.1.

The spatial price and occupancy rate distributions in New York City are clustered among neighborhoods. The highest prices and occupancy rates are in Manhattan and Brooklyn Heights, and the lowest are in the Bronx and Queens.

Under the assumption that the probability of cancellations is the same across ethnicities.

For listings scraped before 2019, the daily price contained in the calendar data is missing whenever a listing is unavailable. Otherwise, one could build a daily panel and explore the rising price variation due to weekends, holidays, and other factors.

Figure 3.3.1: Spatial distribution of outcomes



(a) Price per Night

(b) Occupancy Rate

Note: Lighter colours indicate low values in the outcome variable. Numbers are averages over all the months in the panel (44 months).

3.3.3 Ethnic prediction

We determine the perceived ethnicity of each Airbnb host by fine-tuning an image classification algorithm based on state-of-the-art Vision Transformer (ViT) models (Dosovitskiy et al., 2020). The model analyzes facial features from profile pictures and returns the predicted probabilities for each ethnicity, e.g. Asian, Hispanic, Black and White. We then transform these probabilities into a unique class label by assigning the ethnicity with the highest predicted probability among the four . We applied a similar training pipeline and fine-tuned the model to extract three other facial features, i.e. host perceived gender, age, and smiling. The detailed description of the model training and testing pipeline is in Appendix 3.D

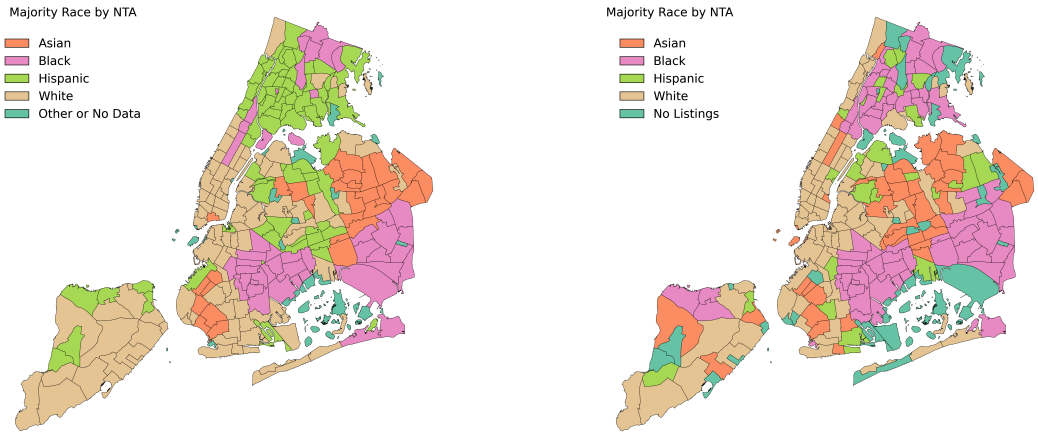
As an additional accuracy check, in Figure 3.3.2 we plot the spatial distribution of predicted hosts' ethnicities together with the ethnic distribution as documented in the New York census of 2020. The ethnic pattern aligns with nationally representative data, reassuring about the sample representativeness.

In the difference-in-difference approach, we additionally restrict the sample to listings where the predicted ethnicity and gender remain constant during the

Specifically, we use the softmax function.

analyzed period. The reasons for this change are related to the change in profile pictures. Approximately 10% of the listings are excluded due to this restriction.

Figure 3.3.2: Spatial distribution of Hosts ethnicity. NYC census (2020) and Airbnb hosts



(a) NYC Census (2020)

(b) Predicted Host ethnicity

Note: In the figure, each NTA neighbourhood takes the colour of the majority true ethnicity of NYC residents (left) and the predicted ethnicity of Hosts (right)

3.3.4 Controls

The data from Inside Airbnb provide all publicized information potential guests can use to select their preferred listing. Naturally, our econometric models include most of this information as control variables. The summary statistics of all control variables are in Appendix 3.F. Here, we summarise the main macro-categories.

Listing characteristics This category comprises all listing information visible on the listing pages. These are the maximum number of guests that could be hosted, with and without additional fees, the number of bedrooms, bathrooms, and beds, the cleaning fees, whether the instant booking option is activated, the minimum of nights that have to be booked, number and type of amenities, whether house, apartment, loft or townhouse, whether shared or private room or a full house, the security deposit and type of cancellation policy.

Host characteristics This category includes all the features attached to the Host: being a super host, whether the Host's identity has been verified, months of experience as an Airbnb host, number of listings, Host's gender, smile, age, face dimension in the profile picture; informativeness of the host self-description; if the Host requires the guest profile picture or the guest's phone number and whether the host has the instant booking option active.

Geographic controls We combine listing approximate geographic coordinates (latitude and longitude) with geographic information system software (GIS) to pinpoint to which Neighborhood Tabulation Area (NTA) and Public Use Microdata Area (PUMA) of NYC each listing belongs. For security reasons, Airbnb does not disclose the exact geographic coordinates of the listing before booking but instead a 150-meter radius. For each listing, a latitude and longitude corresponding to the centroid of the 150-meter radius is assigned. In total, we include 55 different PUMAs and 240 NTAs in NYC.. In addition, Using GIS and Metropolitan Transportation Authority data (Metropolitan Transportation Authority, 2019), we create a variable for the distance of each listing to the nearest subway station in kilometers. Moreover, we use OpenStreetMap data (OpenStreetMap, 2021) to create variables for the distance of each listing to the nearest supermarkets. We also create variables for the number of stores, bars and pubs, restaurants, and touristic places/attractions within a 500 meters radius around each listing.

Guest Reviews We process guest reviews for each listing page. This information is very relevant, as other studies highlight the influence of the content and number of reviews in reducing inaccurate statistical discrimination against ethnic minority hosts (see Laouénan and Rathelot, 2022). We extract the sentiment of the reviews for each listing (i.e., if the feedback is positive, neutral, or negative) and the review language. To do so, we apply a large language model called RoBERTa (Liu et al., 2019). This model proved to perform remarkably well in natural language processing tasks. After classifying sentiment scores for each review, we create the average

The NTAs are subsets of the PUMAs.

Since the reported location of each listing can change over time due to randomization of location, we take the average of the geographic controls for each listing.

monthly sentiment of each listing. Finally, we store the monthly number of reviews, monthly stock of reviews, and the additional number of reviews in a given month, i.e., the differences in the "stock" of reviews in a month with respect to the previous month.

3.4 Empirical Strategy

This paper aims to answer two main research questions. The first is whether the ethnicity of Airbnb has a role in predicting their occupancy rates and prices and which sign it has. The second is whether the 2018 Airbnb anti-discrimination policy effectively mitigates this impact. We construct a large set of control variables and ethnicity indicators to estimate an Ordinary Least Squares (OLS) regression and obtain the residual average impact of host ethnicity on listings' occupancy rates and prices. Second, we estimate the policy effect on occupancy rates and prices over time (pre vs. post-policy) between different ethnic groups (White vs. Black/Asian/Hispanic hosts) by applying a DiD estimator using a Two-Way Fixed Effects model (TWFE). We do not have a control group in this setting, as the policy was simultaneously implemented for all the hosts. Section 3.6.1 discusses the effects we capture with our estimation strategy. Third, we implement an Event-Study (ES) approach of our DiD estimator to assess the dynamics of the policy effects.

For both versions of the DiD, we can only identify the policy's short-term effects in the six months after its implementation. Widening the time period would significantly restrict the sample to well-established and experienced hosts unlikely to suffer from discrimination. Indeed, Airbnb is a dynamic platform where new hosts enter while others leave the platform at a high turnover.

Moreover, as the parallel trends assumptions of the standard DiD do not always perfectly hold, as a robustness test, we implement the Synthetic Difference-in-Differences (SDID) approach, a state-of-the-art extension from the traditional DiD to solve the problem of no pre-trends (Arkhangelsky et al., 2021).

This means that the first month in our sample (May 2016) has the average sentiment of all reviews given up to May 2016. We do not create the average sentiment of only reviews made in May 2016 because old reviews are not deleted, and potential guests can access all reviews up to that date.

The reason why we do not use only the SID – but only as a robustness check – is because it requires a perfectly balanced panel, and therefore, we would get the same problem of estimating the effect for mainly well-established hosts.

3.4.1 OLS Estimation

To measure the magnitude of ethnic disparities in our observational setting, we rely on the simplest econometric approach proposed by the discrimination literature throughout the years (Blank et al., 2004, Edelman and Luca, 2014).

Specifically, we run a P-OLS regression with ethnicity dummies as our main independent variables, with White hosts as our omitted category. The two main dependent variables are occupancy rates and (log) prices. We run six different specifications with increasing controls to assess the sensitivity of the main ethnic coefficient when additional control variables.

The equation takes the following form:

$$Y_{ijt} = \alpha + \beta E_i' + \delta_1 X_{it}' + \delta_2 Q_{jt}' + \delta_3 Z_i' + \delta_4 W_j' + \gamma_j + \theta_t + \varepsilon_{ijt} \tag{3.1}$$

The outcome variable (Y_{ijt}) is the occupancy rate or the log price per night in US$ of host i from their Airbnb listing j at month t. The vector E_i' is composed of our three main independent dummy variables, which indicate the ethnicity of the host (i.e., Asian, Black, or Hispanic; White is the omitted category). Therefore, β in equation (1) indicates by how many percentage points the average occupancy rates and price of listings offered by Asian, Black, or Hispanic hosts differ from that of White hosts. The vector X_{it}' comprises time-variant host controls, while the vector Q_{jt}' is composed of time-variant listing controls. Moreover, the vector Z_i' comprises time-invariant host controls, while the vector W_j' is composed of time-invariant listing controls. We also include neighborhood (γ_j) and time (θ_t) fixed effects. We do not include listing fixed effects in this model since our main independent variables (i.e., the ethnicity of each Airbnb host) are time-invariant, and fixed effects would cancel out its effect. Finally, ε_{ijt} denotes the idiosyncratic disturbance term. Standard errors are clustered at the NTA level.

3.4.2 Difference-in-Differences Estimation

This paper assesses the short-term impact of the Airbnb anti-discrimination design transformation implemented in October 2018. The changes concerned reducing the

We do not take the log of the occupancy rate variable, since there are many listings with an occupancy rate of zero in some months.

displayed size of users' profile pictures on Airbnb users' pages from 250 to 104 pixels. To evaluate the impact of this transformation, we use two-way fixed effects models in a Difference-in-difference design. This model allows us to estimate the interaction effect of being from any ethnic minority on outcomes before and after this design change, controlling for any time-invariant and variant listing characteristics.

We highlight that this setting does not meet the requirements of a standard difference-in-differences. Indeed, the Airbnb design transformation affects all hosts, and no unaffected control group exists. Without a control group, we cannot estimate how the policy influenced the overall occupancy rate and hosts' prices through Airbnb user usage variations. Only in the scenario of a static demand, where only guest composition affects occupancy rates for each ethnic group, would our estimate correctly identify the overall effect of the Airbnb anti-discrimination policy.

However, the recent literature has proven that, under certain assumptions, our model identifies how heterogeneous the policy's effect is between White and ethnic minority hosts (Shahn, 2023). A negative policy coefficient still demonstrates adverse policy outcomes, implying that the policy effects are greater for White hosts than minority hosts.

As in the standard DiD, there are two key identifying assumptions: (1) parallel trends (i.e., there are no trend differences between ethnic minorities and white hosts before the policy) and (2) no anticipation (i.e., hosts did not know or did not react to the policy before she was implemented). We discuss them extensively in section 3.6.1.

We only include the first six months before and after October 2018 to have a comparable period around the policy implementation date. The sample size decreases to 10,113 hosts and 12,633 listings. The difference-in-differences equation takes the following form:

$$Y_{ijt} = \alpha + pp_t + \lambda E_i * pp_t + \delta_1 X'_{it} + \delta_2 Z'_{jt} + \gamma_j + \theta_t + \varepsilon_{ijt} \tag{3.2}$$

where pp_t stands for post-policy and is a dummy variable that is unity for months after October 2018 and zero otherwise. As in the P-OLS, the vector E'_i is composed of our three main independent dummy variables, which indicate the ethnicity of the host, but now we interact it with pp_t. The interaction of both terms, our coefficient of interest (λ), indicates whether Asian, Black, and Hispanic hosts were

impacted differently by the policy than White hosts. The vectors X'_{it} and Z'_{jt} comprise exogenous time-variant host and listing controls, respectively. The (θ_t) indicates time-fixed effects, while the γ_j is a set of listing dummies (i.e., listing fixed effects) The key parameter of interest λ captures how the policy impact differs between ethnic minorities and white hosts.

3.4.3 Event Study Estimation

In addition, we explore the dynamics of the policy effect using an event study (ES) approach. This approach provides a more detailed analysis of the immediate and longer-term effects of the policy change on the occupancy rate and price disparity between minority hosts and whites. In practice, we estimate the following equation:

$$Y_{ijt} = \alpha + \beta E'_i + \sum_{k=-6}^{6} \lambda_k \cdot \text{pp}_t^k \cdot E'_i + \delta_1 X_{it} + \delta_2 Z_{jt} + \gamma_j + \theta_t + \varepsilon_{ijt} \tag{3.3}$$

The specification includes 6 pre-policy effects ($\beta - 1$, $\beta - 2$, ...,$\beta - 6$) and 6 post-policy (lag) effects ($\beta + 1$,$\beta + 2$, ..., $\beta + 6$) capturing the differential occupancy rate trend between Black and White groups for each month from April 2018 to March 2019. If the design transformation increases the occupancy rate disparity between Black and White hosts, the post-policy β_t's coefficients will be positive. The model includes monthly fixed effects θ_t and listing fixed effects μ_j to capture time-invariant listing specific factors. Finally, as previously defined, our regression incorporates the time-variant covariates vector X_{it} and Z_{jt}. ϵ_{ijt} denotes the idiosyncratic disturbance term. Standard errors are clustered at the listing level. This approach's advantage is that the interactions of post-treatment time dummies with the ethnicity indicator reflect the dynamics of the occupancy rate disparity after the design change. The lag coefficients indicate whether the treatment effect diminishes, remains constant, or grows over time.

We use the fixed effects estimator (within estimator). Still, the notation with dummies is simpler, and the within and least squares dummy variable estimator results are identical.

3.5 Ethnic Disparities on Airbnb

We begin by assessing the underlying ethnic disparity in occupancy rates and prices that the anti-discrimination policy attempted to mitigate. How does the ethnicity of the hosts impact the occupancy rates and overnight prices of Airbnb listings? Table 3.5.1 provides multivariate regression estimates of the main effects of host ethnicity on listings' occupancy rates and (log) prices per night for several model specifications. We define White hosts as the omitted category. Therefore, we interpret each ethnicity coefficient in the regression as the residual impact with respect to White hosts. The first column (Model 1) reports the raw differential in occupancy rate and daily log prices without controlling for differences in observable listings and host characteristics. We observe a significant negative impact on both outcomes of being from any ethnic minority compared to White hosts. This is most pronounced for Black hosts, where the gap in occupancy rates is 13.2% and in prices is 27.1%. For Asian hosts, the occupancy rates gap is 4.4%, and the price gap is 13.3%. Hispanic hosts have the lowest gap in both outcomes, at 3.7% and 8.4%, in occupancy rates and prices, respectively.

A major source of heterogeneity across listings is their geographic location. Also, the time of the year explains much of the occupancy rate and price variation. To account for these variations, Model 2 includes dummies for the Neighborhood Tabulation Areas (NTA) where the listing is located, time dummies and other geographic controls. Including geographic and time controls reduces the residual ethnic occupancy rate and price gaps for all minorities. For Black Hosts, the occupancy rate gap reduces from 13.2% to 9.1%; interestingly, the residual price gap for Black Hosts reduces to 1.5% and is not significant anymore. However, for Asian Hosts, the occupancy rate gap reduces from 4.4% to 2.8%, and the residual price gap reduces to 7.3%, remaining significant at the 1% level. For Hispanic Hosts, the residual occupancy rate gap reduces to 1.7% and the residual price gap These results suggest the geographic position and time of the year almost fully explain the price differentials for Black Hosts, but not their occupancy rate gaps.

Controlling for listing observable characteristics further reduces the residual occupancy rates and price gap for all ethnic minorities (Model 3). Missing information on the number of bathrooms, bedrooms, and beds in some listings slightly

reduces the number of observations. The residual occupancy rate gap shrinks to 8% for Black Hosts, 2% for Asian Hosts, and 2.2% for Hispanic Hosts. The residual price gap reduces significantly for Hispanic hosts (0.4%) and turns insignificant. The price gap for Asian hosts reduces to 1.9% but remains statistically different from zero at the 1% level.

Models 4 and 5 factor in Host characteristics and Guest review controls, respectively. Observations drop further as new listings have no review information. The inclusion of this set of control variables reduces the residual occupancy rate gap to 7% for Black Hosts, 1.4% for Asian Hosts, and 1.8% for Hispanics, with all coefficients at the 1% significant level. The residual price gap is 1.6% for Asian Hosts, at the 1% significant level. For Black and Hispanic Hosts, the price gap remains close to zero and not significant.

In Model 6, we include the log prices and occupancy rates. Due to endogeneity concerns, we only add prices and occupancy rates in the last step. Results are similar to those of Model 4 for both outcomes. Only for Asian Hosts, controlling for prices increased the residual occupancy rate gap to 1.7% instead of reducing it. The price coefficients in the regression (not reported) indicate that if prices increase by one percentage point, the predicted occupancy rates fall by approximately 20%, at the 1% significant level. On the other hand, for Black Hosts, controlling for occupancy rates increases the residual price gap to 1.4%.

The finding that minority hosts maintain comparable pricing to White hosts despite lower occupancy rates is counter-intuitive. Typically, in a competitive market, the rule of supply and demand would predict suppliers to lower their prices when facing a lower demand. We provide two potential explanations. First, the ”Smart Pricing option” Airbnb provides to its hosts may play a role. This option suggests a market price that does not consider guest discrimination. The price suggestions are given based on the local area demand and the characteristics of the listings rather than individual host attributes. This implies that Black and White hosts with comparable listings will receive the same price suggestion if their listings are in the same neighbourhood. The second explanation considers the information asymmetry among hosts. It is plausible that Airbnb hosts are unaware of the occupancy rates of their White competitors, as they may observe only their prices.

To contextualize the implications of our findings, we estimate the annual revenue

Table 3.5.1: P-OLS Results for Impact of Host Ethnicity on Listings' Occupancy Rates and Log Prices

	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 4 (4)	Model 5 (5)	Model 6 (6)
Occupancy Rate						
Asian	-0.044***	-0.028***	-0.020***	-0.016***	-0.014**	-0.017***
	(0.007)	(0.006)	(0.004)	(0.004)	(0.006)	(0.004)
Black	-0.132***	-0.091***	-0.080***	-0.072***	-0.070***	-0.063***
	(0.009)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)
Hispanic	-0.037***	-0.025***	-0.022***	-0.018***	-0.018***	-0.014***
	(0.006)	(0.005)	(0.004)	(0.004)	(0.006)	(0.004)
Observations	574,316	574,316	571,287	502,475	274,210	398,228
Adjusted R^2	0.015	0.107	0.160	0.170	0.193	0.223
(Log) Price per night						
Asian	-0.132***	-0.073***	-0.019**	-0.019**	-0.016**	-0.019***
	(0.021)	(0.013)	(0.007)	(0.008)	(0.007)	(0.007)
Black	-0.271***	-0.015	0.005	0.004	-0.002	-0.014*
	(0.034)	(0.016)	(0.008)	(0.009)	(0.009)	(0.009)
Hispanic	-0.084***	-0.017*	-0.004	-0.005	-0.004	-0.007
	(0.015)	(0.010)	(0.006)	(0.006)	(0.007)	(0.007)
Observations	574,137	574,137	571,108	502,318	398,228	398,228
Adjusted R^2	0.020	0.310	0.728	0.730	0.758	0.767
Time Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Neighborhood Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Geographic Controls	No	Yes	Yes	Yes	Yes	Yes
Listing Controls	No	No	Yes	Yes	Yes	Yes
Host Controls	No	No	No	Yes	Yes	Yes
Review Controls	No	No	No	No	Yes	Yes
Prices/Occupancy rate	No	No	No	No	No	Yes

Notes: The table reports pooled OLS results where the dependent variables are occupancy rates and log price per night. The independent variables are dummies that indicate the effect for Asian, Black, or Hispanic hosts. The omitted category is White hosts. Model 2 adds time dummies, neighbourhood fixed effects, and geographic controls. Model 3 adds listing characteristics, and Model 4 adds host controls. Model 5 adds review controls. Finally, Model 6 adds the prices and occupancy rates, respectively. Standard errors are clustered at the PUMA level. $***p < 0.01, **p < 0.05, *p < 0.1$. *Source*: Author's calculations.

disparity experienced by Black Airbnb hosts relative to their White counterparts. This calculation aims to quantify the potential earnings deficit for Black hosts over a year, attributable to guests' ethnic preferences. Our analysis indicates that, on average, Black hosts earn approximately $4,110\$$ less annually than White hosts. While results regarding Asian hosts show a much smaller magnitude, varying with the analytical model employed, it is still feasible to estimate their annual

revenue shortfall using a similar approach. In contrast, our analysis does not yield substantial evidence to suggest a similar revenue impact for Hispanic hosts, as their occupancy rates and pricing appear comparable to White hosts.

3.5.1 Mechanisms

Our analysis revealed a persistent residual ethnic gap in occupancy rates, which remained unexplained by the geographic location, time of the year, and observable characteristics of listings and hosts. Conversely, the ethnic price gap notably reduces when more observable characteristics are incorporated into the econometric model.

This section delves into potential mechanisms driving the residual occupancy rate disparity. Should statistical discrimination by guests be a contributing factor, we expect to find stronger disparity for newer listings or those with fewer reviews. Without feedback from prior guests, new visitors might default to relying on ethnic stereotypes to assess listing quality. Conversely, taste-based discrimination may be more influential in settings requiring increased host-guest interaction, such as private rooms and shared accommodations. In this context, guests with prejudicial attitudes towards ethnic minorities will incur a disutility from staying with minority hosts, and they will be willing to pay a higher price for staying with whites.

Another dimension to consider is that hosts might also discriminate against guests. If minority hosts discriminate and refuse bookings at higher rates than White hosts, their occupancy rates would be lower. Regarding this point, the seminal paper of Edelman et al., 2017 reveals similar rates of rejection across host ethnic groups, type of property, and gender. To investigate this aspect further, we restrict our analysis to hosts who have enabled the Instant Booking feature, ensuring automatic guest acceptance and eliminating the potential for selective booking acceptance from hosts.

We test for these mechanisms by performing heterogeneity analysis on several sub-samples according to the number of reviews (below median vs. above median), the nature of the listing (entire property vs. private/shared property), and the instant booking option (activated vs. not activated). For each sub-sample, we report

The estimated annual revenue shortfall for hosts belonging to ethnicity i is determined by the formula: $Rev_i = 365[(APW AOW) - (APW(1 - EPD_i) * (AOW - EOR_i))]$, where APW represents the average listing price of White hosts, AOW denotes the average occupancy rate of White hosts at this threshold, EPD_i is the calculated price differential between White hosts and hosts of race i as per Model 5, and EOR_i indicates the estimated occupancy rate differential for the same groups.

the residual occupancy rate gap estimates for each ethnic minority. White hosts are the reference group, and their coefficient is not reported. The baseline model is Model 4, specified in the previous section.

Table 3.5.2 reports results from the heterogeneity analysis, suggesting that a mixture of taste-based and statistical discrimination lies at the core of the residual ethnic disparity in occupancy rates. In the sub-sample of listings with a lower than median number of reviews (Column 2), i.e., less than 9 reviews, the residual occupancy rate disparity increases from 7.2% to 7.8% for Black hosts, slightly for Asians but not for Hispanics. Conversely, In the sub-sample of listings with a higher median number of reviews (Column 3), the residual occupancy rate disparity reduces to 5.9% for Black Hosts. These results highlight a stronger and positive impact of the number of reviews for Black Host, but not for other ethnic minorities. Next, we stratified the sample by different types of property, i.e. entire apartments or private/shared rooms, with less and more host-guest interaction intensity, respectively. Results reveal that when the property requires more interaction (Column 5), the occupancy rate disparity, again mainly for Black Hosts, increases from 7.2% to 8.3%. With less interaction (Column 4), Black hosts' disparity reduces to 6.6%. Lastly, by splitting the sample among hosts with and without instant booking, we find that the occupancy rate disparity shrinks when the instant booking is active (Column 6), but not enough to suggest an effect of discrimination from the Hosts.

Another mechanism that could explain the residual ethnic disparities is ethnic matching between guests and hosts. Ethnic matching arises because attitudes, customs, tastes, and values, which often foster friendships, are frequently associated with, or even rooted in, ethnicity (Leszczensky & Pink, 2019). Consequently, if hosts' listing demand depends solely on their own and their guests' ethnicity, hosts are not competing in the same market. To investigate this mechanism, we scraped the URLs of guests' profile pictures for those guests who left a review on the listing page. We classified the ethnicity of guests for a smaller subset of the listings (7,711). This reduction in sample size was due to the time elapsed between the main analysis and this additional analysis conducted in 2023. Therefore, we had to focus only on listings in the main sample still active in 2023. This corresponds to around 12% of the main sample of 2016-2019.

For each listing, we regress the share of reviews written by guests of a given eth-

nicity on a dummy for the host ethnicity, controlling for the location, the observable characteristics of the listing and the host, and its price. Table 3.5.3 illustrates the result. We find evidence for some ethnic matching, especially among black hosts and guests: a host classified as Black is 4 percentage points more likely to have a review from a guest also classified as Black. Conversely, the share of White guests negatively correlates with a host being Black and a host being Asian. This means that, on average, compared to White hosts, Black and Asian hosts receive fewer reviews from White guests compared to their White host counterparts.

Table 3.5.2: Heterogeneity analysis for Impact of Host Ethnicity on Listings' Occupancy Rates

	Baseline	Below Median	Above Median	Entire H/A	Private/Shared Room	Instant Booking
	(1)	(2)	(3)	(4)	(5)	(6)
Asian	-0.016***	-0.018*	-0.017**	-0.020***	-0.015*	-0.024**
	(0.004)	(0.004)	(0.009)	(0.006)	(0.006)	(0.007)
Black	-0.072***	-0.078***	-0.059***	-0.066***	-0.083***	-0.060***
	(0.006)	(0.013)	(0.009)	(0.008)	(0.008)	(0.010)
Hispanic	-0.018***	-0.012	-0.012*	-0.019***	-0.015*	-0.017*
	(0.004)	(0.004)	(0.010)	(0.004)	(0.006)	(0.007)
Observations	502,475	90,731	182,524	251,732	237,539	168,241
Adjusted R^2	0.170	0.229	0.206	0.170	0.193	0.214

Notes: The table shows heterogeneous effects of hosts' ethnicity on occupancy rate disparity. The independent variables are dummies that indicate the effect for Asian, Black, or Hispanic hosts. The omitted category is White hosts. The baseline model refers to Model 4 in Table 1. $***p < 0.01, **p < 0.05, *p < 0.1$. *Source*: Author's calculations.

Table 3.5.3: Ethnic matching between guests and hosts

	Share of White Guests	Share of Asian Guests	Share of Balck Guests	Share of Hispanic Guests
Asian	-0.019**	0.013*	-0.000	0.002
	(0.009)	(0.007)	(0.004)	(0.007)
Black	-0.036***	-0.005	0.040***	0.000
	(0.011)	(0.007)	(0.006)	(0.006)
Hispanic	-0.008	-0.000	0.005	0.005
	(0.008)	(0.006)	(0.005)	(0.006)
Observations	32,851	32,851	32,851	32,851
Adjusted R^2	0.071	0.028	0.115	0.008

Notes: OLS regression of share of guests from a given ethnicity on host ethnicity dummies. Controls include neighbourhood FE, geographic controls, property characteristics, and host characteristics. Standard errors are clustered at the property level. $***p < 0.01, **p < 0.05, *p < 0.1$. *Source*: Author's calculations.

3.6 Evaluating Airbnb Anti-Discrimination policy

So far, we have empirically identified the extent of the ethnic disparity in two main outcomes: occupancy rates and prices per night. In this section, we focus on the short-run effects of the 2018 Airbnb anti-discrimination policy, which aims to reduce such ethnic differentials.

Figure 3.6.1 shows the standard difference-in-differences coefficient estimates for occupancy rates and prices. As before, we treat White hosts as the control group and use it as the reference category. In other words, we compare the evolution of the occupancy rates and prices between each ethnic minority and White hosts before and after the design transformation.

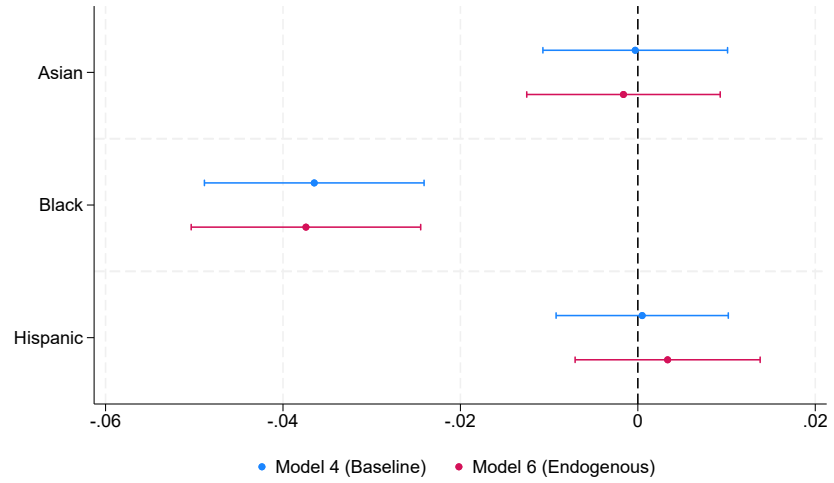
Panel (a) indicates an increase of 4 percentage points in the occupancy rate gap of Blacks with respect to Whites. The coefficient is statistically significant at the 5% significance level. For Asians and Hispanics, the coefficients are close to zero and are not statistically significant. Panel (b) illustrates the coefficients for (log) price per night. The estimates indicate no change in price differential between ethnicity's post-policy intervention. This null effect on prices supports our previous findings. Hosts do not adjust their prices in response to reduced demand. The residual occupancy rate gap likely results from a change in guest booking behavior due to the picture dimension shrinkage.

Next, we look at the policy effect dynamics for occupancy rates and prices six months before and after implementation. Figure 3.6.2 shows the results of the DiD ES approach (i.e., the evolution of the occupancy rate and price gap between each minority and White hosts before and after the policy intervention date). Only for Black hosts (panel c), the estimates show a persistent negative effect on occupancy rates in the six months after the policy date. As already anticipated, the figure clearly shows that one month before the policy date (i.e., April 2018), the parallel trend assumption does not hold, and the estimate for Black hosts is significantly different from zero. For the other ethnic groups, there is no dynamic, and the estimates remain close to zero and not significant. The dynamic appears flat for prices, implying the price is not responsive to the new design.

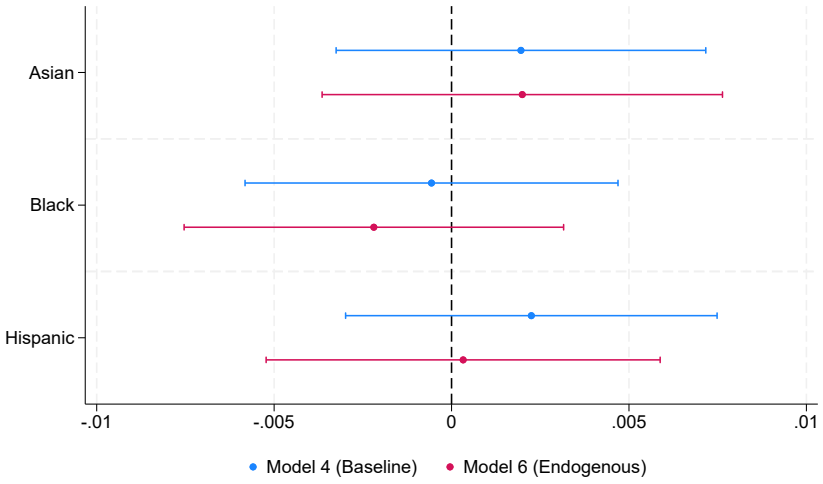
As we have some pre-trends, we also report the SDiD estimates. Figure 3.6.3 shows our SDiD ES approach estimates. The results are similar to our standard DiD ES. After October 2018, the occupancy rate gap between Black and White hosts increased substantially, as shown in panel b. For Asians and Hispanics (panels a and c), the effect remains nearly zero and is mostly insignificant.

One possible explanation for the smaller Black-White occupancy rate gap in December 2018 could be that NYC is very popular for spending Christmas and New-Years-Eve. Therefore, guests might have booked an Airbnb way in advance to guarantee their stay in NYC over the holidays.

Figure 3.6.1: DiD - Impact of Airbnb Policy on Listings' Occupancy Rates and Log Prices



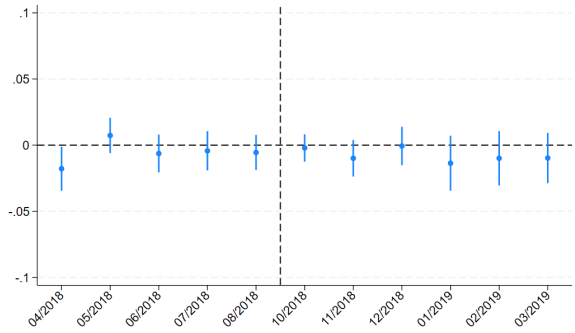
(a) Occupancy rates



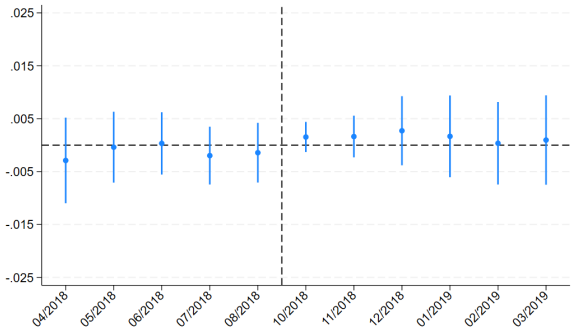
(b) Log Prices

Note: The figures report the estimates from the DiD estimation strategy for occupancy rate (panel a) and log prices (panel b). The independent variables consist of interactions between host ethnicity dummy variables and a post-policy dummy variable, which takes the value of unity for months following the implementation of the anti-discrimination policy. The omitted category is White hosts. The baseline model includes all time-variant controls used in Model 4 of the P-OLS. The endogenous model includes prices and occupancy rates, respectively. Standard errors are clustered at the NTA level.

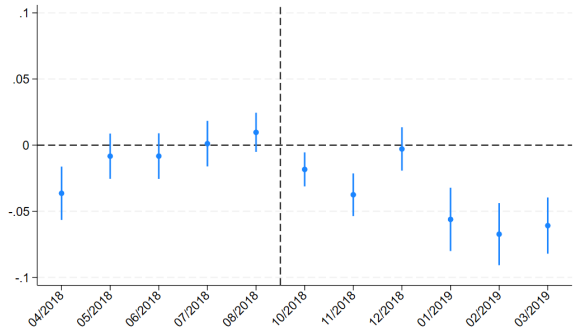
Figure 3.6.2: Event-Study - Impact of Airbnb Policy on Listings' Occupancy Rates and Prices



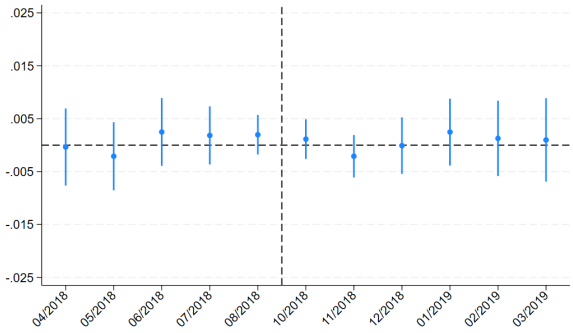
(a) Occupancy Rates (Asian Hosts)



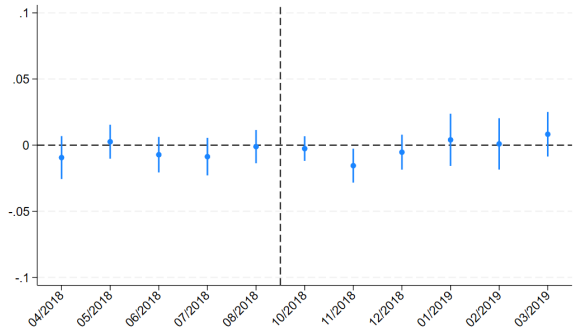
(b) Log Prices (Asian Hosts)



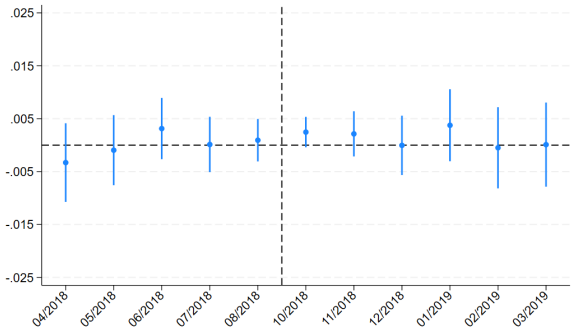
(c) Occupancy Rates (Black Hosts)



(d) Log Prices (Black Hosts)



(e) (Occupancy Rates (Hispanic Hosts)

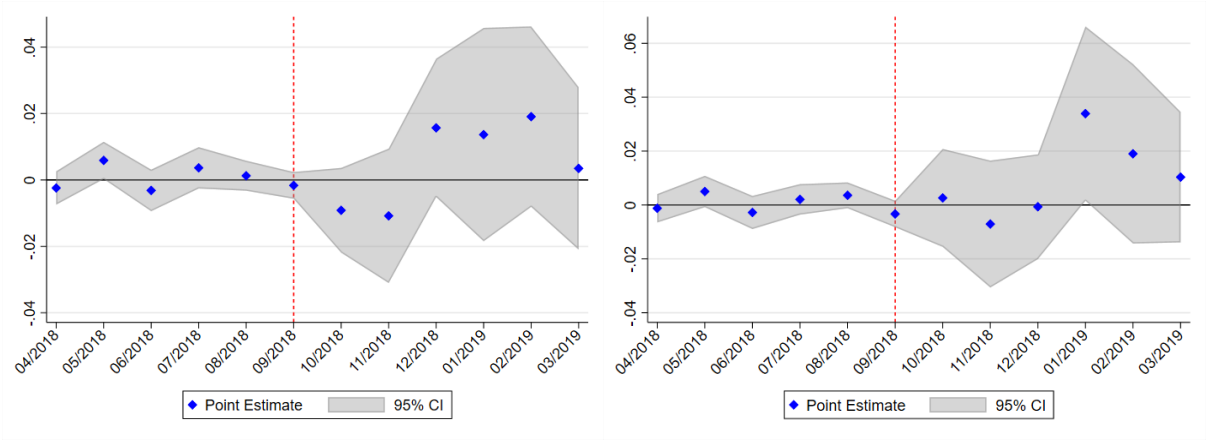


(f) Log Prices (Hispanic Hosts)

Note: ES estimates of the Airbnb policy on occupancy rates for the three ethnic groups. 90% confidence interval. Standard errors clustered at the NTA level. Dotted vertical lines: Airbnb policy starting date.

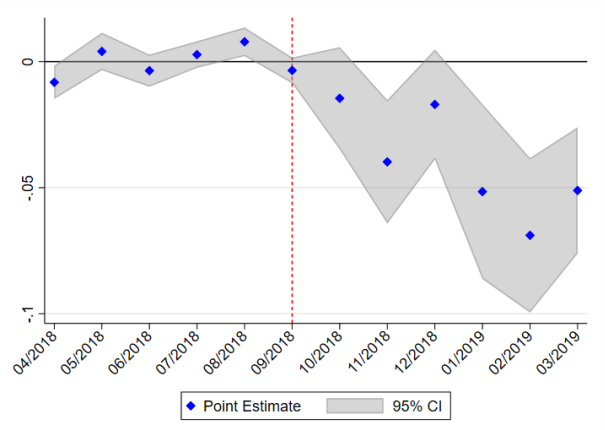
Source: Authors own calculation.

Figure 3.6.3: Event.Study - Impact of Airbnb Policy on Listings' Occupancy Rates



(a) Asian hosts

(b) Hispanic hosts



(c) Black hosts

Note: SDiD Event Study estimates of the Airbnb policy on occupancy rates for the three ethnic groups. 90% confidence interval. Standard errors clustered at the NTA level. Dotted vertical lines: Airbnb policy starting date.

3.6.1 Difference-in-Differences assumptions

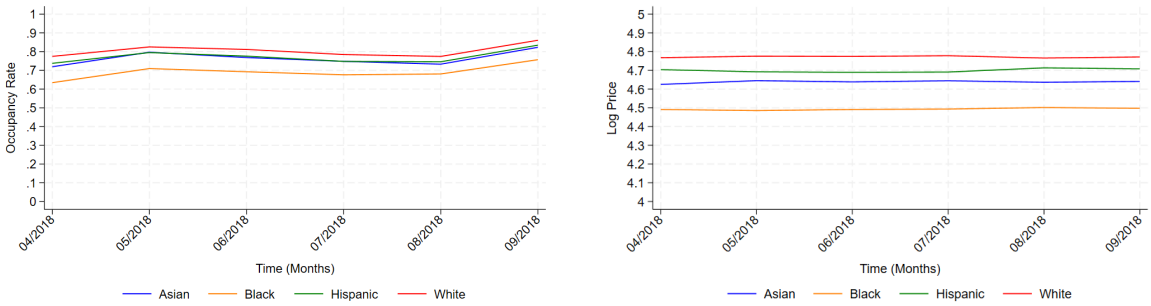
The two main assumptions of the DiD approach are (i) no pre-trends and (ii) no anticipation. This means that, in the pre-policy period (6 months in total), the trend in outcomes variables for ethnic minority and White hosts should follow parallel trends, and hosts should not anticipate any effect of the policy and adapt to it before its actual implementation. As shown in Figure 3.6.4, just visually, it appears that all the ethnicities have similar trends in occupancy rates.

However, Table 3.6.1 shows the joint parallel trend test does not hold for Asian and Black hosts. As previously mentioned, due to the existence of some pre-trends, we use the SDiD (Arkhangelsky et al., 2021) as a robustness check. For (log) prices,

the parallel trend assumption holds at each point (see Figure 3.6.4 Panel b).

To check for anticipation effects, we look at rates of profile picture changes in the period before the policy. Figure 3.6.5 shows the share of hosts that change their profile picture each month. This share is generally very low, with no substantial differences across ethnicity. We interpret this as supporting evidence for no anticipation of the policy. If hosts expected to be negatively impacted by the new design, they should have uploaded a new and clearer profile picture.

Figure 3.6.4: Parallel Trends by Ethnicity



(a) Occupancy Rates

(b) Log Prices

Note: Line plot for the monthly occupancy rates and prices from April 2018 to September 2018, for each ethnic group.

Source: Authors own elaboration

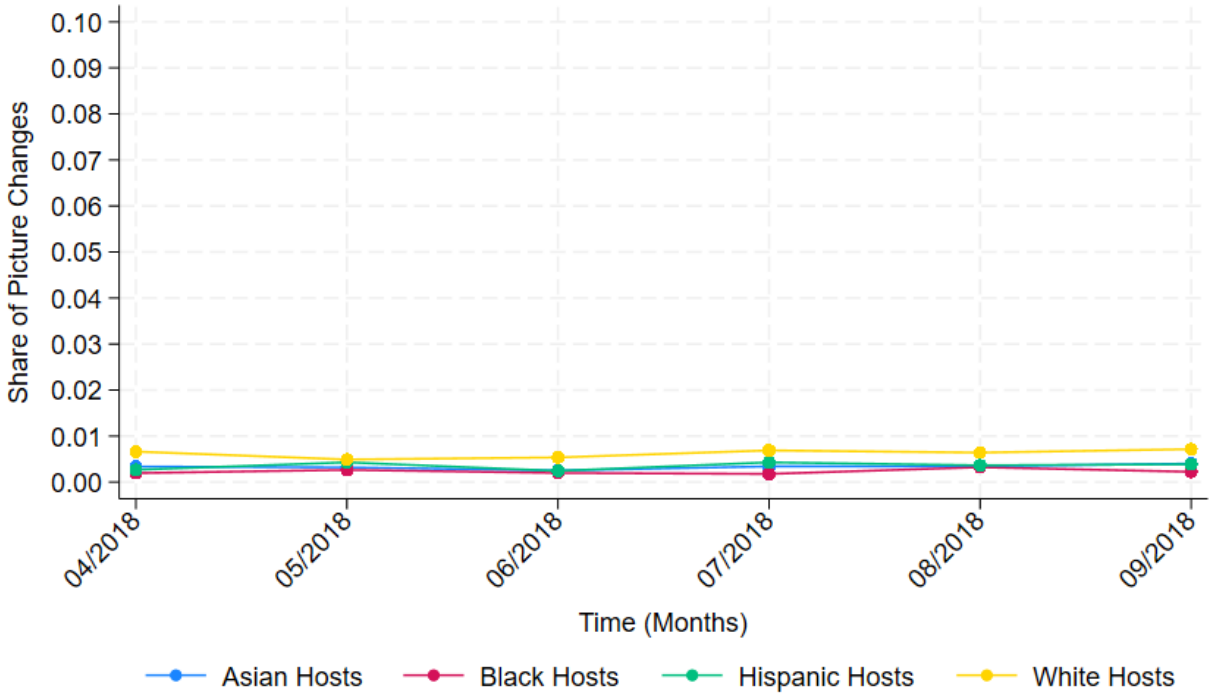
Table 3.6.1: Parallel Trends Test by Ethnicity

Ethnicity	**F-stat (Occ. Rate)**	**Prob > F (Occ. Rate)**	**F-stat (Log Price)**	**Prob > F (Log Price)**
Asian	2.56	0.025	0.360	0.873
Black	3.93	0.002	0.790	0.560
Hispanic	1.03	0.398	1.070	0.376

Note: The table reports the results of a joint parallel trends test for subgroup parallel trends

Source: Authors calculation

Figure 3.6.5: No Anticipation



Source: Authors own elaboration.

3.6.2 Heterogeneity Analysis

The Airbnb anti-discrimination policy increases the occupancy rate disparity between Black-White hosts. While exploring the mechanism in section 3.5, we have found mild evidence that driving the ethnic disparities is a mixture of taste-based, inaccurate statistical, and screening discrimination by guests. Black hosts with a lower than median number of reviews and who offer private rooms instead of entire apartments suffer greater penalties in occupancy rates. Therefore, we also hypothesize that if the policy had any effects, it would mainly target these hosts.

We perform heterogeneity analysis to identify the sub-samples of hosts for which the policy had the greatest impact. Table 3.6.2 shows the difference-in-differences estimates for different sample strata based on the number of reviews (below vs. above median) and the property type (entire apt. vs. private rooms). We found minimal policy effect differentials among these groups. Contrary to our expectations, the policy appears to have increased the Black-White disparity more for Black hosts who offer entire apartments rather than, as expected, private rooms, which involve

more hosts-guest interactions.

Table 3.6.2: Heterogeneity analysis for the impact of Airbnb Policy on Occupancy Rates

	Baseline	Below Median	Above Median	Entire H/A	Private/Shared Room
	(1)	(2)	(3)	(4)	(5)
Asian	-0.000	0.005	-0.000	-0.012	0.011
	(0.005)	(0.008)	(0.008)	(0.007)	(0.008)
Black	-0.036***	-0.038***	-0.030***	-0.056***	-0.018
	(0.006)	(0.013)	(0.009)	(0.008)	(0.008)
Hispanic	0.000	-0.003	0.006	-0.002	0.006
	(0.005)	(0.007)	(0.007)	(0.006)	(0.008)
Observations	100,233	45,950	47,022	51,865	46,285
Adjusted R^2	0.114	0.079	0.166	0.127	0.109

Notes: The table shows heterogeneous effects of hosts' ethnicity on occupancy rate disparity. The independent variables are dummies that indicate the effect for Asian, Black, or Hispanic hosts. The omitted category is White hosts. The baseline model refers to controls used in Model 4 in Table 1. $***p < 0.01, **p < 0.05, *p < 0.1$.

Source: Author's calculations.

3.6.3 Additional Outcomes

Thus far, our focus has primarily been on occupancy rates and prices. Our empirical findings reveal a statistically significant occupancy rate disparity between Black and White hosts, which increased after Airbnb's new design. Surprisingly, we observe minimal price disparities between hosts, unchanged following the policy intervention. However, beyond prices, hosts on the platform can modify certain elements attached to their listings, such as their profile pictures, self-descriptions in their bios, and the range of amenities offered in the property. Conversely, other features remain exogenous, set automatically by the platform or determined by guests' feedback, hence not modifiable directly from the hosts.

Hosts may opt to adjust the more flexible, less cost-intensive features in response to shifts in occupancy rates rather than reduce prices and potentially impact their overall profits. To explore this hypothesis, we estimate the DiD (ES) for a further outcome, the log of the number of listed amenities associated with each listing in a

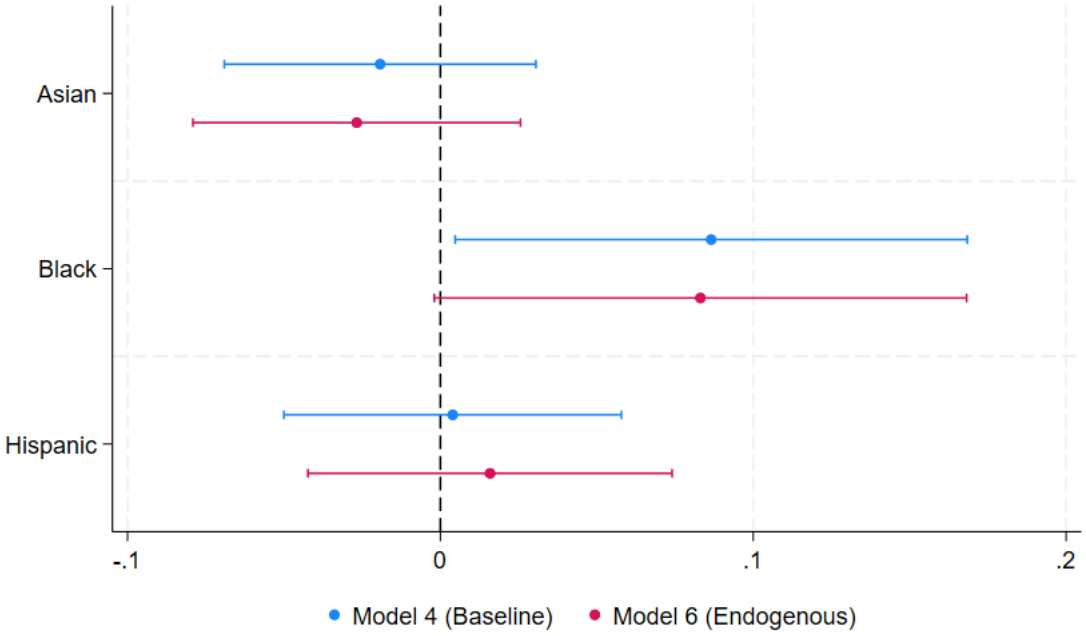
given month.

Figure 3.6.6 reports the DiD (ES) results for the number of amenities. We found a significant positive effect on the number of amenities offered. Specifically, there's a marked increase of about 4 percentage points in the number of basic amenities provided by Black hosts, a change not observed among Asian and Hispanic hosts. This finding suggests a proactive strategy by Black hosts to enhance their listings' appeal after facing a drop in demand. Figure 3.6.8 (panel b) illustrates these trends through an event study estimation, showing the dynamics of the estimated coefficients until six months after the policy change.

Another variable that we have thus far not analyzed as an outcome variable is the number of reviews. This variable is indeed an endogenous choice of guests and, therefore, is likely an underestimate of the true number of listing bookings in a given month. However, we can analyze it to observe whether the policy has any short-term impact on the number of reviews received by hosts. Figure 3.6.7 reports the estimate for the DiD estimator. Interestingly, we observe a high and statistically significant negative effect of a host being Black on the number of monthly reviews in the aftermath of the policy implementation.

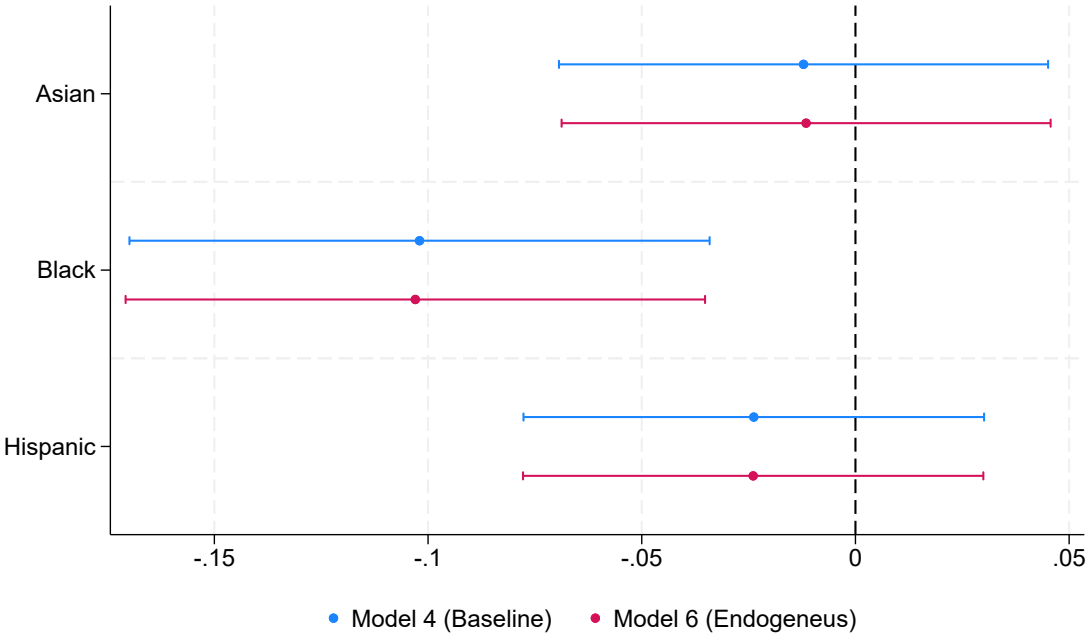
This additional analysis supports the main findings that the Airbnb policy had differing impacts on hosts based on ethnicity. While the occupancy rates gap between Black minority and White hosts increased significantly after the policy change, price disparities remained minimal and unchanged. Black hosts adjusted by increasing the number of listing amenities rather than lowering prices. Finally, the significant negative impact on the number of reviews for Black hosts post-policy further corroborates the increased occupancy rate gap we find.

Figure 3.6.6: DiD - Impact of Airbnb Policy on Number of Amenities



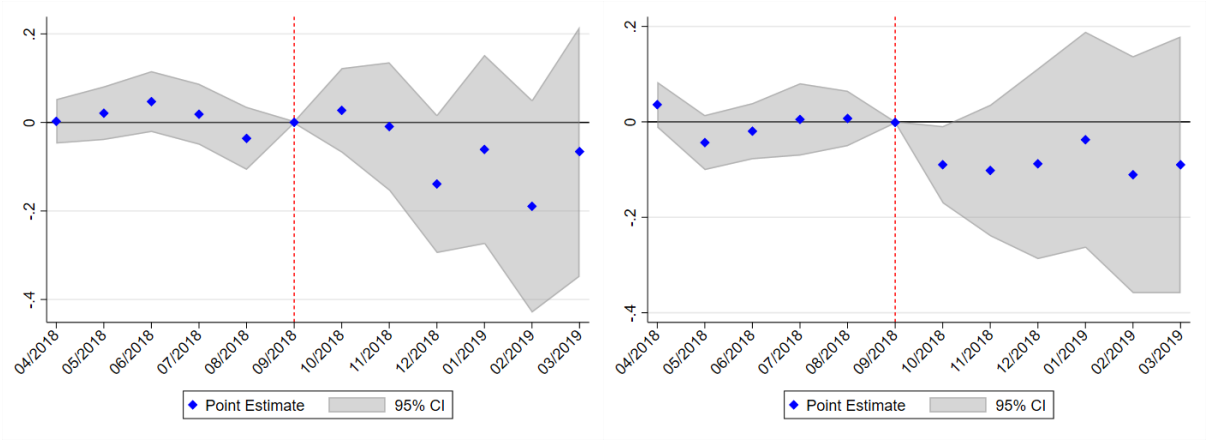
Note: See figure 3.6.1

Figure 3.6.7: DiD - Impact of Airbnb Policy on Number of Reviews



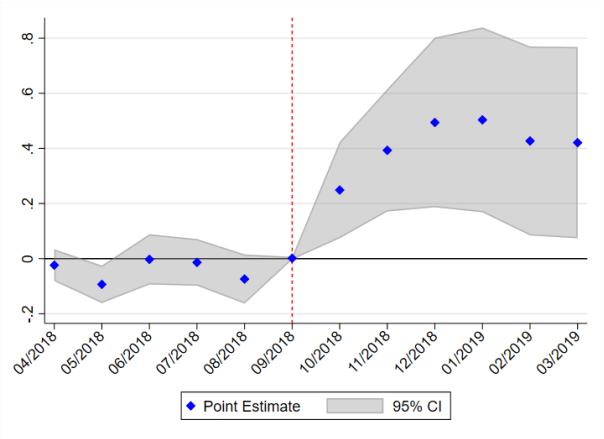
Note: See Figure 3.6.1

Figure 3.6.8: Event.Study - Impact of Airbnb Policy on Listings' Number of Amenities



(a) Asian hosts

(b) Hispanic hosts



(c) Black hosts

Note: SDiD Event study approach estimates the Airbnb policy on the number of amenities for the three ethnic groups. 90% confidence interval based on standard errors clustered at the NTA level. Vertical lines: Airbnb policy starting date.

Source: Authors own calculation.

3.6.4 Mechanisms

The adverse impact of the Airbnb design transformation on the occupancy rate disparity between Black and White hosts warrants a deeper exploration of the underlying mechanisms. After the policy, we hypothesize that guests can still discern a host's skin color, especially darker tones, even with smaller profile pictures. At the same time, smaller profile pictures increase uncertainty and inaccuracy in detecting other facial features that convey trustworthiness, such as a smile. In turn, this increased uncertainty will negatively affect occupancy rates.

Supporting this hypothesis, Ert and Fleischer, 2019 demonstrates that a smile

in a profile picture significantly enhances perceptions of a host's attractiveness and trustworthiness, positively influencing occupancy rates. Also, the pooled regression analysis results, which adjust for observable listing characteristics and geographic variables, reveal that a host's smile positively correlates with occupancy rates by 2.14 percentage points, a statistically significant finding at the 1% level.

In the following, we undertake a series of additional analyses to explore the proposed hypothesis. We aim to simulate the effect of smaller profile pictures on the human ability to detect facial clues. We use our fine-tuned face classification algorithm to predict ethnicity and smiles in images resized to match the dimensions after Airbnb's policy change. We pick the images from the training dataset. We measure the average prediction entropy and accuracy in scenarios with normal-sized and reduced-sized images. We then take the differences between the two scenarios and estimate the p-values.

Prediction entropy is calculated using the Shannon entropy formula, commonly employed in information theory, represented as $H(X) = -\sum_i p(x_i) \log p(x_i)$, where $p(x_i)$ is the probability of feature x_i, i.e., being black or smiling. This metric measures the uncertainty in information extracted from a signal: higher entropy indicates greater uncertainty. For accuracy assessment, we compare the model's predictions against the ground truth stored in the training dataset and take the average number of times the prediction equalizes the true value.

The results in Table 3.6.3 reveal two key findings. Consistent with our hypothesis, prediction entropy increases, and prediction accuracy decreases with smaller pictures across all facial features analyzed. Secondly, and notably, the feature "Black" shows the lowest entropy and highest accuracy, both before and after the reduction in image size. This suggests that our model, designed to replicate human perceptions, is most effective in identifying the "Black" feature among the facial characteristics we studied. While these results support our hypothesis, it is important to note that the algorithm's performance cannot be entirely generalized to human performance.

Table 3.6.3: Prediction Entropy and Accuracy from Airbnb Policy Simulation

	Normal	Compressed	Statistic	p-values
	(1)	(2)	(3)	(4)
Prediction Entropy				
Smile	0.23	0.26	9.01	0.00
Black	0.07	0.15	12.89	0.00
White	0.11	0.16	13.11	0.00
Hispanic	0.31	0.36	6.45	0.00
Asian	0.18	0.26	12.03	0.00
Prediction Accuracy				
Smile	0.90	0.88	-6.06	0.00
Black	0.98	0.92	-8.86	0.00
White	0.97	0.94	-8.11	0.00
Hispanic	0.86	0.82	-3.42	0.00
Asian	0.93	0.87	-7.58	0.00

Note: In information theory, the entropy measures the information uncertainty in a distribution: higher entropy implies higher uncertainty. The accuracy measures the distance of the algorithmic prediction from the human classification. *Source*: Author's calculations.

3.7 Discussion and Robustness Checks

In this section, we examine the robustness of the main results. First, we investigate the potential for omitted variable bias in the OLS regression using the Oster approach, as presented in Oster (2017). Second, we describe additional analysis of confounding factors for the OLS regression. These are profile picture quality and apartment quality. Indeed, these two variables may correlate with the host's ethnicity and the outcome variables, biasing our coefficients. Third, we evaluate the OLS results' sensitivity to different cutoff thresholds in the ethnicity prediction probability. Finally, we provide evidence that what drives the DiD results are not supply or demand shocks on the platform.

Omitted Variables This paper estimates the residual ethnic disparities in two outcomes after controlling for the entire set of observable characteristics from the Airbnb website. This simple statistical approach to measuring discrimination has statistical challenges and limitations.

The gold standard for measuring the causal effect of immutable traits such as ethnicity is to manipulate this trait, or the perception of it, randomly and estimate the effect of such intervention on the outcome. In our setting, however, to interpret β as the causal effect of hosts' ethnicity on the outcomes, we must assume we have included all relevant observable factors correlating with the outcomes and differing systematically across ethnicity. If that is not the case, our estimates will suffer from omitted variable bias. As shown in Table 3.5, the estimated coefficients for ethnic minorities tend to get closer to zero as we add more controls. This raises the question of what would happen to the estimated coefficients if we could add all potentially relevant unobserved controls.

Due to this potential omitted variable bias (OVB), we follow an approach proposed by Oster, 2017, which builds up on an idea proposed by "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools", 2005.

This approach aims to measure how large the effect of unobservable variables would have to be to offset the estimated coefficients under the assumption that the selection of observables is proportional to the selection of unobservables. The parameter δ, which is reported when applying this approach, shows how large the selection of unobservables would have to be to cancel out the estimated impact of

host ethnicity on the outcome variables of interest. For example, $\delta = 2$ would mean that unobservables have to be twice as important as observables to cancel out the estimated coefficients. According to Oster, $\delta = 1$, is an appropriate cutoff to define whether the results are robust or not. The other important information we report is the identified set. If the identified set does not include zero, the estimated coefficient can be considered robust to OVB. For further details about Oster's approach and derivations of δ and the identified set, see Oster, 2017.

The results using Oster's approach are reported in Table 3.G.1. To apply this method, the R-squared from the uncontrolled regression (i.e., the regression where we do not control for anything but the independent variable of interest) needs to be the same for all models compared to the uncontrolled regression. Therefore, we restrict our sample so that only the observations included in Model 4 are used to estimate the coefficients of the other models. Our results show that the estimated occupancy rate coefficients are still robust after applying Oster's approach. As one can see, δ is always above the appropriate cutoff of 1, proposed by Oster. The identified sets also exclude zero, implying that the bias-adjusted coefficients with the selected upper bounds on δ and R_{max} do not change sign significantly relative to the estimated coefficients in the main models. However, the estimated log price coefficients are no longer robust after applying Oster's approach. Nonetheless, δ is always close to 1, and only the identified Black hosts include zero.

Alternative Ethnicity Definitions Alternative definitions of outcomes and ethnicity variables tend to provide very similar results. In all these robustness exercises, the baseline model refers to the OLS regression with geographic controls, time dummies, listings, and host characteristics (model 4).

We first assess the change in the estimated coefficients with different probability thresholds in the ethnicity prediction. Our face classification model returns the probability of a host belonging to each ethnicity. In the main analysis, we assign to hosts the ethnicity with the highest predicted probability. To test the robustness of our results, we adjusted the threshold, considering only classifications where the model's confidence for a given ethnicity exceeded 50%, 60%, 70%, and 80% of

This means that observations of any control variable with missing values are not used in estimations of any model.

The choice of using the highest predicted probability adds a bit of noise to our ethnicity classification, but it maximizes the number of observations.

probability. This adjustment aimed to restrict our sample to hosts for which the confidence for the predicted ethnicity is higher than the given threshold.

Table 3.G.2 illustrates the results for the OLS estimator on occupancy rates. In both cases, coefficients are robust to different specifications. Increasing the threshold around the ethnicity predictions results in slightly higher occupancy rate disparity for all ethnicities, especially for Black hosts. This implies that the more certain our algorithm (and potentially the prospective guests) are about the host's ethnicity, the higher the residual impact of host ethnicity on occupancy rates. This means that our results are rather conservative and, if anything, underestimate the true effect.

As a second robustness test, we substitute the ethnicity dummies with a normalized black/white continuous measure, as proposed in

Face Dimension in Profile Pictures We have shown that hosts do not change their profile pictures in the pre-policy period (see Figure 3.6.5). Nevertheless, it is plausible that systematic differences in profile pictures between ethnicities also existed before the policy. For instance, it could be the case that all White hosts had clearer and more professional profile pictures compared to Black hosts. Such pre-existing disparities might contribute to the observed strong ethnic disparity and the strongest policy's effect on Black hosts.

To investigate this scenario, we analyze each profile picture by measuring the proportion of the total image occupied by the face. In full-body pictures or non-frontal faces, the face percentage will be lower than in cases with frontal and close-up photos. Therefore, we examine whether systematic differences in the face percentage within images existed among ethnic minorities and White hosts before and after the policy. Results from this analysis revealed no systematic difference between Black and White hosts in terms of face percentages.

Listing Quality and Room Types The main analysis in this paper uses the information at the listing and host levels collected from May 2016 to December 2019. A limitation of this dataset is the lack of any information related to the apartment pictures hosts upload on their listings pages. If Black hosts systematically show worst-quality apartments in their listing pictures, the model suffers from an omitted

Results not yet reported.

variable problem.

We collected and analyzed additional listing information in a later period (2023-2024) from the original sample (2016-2019). Likely due to the impact of the COVID-19 pandemic and more restrictive policies for Airbnb imposed by NYC, only 13% of the listings from the original sample remained active during this period. We found no access to any further information about listings that were removed from the platform. Therefore, this additional analysis assesses only descriptively, at present, whether systematic differences exist in room types and overall apartment quality across ethnic groups.

We extracted the raw data from our main dataset by selecting a stratified random sample of 1,000 active hosts, comprising 250 hosts from each ethnic group. Using the listing URL information, we downloaded images of the apartments at the current time. We then fine-tuned an image classification model to identify the types of rooms depicted in each image (i.e., bathrooms, bedrooms, kitchens, living rooms, dining rooms, and exteriors). The model was trained using a pre-classified apartment imagery dataset provided by Poursaeed et al., 2018.

The model achieved an accuracy of 87% on the test sample and 95% on the training sample. Using this model, we classified the room types in each host's apartment imagery. We calculated the ethnic differentials in the likelihood of displaying a picture for each room type. For example, whether a Black is as likely to display a bathroom as a White host. The results of a simple regression analysis, where the outcome variable is a binary indicator of whether a given room type is present and the independent variables include ethnic group dummies (with White hosts as the reference group). This analysis reveals that Black hosts are less likely to display living or dining rooms than White hosts. These room types correlate with the type of apartment (e.g., an entire apartment or a private room), which is a variable we control for. Other types of rooms, such as bathrooms and bedrooms, appear as likely between ethnicity.

In a second descriptive analysis, we assess the quality of the apartments using the same images. Given that apartment quality can be highly subjective, it is challenging to quantify it as straightforwardly as variables such as ethnicity or the presence of a smile in a host's profile picture. Any predictive model would likely

Not yet reported.

produce noisy and imprecise results.

Consequently, we adopted an alternative approach: we fine-tuned a generative text model to describe the content of each image objectively. We then utilized a text classification model to predict the sentiment of these descriptions—whether positive, negative, or neutral. This predicted sentiment serves as a proxy for the quality of the room type depicted in the images. We then compared the average sentiment across hosts of different ethnicities. The results reveal no statistical difference in sentiment across ethnicities.

Supply and Demand In Section 3.4.2, we observe that we are not in a standard DiD setting. The new design was implemented simultaneously across all hosts' ethnicity, so we do not have a control group in the classical sense. This implies that we cannot identify the causal effect of the policy on each ethnic group's demand. However, suppose there is no supply (i.e., the number of listings offered on Airbnb) and demand (i.e., the number of guests on Airbnb) policy effects. In that case, our estimates will identify the policy's overall effect on each ethnic group's demand. In this static case, where all the market conditions stayed the same, our results might reflect a policy effect on the distribution of guests across hosts depending on their ethnicity – for example, a shift of guests from Black to White hosts.

The supply shock might have occurred if an over-proportionally high number of Black Hosts had registered on Airbnb because they thought they would have better chances after the new design. This shock might have created extra competition for Black hosts in a setting of strong segregation and homophily, explaining the increased outcome gap with White hosts. A demand shock due to the policy appears less likely. However, in the case of a considerable increase in the number of people using Airbnb, our causal estimates would not capture the effect of this shock.

We test this hypothesis by graphically assessing the average monthly listings active for each ethnicity and estimating the overall policy responses of occupancy rates and log prices. Figure 3.G.1 in Appendix 3.G shows that the supply of listings offered on the Airbnb platform stayed relatively constant over time and across ethnicity. Excluding the possibility of a supply shock driving our main result. Table 3.G.3 in Appendix 3.G reports that the overall demand (i.e., occupancy rates) is also constant in the period analyzed. This shows that there was no significant demand

Not yet reported.

shock on the platform.

This is supporting evidence that, due to the policy, it is more likely that guests' preferences have changed, and discriminatory behavior toward Black hosts has increased.

3.8 Conclusion

This paper investigates the effects of Airbnb's anti-discrimination policies on ethnic disparities among hosts. We initially identified a notable occupancy rate disparity. Black hosts experience a 7.2 percentage points lower occupancy rate than White hosts, even after adjusting for location and observable property and host characteristics. Moreover, we estimate a 4 percentage point increase in the Black-White occupancy rate disparity following the implementation of the Airbnb anti-discrimination policy. This adverse effect likely stems from the policy's reduction of guests' ability to discern positive facial cues in profile pictures while still permitting recognition of skin color, disproportionately disadvantaging Black hosts. In an adaptive response, Black hosts appear to enhance the amenities offered in their listings. Asian and Hispanic hosts are not affected by the anti-discrimination policy in any direction.

Several critical insights and a potentially serious dilemma for platform designers and policymakers emerge from our study. First, reducing profile picture prominence has amplified ethnic gaps in occupancy rate. Limiting the information available to guests to assess prospective hosts has potentially increased reliance on statistical or inaccurate statistical discrimination. This type of discrimination is partially driven by guests' imperfect information regarding the quality of properties, a situation exacerbated by the platform's feedback system, which disadvantages new hosts with few reviews. Profile pictures are crucial in countering these biases by offering insights into hosts' reliability and trustworthiness. Therefore, increasing the transparency and breadth of information provided to guests, including through profile pictures, could help mitigate ethnic disparities by addressing and correcting beliefs.

This study contributes to the literature concerning discrimination in digital platforms, underscoring the challenges and potential strategies to enhance inclusivity and diversity. As online interactions become increasingly prevalent, our findings

advocate for a balanced approach to information disclosure that fosters trust while avoiding the reinforcement of biases. This approach could prove beneficial across diverse digital platforms. Future research should examine the effective implementation of such strategies and assess their wider effects on ethnic disparities within the digital economy.

References

Agan, A., & Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, *133*(1), 191–235.

Airbnb. (2022a). About us [Available online at: https://news.airbnb.com/about-us/, last accessed on 2023-07-25].

Airbnb. (2022b, December). A six-year update on airbnb's work to fight discrimination.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, *111*(12), 4088–4118.

Arrow, K. J. (1973). The Theory of Discrimination. In *Discrimination in labor markets* (pp. 3–33). Princeton University Press.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Behaghel, L., Crépon, B., & Le Barbanchon, T. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics*, *7*(3), 1–27.

Bente, G., Baptist, O., & Leuschner, H. (2012). To buy or not to buy: Influence of seller photos and reputation on buyer trust and purchase behavior. *International Journal of Human-Computer Studies*, *70*(1), 1–13.

Blank, R. M., Dabady, M., Citro, C. F., & Blank, R. M. (2004). *Measuring racial discrimination*. National Academies Press Washington, DC.

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2023). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, 1–45.

Bohren, J. A., Hull, P., & Imas, A. (2022). *Systemic discrimination: Theory and measurement* (tech. rep.). National Bureau of Economic Research.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224.

Cornell, B., & Welch, I. (1996). Culture, information, and screening discrimination. *Journal of political Economy*, *104*(3), 542–571.

Cox, M. (2017). Inside Airbnb.

Doleac, J. L., & Stein, L. C. (2013). The Visible Hand: Race and Online Market Outcomes. *The Economic Journal*, *123*(572), F469–F492.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Edelman, B., & Luca, M. (2014). *Digital Discrimination: The Case of Airbnb.com* (Harvard Business School Working Papers No. 14-054). Harvard Business School.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22.

Einav, L., Farronato, C., & Levin, J. (2016). Peer-to-Peer Markets. *Annual Review of Economics*, *8*(1), 615–635.

Ert, E., & Fleischer, A. (2019). What do Airbnb Hosts Reveal by Posting Photographs Online and How Does it Affect their Perceived Trustworthiness? *Psychology and Marketing*, *37*(5), 630–640.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism Management*, *55*, 62–73.

Fisman, R., & Luca, M. (2016). Fixing discrimination in online marketplaces. *Harvard Business Review*, *94*(12), 88–95.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review*, *90*(4), 715–741.

Gössling, S., Larson, M., & Pumputis, A. (2021). Mutual surveillance on airbnb. *Annals of Tourism Research*, *91*, 103314.

Guttentag, D. (2013). Airbnb: Disruptive Innovation and the Rise of an Informal Tourism Accommodation Sector. *Current Issues in Tourism*, *18*(12), 1192–1217.

Hossain, M. (2021). The Effect of the Covid-19 on Sharing Economy Activities. *Journal of Cleaner Production*, *280*, 124782.

Internet archive [Accessed on [11.08.23]]. (2023).

Jaeger, B., Sleegers, W. W., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets [Replications in Economic Psychology and Behavioral Economics]. *Journal of Economic Psychology*, *75*, 102125.

Lang, K., & Spitzer, A. K.-L. (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives*, *34*(2), 68–89.

Laouénan, M., & Rathelot, R. (2017). *Ethnic Discrimination on an Online Marketplace of Vacation Rental* (CAGE Online Working Paper Series No. 318). Competitive Advantage in the Global Economy (CAGE).

Laouénan, M., & Rathelot, R. (2022). Can information reduce ethnic discrimination? evidence from airbnb. *American Economic Journal: Applied Economics*, *14*(1), 107–132.

Leszczensky, L., & Pink, S. (2019). What drives ethnic homophily? a relational approach on how ethnic identification moderates preferences for same-ethnic friends. *American Sociological Review*, *84*(3), 394–419.

Levy, K., & Barocas, S. (2017). Designing against discrimination in online markets. *Berkeley Technology Law Journal*, *32*(3), 1183–1238.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marchenko, A. (2019). The Impact of Host Race and Gender on Prices on Airbnb. *Journal of Housing Economics*, *46*, 101635.

Metropolitan Transportation Authority. (2019). Subway Stations.

Murphy, L. W. (2016). Airbnb's work to fight discrimination and build inclusion. *Report submitted to Airbnb*, *8*, 2016.

OpenStreetMap. (2021). Overpass-Turbo.

Oster, E. (2017). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business and Economic Statistics*, *37*(2), 187–204.

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, *62*(4), 659–661.

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*(4), 667–676.

Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. (2005). *Journal of Political Economy*, *113*(1), 151–184.

Shahn, Z. (2023). Subgroup difference in differences to identify effect modification without a control group. *arXiv preprint arXiv:2306.11030*.

Valfort, M.-A. (2018). Do anti-discrimination policies work? *IZA World of Labor*.

Appendix

3.A Airbnb Anti-Discrimination policy

Figure 3.A.1: Host Profile Pictures-before (left) and after (right) the design change

Note: The figure does not show the exact sizes of the profile pictures, but the proportion represents the design change from 256 to 104 square pixels.
Source: Authors own elaboration

3.B Sample selection

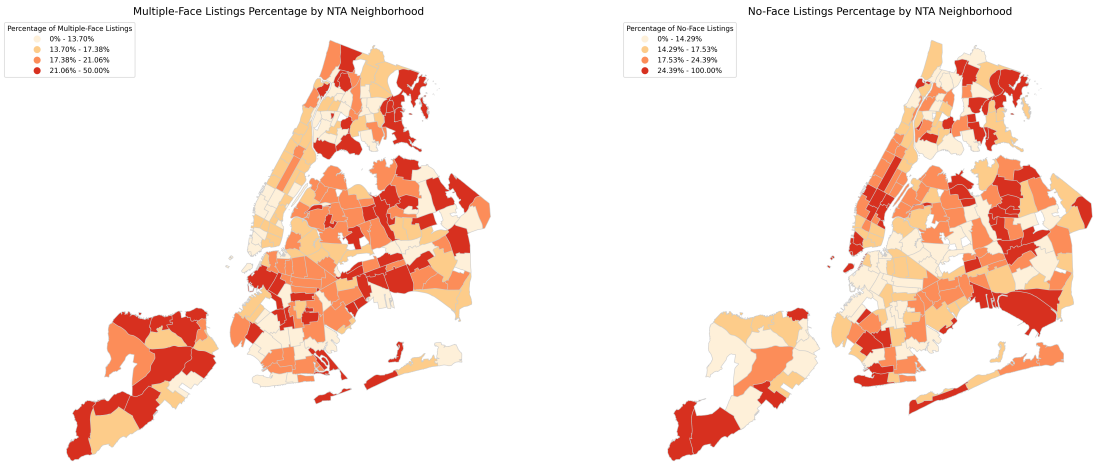
Table 3.B.1: Samples Selection

Initial Number of obs.	1,525,361
Criteria	**Percentage Loss (%)**
Inactive listing	4.46%
Invalid profile picture URL	41%
Picture URL missing	0.12%
Face ethnicity and Name ethnicity mismatch	0.83%
No Human Face detected	9.56%
Multiple Faces detected	10.22%
Pooled OLS number of obs.	574,316
Picture URL changes	36.02%
Race prediction changes	6.83%
Gender prediction changes	3.3%
Not in the sample before and after the policy (DID 6-months)	39.37%
Excluding months outside the reference period (DID 6-months)	71.2%
DID (6-months) number of obs.	99,820

Notes: The table illustrates, for each selection criteria, the percentage lost relative to the initial number of observations.

3.C Spatial Patterns in Host Profile Pictures

Figure 3.C.1: Spatial Distribution of host profile picture characteristics



(a) Multiple faces

(b) No face

Notes: The figures report the percentage of profile pictures characterized by multiple faces (left panel) or no faces (right panel)) over the total number of listings in a given neighborhood. The scale differs between the two panels

Table 3.C.1: Moran I results for No Face and Multiple Faces

	No face	**Multiple Faces**
Moran's I	0.0544	0.0833
p-value	0.107	0.036

Notes: Moran's I measure spatial autocorrelation. Values range from -1 (indicating perfect dispersion) to +1 (indicating perfect clustering). A value of 0 indicates random spatial patterning. The p-value tests the null hypothesis that the observed pattern is random. Lower p-values indicate significant spatial autocorrelation.

3.D Face classification

The Inside Airbnb data provides each host profile picture's Uniform Resource Locator (URL). Previous studies (see Edelman and Luca, 2014, Wang et al., 2015, Kakar et al., 2018 and Marchenko, 2019) also used profile pictures to identify the ethnicity of Airbnb hosts. However, they manually coded the perceived race and other demographic characteristics, making the process time-consuming and not completely replicable.

To fine-tune the ViT model, we engaged five people with diverse backgrounds, including country of birth, age, education, and occupation. We ask them to classify ethnicity, age, gender, and smiling on a random sample of 5,000 Airbnb profile pictures. The mode of these classifications serves as the ground truth in our training dataset.

Therefore, we augmented the training data with 1,354 face images from the Chicago-Face database, which includes self-declared ethnicity and other demographics, such as age and gender (Ma et al., 2015) and 2,223 images from the 10k US Adult Faces Database (Bainbridge et al., 2013).

The ethnic classification achieved an overall accuracy of 92% in the training data and 86% in the test data, competing with state-of-the-art results in ethnic classification models (see, e.g., Abdulwahid, 2023).

The gender classification achieves 95% accuracy in a test sample (97% in the training sample). For smiling, we got 88% accuracy in the test sample and (90% during training) while for the age groups, we reached 82% accuracy in the test sample (93% during training). We provide detailed predictive accuracy metrics for

About 39.4% of our panel's pictures have an invalid URL. Invalid URLs may be caused by hosts changing their profile pictures occasionally. Either the host deleted the old picture or decided to delete its account.

predicted features in Appendix 3.E.

In addition, we use each host's first name to double-check the algorithmic classification for ethnicity and gender. To do so, we use data from three different sources: 1. Worldwide Gender-Name Dictionary, a data set provided by Raffo and Lax-Martinez (2018) which includes 6.2 million names classified by their perceived gender, 2. state-specific data for New York from the Social Security Agency (2020), and 3. a data set provided by Hayes and Mitchell (2020). When the classification of profile pictures completely disagrees with the classification by name, we exclude the observation from the panel. These selection criteria leave us with a final panel that contains 44 months with 574,316 listing x months observations. Table 3.B.1 illustrates the sample selection steps. The panel is unbalanced: some properties enter the system, and others exit.

3.E Performance Metrics Face Classification

Table 3.E.1: Precision, Sensitivity, F_1-Score and Balanced Accuracy of Fine-tuned ViT model (Ethnicity)

		Ethnicity		
Category	Precision	Sensitivity	F1-Score	Balanced Accuracy
Asian	91%	90%	91%	94%
Black	96%	98%	97%	98%
Hispanic	87%	82%	85%	90%
White	94%	97%	95%	96%
Overall Accuracy:	92 %	Observations:	13,037	
		Gender		
Women	97%	97%	97%	97%
Men	96%	97%	97%	97%
Overall Accuracy:	98%	Observations:	13,037	
		Smile		
No	86%	92%	89%	90%
Yes	93%	89%	91%	90%
Overall Accuracy:	90 %	Observations:	13,037	

3.F Descriptive Statistics

Table 3.F.1: Descriptive Statistics of Host Characteristics

	Asian	Black	Hispanic	White
Host is a Superhost	0.16	0.20	0.16	0.18
	(0.37)	(0.40)	(0.37)	(0.38)
Host's Identity Verified?	0.56	0.52	0.56	0.63
	(0.50)	(0.50)	(0.50)	(0.48)
Months of Experience	41.77	40.01	43.16	47.32
	(26.49)	(26.80)	(26.86)	(25.81)
Host about words	28.42	34.94	30.74	31.68
	(38.47)	(37.31)	(35.68)	(34.69)
Host about information	2.11	1.96	2.14	2.29
	(1.45)	(1.42)	(1.49)	(1.50)
Instantbook	0.34	0.42	0.33	0.28
	(0.47)	(0.49)	(0.47)	(0.45)
Female	0.56	0.59	0.52	0.53
	(0.50)	(0.49)	(0.50)	(0.50)
Age: $<$30	0.46	0.18	0.30	0.26
	(0.50)	(0.38)	(0.46)	(0.44)
Age: 30-45	0.44	0.59	0.54	0.46
	(0.50)	(0.49)	(0.50)	(0.50)
Age: 45-60	0.08	0.20	0.14	0.25
	(0.27)	(0.40)	(0.34)	(0.43)
Smile	0.53	0.57	0.56	0.64
	(0.50)	(0.49)	(0.50)	(0.48)
Number of Listings	12561	7505	13136	25339
Number of Hosts	7762	4876	8468	17206
N	112055	79430	122478	260625

Table 3.F.2: Descriptive statistics of Listing characteristics

	Asian	Black	Hispanic	White
Number of Guests	2.79	2.87	2.85	2.89
	(1.89)	(1.88)	(1.87)	(1.86)
Number of Guests Included	1.50	1.62	1.52	1.54
	(1.10)	(1.25)	(1.10)	(1.14)
Number of Bathrooms	1.13	1.11	1.14	1.15
	(0.42)	(0.37)	(0.40)	(0.46)
Number of Bedrooms	1.13	1.18	1.16	1.18
	(0.69)	(0.70)	(0.74)	(0.75)
Number of Beds	1.52	1.57	1.55	1.57
	(1.06)	(1.15)	(1.07)	(1.08)
Cleaning Fee	47.57	42.27	50.78	52.91
	(50.60)	(44.05)	(51.06)	(52.60)
Extra Guests Charge	15.28	17.51	15.38	15.73
	(22.62)	(22.20)	(23.19)	(24.65)
Minimum Stay	5.36	4.29	4.97	5.08
	(14.91)	(12.07)	(10.99)	(12.78)
Number of Amenities	18.77	20.01	18.92	18.92
	(8.71)	(9.72)	(9.04)	(8.80)
Apart/Lofts/Townh/Condos	0.86	0.78	0.87	0.90
	(0.35)	(0.42)	(0.33)	(0.30)
Houses	0.09	0.15	0.09	0.06
	(0.29)	(0.36)	(0.28)	(0.24)
Entire Apartment/House	0.46	0.43	0.51	0.55
	(0.50)	(0.50)	(0.50)	(0.50)
Private Room	0.50	0.53	0.47	0.43
	(0.50)	(0.50)	(0.50)	(0.49)
Shared Room	0.03	0.04	0.02	0.02
	(0.18)	(0.19)	(0.15)	(0.14)
Security Deposit	161.95	123.17	173.24	183.65
	(356.99)	(248.15)	(401.87)	(405.63)
Number of Listings	12561	7505	13136	25339
Number of Hosts	7762	4876	8468	17206
N	112055	79430	122478	260625

Table 3.F.3: Descriptive Statistics of Guest Reviews

	Asian	Black	Hispanic	White
Review Scores Rating	93.44	92.95	93.57	94.19
	(7.94)	(8.39)	(8.06)	(7.27)
Review Scores Accuracy	9.57	9.54	9.58	9.62
	(0.81)	(0.85)	(0.80)	(0.75)
Review Scores Cleanliness	9.21	9.26	9.26	9.28
	(1.07)	(1.03)	(1.03)	(0.98)
Review Scores Check-In	9.73	9.71	9.73	9.77
	(0.68)	(0.71)	(0.68)	(0.61)
Review Scores Communication	9.74	9.71	9.75	9.80
	(0.68)	(0.73)	(0.69)	(0.59)
Review Scores Location	9.50	9.22	9.48	9.57
	(0.78)	(0.88)	(0.80)	(0.72)
Review Scores Value	9.35	9.33	9.36	9.41
	(0.85)	(0.89)	(0.85)	(0.79)
Avg. Monthly Agg. Neg. Sentiment	0.07	0.06	0.07	0.06
	(0.13)	(0.12)	(0.12)	(0.12)
Avg. Monthly Agg. Neu. Sentiment	0.11	0.11	0.11	0.10
	(0.10)	(0.09)	(0.10)	(0.09)
Avg. Monthly Agg. Pos. Sentiment	0.82	0.82	0.82	0.84
	(0.18)	(0.16)	(0.17)	(0.16)
Avg. Cancellation by Host	0.05	0.05	0.05	0.05
	(0.29)	(0.26)	(0.30)	(0.28)
Number of Reviews	23.65	29.00	23.98	26.52
	(39.43)	(45.74)	(39.67)	(43.40)
Number of Reviews Host	56.96	61.89	49.46	46.80
	(124.30)	(108.64)	(106.39)	(103.54)
Number of Listings	12561	7505	13136	25339
Number of Hosts	7781	4873	8501	17162
N	112055	79430	122478	260625

3.G Robustness Checks

Table 3.G.1: Results for Selection on Observables

	Baseline Effect [R^2]	Model 4 [R^2]	R_{max}	δ for $\beta = 0$ given R_{max}	Identified Set
	(1)	(2)	(3)	(4)	(5)
Occupancy Rate					
Asian	-0.042*** [0.003]	-0.016*** [0.161]	0.209	1.607	[-0.050, -0.016]
	(0.006)	(0.004)			
Black	-0.131*** [0.026]	-0.075*** [0.184]	0.239	1.774	[-0.151, -0.075]
	(0.009)	(0.006)			
Hispanic	-0.036*** [0.002]	-0.017*** [0.163]	0.212	2.570	[-0.041, -0.017]
	(0.006)	(0.004)			
Observations (Asian)	325,794	325,794			
Observations (Black)	297,493	297,493			
Observations (Hispanic)	335,414	335,414			
(Log) Price per night					
Asian	-0.144*** [0.010]	-0.020*** [0.737]	0.958	0.456	[-0.181, -0.020]
	(0.021)	(0.008)			
Black	-0.280*** [0.033]	0.005 [0.730]	0.949	-0.040	[-0.369, 0.005]
	(0.036)	(0.009)			
Hispanic	-0.091*** [0.004]	-0.003 [0.736]	0.957	0.108	[-0.118, -0.003]
	(0.006)	(0.004)			
Observations (Asian)	325,794	325,794			
Observations (Black)	297,493	297,493			
Observations (Hispanic)	335,414	335,414			
Time Fixed Effects	No	Yes			
Neighborhood Fixed Effects	No	Yes			
Geographic Controls	No	Yes			
Listing Controls	No	Yes			
Host Controls	No	Yes			
Review Controls	No	No			
Prices/Occupancy rate	No	No			

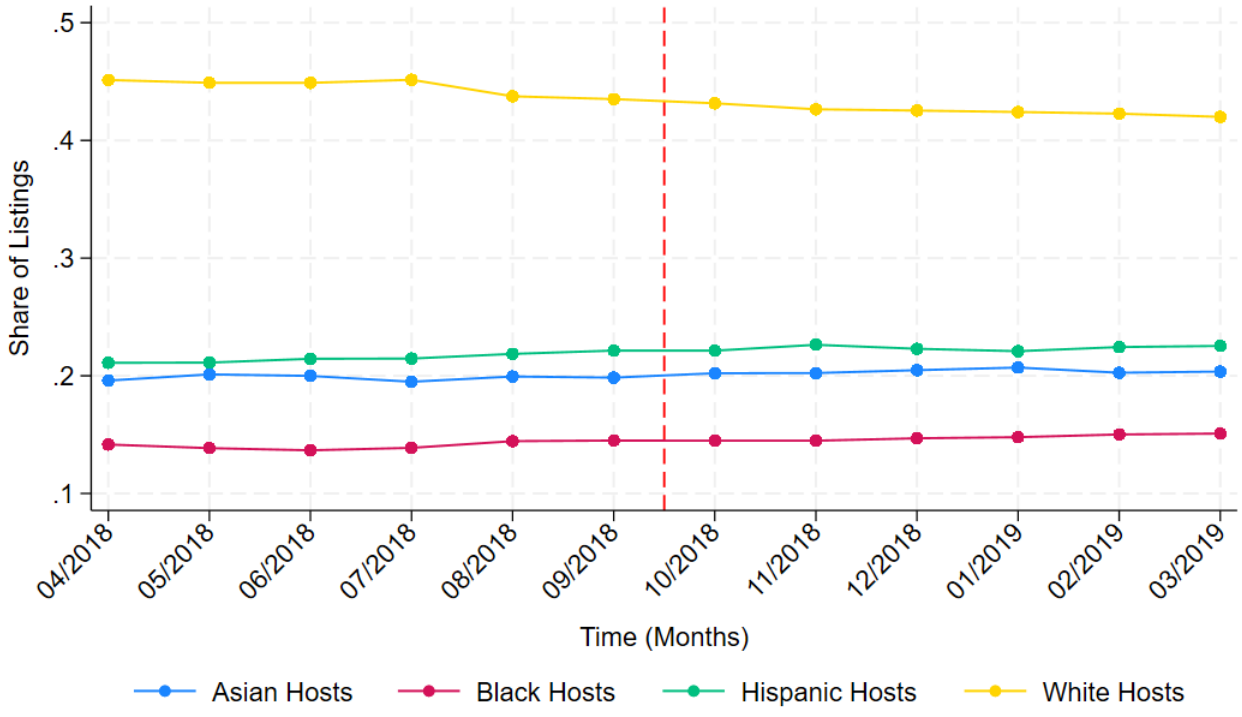
Notes: The table reports pooled OLS results where the dependent variable is occupancy rate or log price per night. The independent variables are dummies indicating the effect for Asian, Black, or Hispanic hosts. The omitted category is White hosts. The table reports the results after applying an approach proposed by Oster, 2017. δ is calculated by assuming $R_{max} = 1.3$(R-squared) and $\beta = 0$ for each racial group individually. The identified set is calculated assuming $R_{max} = 1.3$(R-squared) and $\delta = 1$ for each racial group individually. Standard errors are clustered at the PUMA level. $***p < 0.01, **p < 0.05, *p < 0.1$. *Source*: Author's calculations.

Table 3.G.2: Results for different Ethnicity prediction thresholds

	(1)	(2)	(3)	(4)	(5)
	Baseline	Threshold 50	Threshold 60	Threshold 70	Threshold 80
Asian	-0.016***	-0.017***	-0.016***	-0.014**	-0.017**
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Black	-0.072***	-0.074***	-0.077***	-0.078***	-0.081***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Hispanic	-0.018***	-0.017***	-0.020***	-0.017**	-0.018**
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Observations	502,475	486,098	457,948	429,464	395,169
Adjusted R^2	0.170	0.171	0.172	0.173	0.174

Notes: The table reports pooled OLS results where the dependent variable is occupancy rate. The baseline model refers to Model 4 in the main analysis. Each threshold refers to a sub-sample of hosts where the predicted ethnic probability is higher than the specified threshold. Standard errors are clustered at the PUMA level. $***p < 0.01, **p < 0.05, *p < 0.1$. *Source*: Author's calculations.

Figure 3.G.1: Supply of Listings Before and After the Design Change



Note: The figure illustrates the share of listings active in a given month for each ethnic group. The vertical dotted line indicates the policy implementation date.

Source: Authors' calculation

Table 3.G.3: Results for Demand Shocks After the Policy on Occupancy Rates and Prices

	(1)	(2)
	Occupancy Rate	Log Price
Post-Policy	0.039	-0.004
	(0.035)	(0.011)
Observations	113,252	113,201

Notes: Standard errors are clustered at the listing level.

Source: Author's calculations.

Chapter 4

Spatial Comprehensive Well-Being Composite Indicators based on Bayesian Latent Factor Model: Evidence from Italian Provinces

Montorsi, C., & Gigliarano, C. (2024). Spatial Comprehensive Well-Being Composite Indicators Based on Bayesian Latent Factor Model: Evidence from Italian Provinces. Social Indicators Research, 1-37.

4.1 Introduction

In the socioeconomic literature, we observe a strong consensus that the well-being concept encompasses multiple dimensions and that looking only at economic aspects may distort perceptions, leading to inadequate policy actions (Atkinson & Bourguignon, 1982).

The 2009 report by the Sen-Stiglitz-Fitoussi Commission on the Measurement of Economic Performance and Social Progress marked a milestone in this debate, requiring researchers across the globe to develop new tools for the multidimensional monitoring of well-being (Stiglitz et al., 2009). Since then, the tools used to measure well-being have flourished in Europe and beyond. The standard of living, quality of life, quality of services and many other aspects of well-being have been measured and monitored through an increasing number of specialized indicators. More recently, climate awareness has created new imperatives for the private and public spheres. Air pollution, water quality, particulate matter, and other environmentally related indicators have begun to be assessed throughout Europe, expanding the definition of well-being to include an environmental dimension. In many European countries, these elementary indicators have been integrated into national accounts, expanding policymakers' access to information when designing policies.

Despite these remarkable advances, providing a unique definition of well-being remains a challenge, both on the macro and individual levels. Over the years, scholars have worked to create theoretical frameworks reflecting such multidimensional ideas (see, e.g., Bourguignon and Chakravarty, 2003, and Alkire and Foster, 2011). On the macro level, advanced theoretical models are mainly based on a set (or dashboard) of indicators of demonstrated consistency with the well-being construct. Examples include the OECD Better Life Index (BLI) and the Canadian Index of Wellbeing.

In Italy, the first theoretical framework developed in this debate was the "Equitable and Sustainable Well-Being (BES)" jointly proposed in 2013 by the National Council for Economics and Labor (CNEL) and the Italian National Institute of Statistics (ISTAT).

The emergence of a new multi-dimensional well-being paradigm has been revolutionary but not without drawbacks. Comparing nations or sub-regions with multiple

and arbitrary sub-dimensions of well-being has become a daunting task (Kasparian & Rolland, 2012). This hurdle gives rise to the need for synthesis. Composite indicators (CIs) fulfill this requirement by reducing complex systems into lower-dimension spaces, thus allowing the performance of an individual unit to be evaluated across space and time.

The state-of-the-art aggregation methods for constructing composite indicators entail a broad list of approaches, from simple ones, such as linear aggregation, to more refined ones. Refined empirical indices are built on non-substitutable and non-compensatory indicators and allow for comparison across territorial units (see, e.g., De Muro et al., 2011, Mazziotta and Pareto, 2013, Mazziotta and Pareto, 2018 and Scaccabarozzi et al., 2022).

Although they effectively fulfil their synthesis requirement, most CIs' approaches require researchers to rely on several structural assumptions, for example, lack of uncertainty measure, normative weighting, temporal stability, spatial independence and linearity in the functional form (Ciommi et al., 2017). The approach we propose in this paper addresses three of them.

First, we argue that the normative selection of indicator weights is problematic. For several CIs, the choice of weights comes from expert judgments or is neutral by setting all indicators equally weighted (Mazziotta & Pareto, 2013). This approach exposes indicator weights to the subjectivity of those involved in constructing the CI.

Second, we question the assumption of the spatial independence of elementary indicators across areas. Current methods rely solely on variables from the analyzed area for well-being information, ignoring information from neighboring areas. However, economically speaking, the neighbourhood is not random (Fusco et al., 2018), but instead describes a common culture among enterprises, a shared set of administrative rules on the provincial or regional level and so on, creating spatially aligned clusters, not to mention the detrimental influence of neighbouring factors on the validity of model estimates. In the linear regression framework, spatial correlation creates duplicated information and inflates the variance of the statistical model, damaging the validity of the estimated standard errors (Anselin & Griffith, 1988). As suggested by Fusco et al., 2018, spatial composite indicators bring out inherent local differences by identifying spatial clusters when elementary indicators are well

clustered. Moreover, the lack of attention to the variables' spatial dimension may have significant consequences when assigning weights (Sarra & Nissi, 2020).

Third, traditional indices often lack a measure of uncertainty. This last feature can be problematic if policies or resources' allocation is based on threshold values or composite indicator percentiles (Hogan & Tchernis, 2004).

Researchers solve the weights selection problem relying on data-driven statistical models such as principal component analysis (PCA), factor analytic models Chelli et al. (2015), and Bayesian latent class models (Hogan and Tchernis, 2004; Machado et al., 2009; Ciommi et al., 2020). These weighting methods are helpful when dealing with large data sets to reduce data dimensionality and find common patterns. One critique of this method is that it can accommodate only linear relationships among variables, while it would be reasonable to have non-linear underlying patterns Canning et al., 2013. Nonetheless, when applied to well-being composite indicators, the factor analytic model provides a clear interpretation: the elementary indicators reflect an underlying latent construct interpreted as well-being, and the factor loadings represent each indicator's contribution to this well-being construct (Rijpma, 2016; Ciommi et al., 2020).

Here, we follow the above-outlined approach based on factor models. In addition, we assume that well-being spillovers occur among neighboring provinces, creating well-being levels that are spatially correlated. Since we deal with spatial data, we must reformulate the traditional factor analytic model to incorporate spatial co-variation. We follow Hogan and Tchernis (2004) and Davis et al. (2021) and propose a Bayesian latent factor model for spatially correlated multivariate data. Our Bayesian model confers the distinct advantage of estimating a distribution of well-being for each province instead of single-point estimates, thereby allowing for uncertainty quantification in the estimates. Another advantage of the Bayesian setting is that it can handle missing values with a posterior imputation procedure. In this way, the model directly incorporates the uncertainty caused by missing data into the resulting model's estimates.

This paper proceeds in three steps, as in Ciommi et al. (2020). We first partition the BES elementary indicators into three distinct well-being domains: social, economic, and environmental. We analyze Italian provinces' well-being for each dimension through composite indicators, including spatial correlation among provin-

cial well-being levels. Under the assumption that neighbouring areas influence each other, our proposed method allows us to obtain more precise estimates by exploiting information from neighbouring provinces (Hogan & Tchernis, 2004). In the second step, we consolidate the three well-being dimensions into an overall well-being index for each Italian province. Lastly, we estimate the well-being levels of macro-regions (NUTS 1) and evaluate their evolution over time.

The paper proceeds as follows. Section 4.2 describes the data and summarizes the results from the exploratory spatial analysis. Section 4.3 explains the statistical methodology. Section 4.3.1 presents the estimates from implementing statistical models to 'Province BES' data. Section 4.4 is devoted to concluding remarks.

4.2 Data

This analysis uses data from the Province BES dashboard ('BES at the local level').The Province BES data contains 55 elementary indicators of well-being grouped into 11 macro-domains for the 107 Italian provinces over the period 2004-2021 (ISTAT, 2021). This data source enables well-being monitoring in the Italian territories over time. The presence of missing values, especially in the early and later years, led us to restrict the analysis to 2012 to 2019. We hold elementary indicators with at least one non-missing value for each remaining year. The final set counts 34 elementary indicators. We list and report descriptive statistics for the selected elementary indicators in Appendix 4.A.

Our set of indicators resulted in a missing value percentage of 0.7%, which we then impute with a posterior imputation procedure, as explained in section 4.3.

As in Ciommi et al., 2020, we partition the elementary indicators into three well-being domains: social, economic, and environmental. In doing so, we aim to build composite indicators for each Italian province that summarize the level of well-being in each of these domains.

As mentioned in the introduction, we assume that neighboring provinces have spatially correlated levels of well-being. To test this assumption, we explore the spatial correlation of Province BES indicators through a spatial exploratory data analysis (SEDA). Specifically, we estimate the Moran I test of global spatial corre-

For more references see https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level

lation (see Moran, 1950) and an indicator of the local spatial association (LISA) (see Anselin, 1995). Both approaches test the hypothesis of spatial randomness against the alternative of spatial clustering across each Italian province and BES elementary indicators.

We perform these spatial assessments for each year from 2012 to 2019. The assessment highlights significant spatial auto-correlation for many indicators in all well-being domains. Moran's I coefficients notably differ from zero, indicating spatial solid clustering or patterns. Some indicators show variations in spatial auto-correlation across time. For instance, Life expectancy at birth and Women's political representation in municipalities increased their spatial clustering from 2012 to 2019. Some indicators demonstrate non-significant spatial auto-correlation, reflected in higher p-values (above 0.05), indicating a lack of spatial clustering. For instance, *Public transport networks*, *Specialized doctors*, and *Density of historical green areas*, among others, show no significant spatial patterns in both years. The LISA assessment highlights the highest concentration of spatially correlated observations in East-North and Southern areas. The economic domain has the greatest number of elementary indicators with significant spatial correlation and clustering. Surprisingly, the environmental indicators only have a weak spatial association. The detailed results from the exploratory spatial assessment outlined above are in section 4.B of the Appendix.

This empirical evidence favours our hypothesis that neighbouring provinces share information on socioeconomic development levels. Thus, we estimate latent factor analytic statistical models that flexibly account for spatial correlation in the observed data.

4.3 Bayesian factor model for spatial data

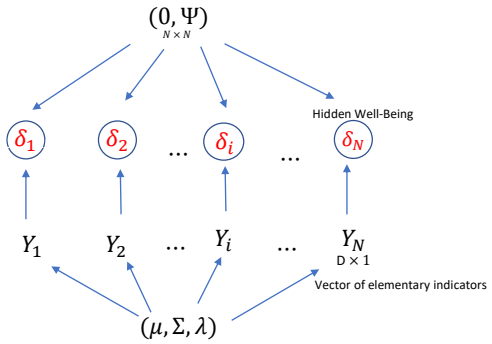
We incorporate spatial information following the Bayesian factor model proposed by Hogan and Tchernis, 2004. This model is based on a latent variable framework, where elementary indicators manifest a hidden construct– the province's well-being.

For province i, where $i = 1, \ldots, N$, with $N = 107$ Italian provinces, let Y_{id} denote the elementary indicator d in province i. The length D of the observed vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iD})$ depends on the well-being domain considered: the social domain has

$D = 20$ indicators, the economic domain has $D = 9$ indicators, and the environmental domain has $D = 5$ indicators.

For each observation i, the latent factor model assumes an L dimensional $(L < D)$ latent variable δ_i that holistically characterizes socioeconomic characteristics. Socioeconomic characteristics, in turn, exemplify through $\boldsymbol{Y}_i$. Here, we assume $L = 1$, hence reducing the model to one latent factor for each province, and represent the model in a hierarchical form as in Figure 4.3.1.

Figure 4.3.1: A graphical representation of a Bayesian hierarchical latent variable model



On the level of observed data, the likelihood is:

$$\boldsymbol{Y}_i \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \delta_i, \Sigma \sim \text{Multivariate Normal}\left(\boldsymbol{\mu} + \boldsymbol{\lambda}\delta_i, \Sigma\right), \tag{4.1}$$

where $\boldsymbol{\mu}$ is a $D \times 1$ mean vector, $\boldsymbol{\lambda}$ is a $D \times 1$ vector of factor loadings, and $\Sigma = \text{diag}\left(\sigma_1^2, \ldots, \sigma_D^2\right)$ is a diagonal matrix measuring residual variation in $\boldsymbol{Y}_i$, implying independence among the elements of $\boldsymbol{Y}_i$ conditionally on δ_i.

In this model, each factor loading is a variance component, i.e. $\lambda_d = \text{cov}(Y_{id}, \delta_i)$. Because residual variances σ_d differ across elementary indicators, factor loadings measure covariance on different scales. They cannot be directly compared to assess the strength of the association between each indicator and provinces' well-being. Instead, following Hogan and Tchernis (2004) and Davis et al. (2021), we can examine squared correlation coefficients $\lambda_d^2/(\lambda_d^2 + \sigma_d^2)$. The squared correlations represent the proportion of variation in latent province well-being explained by each elementary indicator, offering a measure comparable to the weights in a standard weighted average.

On the second level, let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^T$ be the vector of province latent indexes.

The prior distribution is:

$$\boldsymbol{\delta} \sim \text{Multivariate Normal}\left(\mathbf{0}_N, \Psi\right), \tag{4.2}$$

where Ψ is a $N \times N$ spatial covariance matrix with 1's on the diagonal and $\psi_{is} = \text{corr}\left(\delta_i \delta_s\right)$ on the off-diagonal. When $\Psi = I_N$, the model assumes spatial independence. The well-being composite index for province i is summarized by the mean of the posterior distribution of the latent factor δ_i given $\boldsymbol{Y}$ and $\boldsymbol{\mu}, \boldsymbol{\lambda}, \Sigma$.

The prior distributions for the remaining parameters in (4.1) are:

$$\lambda_d \sim \text{Normal}(g, G)(\lambda_1 > 0); \tag{4.3}$$

$$\sigma_d^2 \sim \text{Inverse-Gamma}(\alpha/2, \beta/2); \tag{4.4}$$

$$\mu_d \sim \text{Normal}(0, V_\mu). \tag{4.5}$$

The primary scope of prior distributions is to include subjective opinions on the parameters of interest. However, to let the data speak for themselves and simplify the derivation of posterior distributions, we use conjugate diffuse priors by choosing $g = 0$, $G = 10000$, $\alpha = 1/1000$, $\beta = 1/1000$, and $V_\mu = 1000$.

To include spatial correlation, we work on the spatial covariance matrix Ψ, parametrizing it both marginally and conditionally. The first marginal specification assumes that the generic element ψ_{is} of the prior covariance matrix is

$$\psi_{is} = \text{corr}(\delta_i \delta_s) = \exp(-\omega d_{is}), \tag{4.6}$$

where ω models spatial correlation and $\omega \geq 0$ ensures that $\psi_{is} < 1$; d_{is} is the Euclidean distance between centroids of area i and s and $d_{ii} = 0$ by definition (see Hogan and Tchernis, 2004).

The second way to parametrize the covariance matrix Ψ is through conditional auto-regressive (CAR) specifications of spatial dependency (see Besag et al., 1991). The more general structures are the Gaussian CAR models. These models first require the construction of a set $\mathcal{R}_i$ of areas neighbors of area i. Thus, if we assume the conditional distribution of each δ_i to be

$$\delta_i \mid \{\delta_s : s \in \mathcal{R}_i\} \sim \text{Normal}\left(\sum_{s \in \mathcal{R}_i} \beta_{is} \delta_s, \frac{1}{\alpha_i}\right),$$

then the joint marginal distribution of $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^T$ follows a $\text{Multivariate} - \text{Normal}(\mathbf{0}, B^{-1})$, where B is $N \times N$ spatial covariance matrix with $\{\alpha_1, \ldots, \alpha_N\}$ along the diagonal and $-\alpha_i \beta_{is}$ on the off-diagonal, provided that B is symmetric and positive definite (see Besag, 1974). The β_{is} are general weights defining the influence of province s on the prior mean of δ_i, while α_i represents province-level characteristics such as the number of neighborhoods Hogan and Tchernis (2004).

To ensure that B is positive-definite and symmetric, one or more parameters in the CAR models should be constrained. Here, we consider two different CAR specifications.

Model CAR A defines $\mathcal{R}_i$ as the set of adjacent indicator tracts. R is an adjacency (weight) matrix with $R_{ii} = 0$, $R_{is} = I(s \in \mathcal{R}_i)$ and $R_{is} = R_{si}$. Thus, the model assumes $\beta_{is} = \omega R_{is}$ and $\alpha_i = 1$ (constant), where ω measures the degree of spatial correlation. This leads to the definition

$$B = I_N - \omega R. \tag{4.7}$$

One necessary condition for ensuring that B is positive definite and symmetric is that the ordered eigenvalues $\xi_1, \ldots, \xi_N$ of R satisfy: $\xi_1^{-1} < \omega < \xi_N^{-1}$.

Model CAR B, defines $\mathcal{R}_i$ in the same way as CAR A but here $\beta_{is} = \omega R_{ij} (n_s/n_i)^{1/2}$, and $\alpha_i = n_i$. Where n_i and n_s are the number of neighbours of area i and s, respectively.

For this model

$$B = \text{diag}(n_i) - \omega (n_i * n_s)^{(1/2)} R. \tag{4.8}$$

We estimate the model's posterior distribution using Markov Chain Monte Carlo methods, specifically employing a Gibbs sampling algorithm with Metropolis-Hastings steps to estimate the spatial parameter ω. At each iteration of the algorithm, we draw a sample from the conditional posterior distribution of the model parameters and the latent well-being $\boldsymbol{\delta}$. We use these draws to construct the posterior distributions of all model parameters, discarding the initial samples as a burn-in period. We simulate 6000 draws and "burn" 3000 of them. To obtain our distribution of well-being ranking, we rank the estimates of δ_i in each sampling iteration. The province posterior mean ranking is the mean of a province's rank across all iterations.

A key advantage of this model is that it can handle missing values through

a posterior imputation procedure. The procedure replaces missing elementary indicator values with "draws" from the first level equation conditional on current iterations' "draws" of the latent factor and the other models' parameters (for more details, see Davis et al., 2021).

We carried out a sensitivity analysis to assess the impact of our prior choices on (a) the parameters μ_ω and V_ω of the spatial parameter ω prior distribution, (b) the prior mean and variance, g and G, of the factor loading λ_j, and (c) the prior variance V_μ of the mean μ. We also changed the seed or initial values. Finally, we modified the definition of the spatial topology in the CAR models by increasing the number of neighborhoods and defining the spatial weight matrix R differently. In each case, the resulting estimates remained stable.

The results from this assessment prove the stability of the estimated values to variation in prior choices with a slight degree of instability in the marginal correlation model when changing the prior distribution on the spatial parameter. Data are available upon request.

4.3.1 Economic, social and environmental well-being

We begin by summarizing the posterior distributions of factor loadings (λ_d), residual standard deviations (σ_d), and squared correlations ($\lambda_d^2/(\lambda_d^2+\sigma_d^2)$) for each well-being domain. Next, we present composite indicator estimates for each province, which enable us to examine the extent of divergence in provincial well-being over time and space. To aid in visualizing this heterogeneity, we employ maps that offer an intuitive representation of the spatial distribution of well-being. Furthermore, we compare our data-driven posterior well-being rankings with those obtained through the widely used Mazziotta-Pareto methodology, as extensively documented in the literature (De Muro et al., 2011, Mazziotta and Pareto, 2013). We document the level of agreement between both CIs' approaches. In the following, we focus solely on the results derived from model CAR B, which demonstrated superior models' performances compared to the other spatial models (see section 4.C for models' selection results).

Tables 4.3.1, 4.3.2, and 4.3.3 report the mean posterior estimates of factor loadings, residual standard deviations and squared correlations in the year 2019 (comparisons across spatial models and for 2012 are in Appendix 4.D).

Starting from Table 4.3.1, we find that the leading indicators in the economic domain are *Employment rate*, *Non-participation rate*, and *Youth non-participation rate*, followed by *Pensioners with low pension*. These indicators exhibit high squared correlations, indicating their significant and equal weights in explaining the variation in latent economic well-being. Moving on to the environmental domain (Table 4.3.2), we observe that the indicators *Waste recycling services* and *Separate collection of municipal waste* are the primary drivers of variation in environmental well-being. However, we note that the remaining indicators in this domain have much smaller squared correlations, suggesting a lower impact on environmental well-being.

Finally, turning our attention to Table 4.3.3, we find that the most influential indicator for social well-being is *Graduates mobility*, followed by *People not in education employment or training (neet)* and *People with at least upper secondary education (25–64 years)*. The square correlations for this domain differ significantly across elementary indicators. Suggesting that the elementary social indicators only partially explain social well-being variation across provinces.

Our data-driven approach reveals that elementary indicators have varying weights and contributions to each well-being domain. This result challenges traditional approaches that equally weigh all indicators and emphasizes the need to consider each domain's specific context and characteristics when evaluating provincial well-being.

Figures 4.3.2, 4.3.3, and 4.3.4 illustrate each province's well-being composite indicator and its posterior credibility interval in 2012 and 2019. These figures allow us to assess the variation in well-being trends and rankings of the Italian provinces relative to the Italian mean (represented by the vertical dotted line at 0). Additionally, in Appendix 4.E, we report for each province the posterior distribution quantiles for the three well-being composite indicators.

Examining the figures, we can discern notable patterns in the distribution of well-being across different domains over time. Specifically, in Figure 4.3.2 and Figure 4.3.3, we observe consistent stability in the well-being distribution for the social and economic domains. Only a few provinces exhibit above-average values in the social domain, while most provinces cluster around the mean. This behaviour suggests low polarization in social well-being across Italian provinces. In turn, the economic domain displays more provinces with above and below-average values,

Table 4.3.1: Economic well-being: factor loadings, residual standard deviations, and squared correlations with 95% credibility intervals, on CAR B model, in 2019

Indicator (d)	Factor Loadings (95% CI)	Residual Standard Deviations (95% CI)	Squared Correlations (95% CI)
Employment rate (20–64 years)	1.92 (1.66, 2.20)	0.02 (0.01 , 0.03)	1.00 (1.00 , 1.00)
Non-participation rate	-1.92 (-2.21, -1.66)	0.02 (0.01 , 0.03)	1.00 (1.00 , 1.00)
Youth non-participation rate (15–29 years)	-1.88 (-2.17, -1.63)	0.05 (0.04, 0.07)	1.00 (1.00 , 1.00)
Pensioners with low pension	-1.80 (-2.09, -1.53)	0.14 (0.10, 0.18)	0.99 (0.99 , 1.00)
Youth employment rate (15–29 years)	1.79 (1.52, 2.09)	0.15 (0.11, 0.20)	0.99 (0.99 , 1.00)
Average yearly earnings of employee	1.58 (1.30, 1.91)	0.42 (0.32, 0.56)	0.96 (0.92 , 0.98)
Working days of paid of employee	1.54 (1.29, 1.81)	0.38 (0.28, 0.49)	0.94 (0.90 , 0.97)
Average yearly per-capita pension income	1.48 (1.18, 1.82)	0.42 (0.32, 0.56)	0.92 (0.87, 0.86)
Rate of bank's non-performing loans to households	-1.40 (-1.74 , -1.08)	0.50 (0.38, 0.66)	0.88 (0.81 , 0.94)

Note: Rows indicate the elementary indicators used in the composite indicator's construction. Factor loadings represent the posterior mean of each λ_d in our statistical model. in our statistical model. The numbers in parentheses are the 2.5 and 97.5 quantiles, which define the 95% credibility intervals of the mean. In the Bayesian framework, these values do not indicate significance levels as in the frequentist approach but represent the boundaries that contain 95% of the posterior probability.

Table 4.3.2: Environmental well-being: factor loadings, residual standard deviations, and squared correlations with 95% credibility intervals, on CAR B model, in 2019

Indicator (d)	Factor Loadings (95% CI)	Residual Standard Deviations (95% CI)	Squared Correlations (95% CI)
Waste recycling services	1.46 (1.28, 1.64)	0.01 (0.00, 0.08)	1.00 (1.00, 1.00)
Separate collection of municipal waste	1.35 (1.17, 1.57)	0.16 (0.10 0.22)	0.99 (0.98, 0.99)
Collection of urban waste	0.27 (0.00, 0.55)	0.99 (0.75, 1.30)	0.08 (0.00, 0.25)
Density of historical green areas	0.16 (-0.12, 0.44)	1.03 (0.79, 1.36)	0.04 (0.00, 0.17)
Availability of urban green areas	-0.07 (-0.36, 0.21)	1.04 (0.79, 1.36)	0.02 (0.00, 0.11)

Note: see Table 4.3.1

indicating stronger polarization of economic levels across the Italian surface.

Figure 4.3.4 focuses on the environmental composite indicator. Here, we observe more pronounced variations across the years. Specifically, from 2012 to 2019, a significant decline in environmental well-being is evident across most Italian provinces, leading to heightened polarization in this domain.

Finally, the posterior credibility intervals offer additional insights into the uncertainty surrounding our findings. Over the years, social and economic well-being consistently display relatively narrow credibility intervals for all provinces, indicating low uncertainty in the CIs estimates. On the other hand, there is a gradual decrease in the width of confidence intervals in the environmental dimension over time. This shift reflects an increasing confidence in the CI's point estimate for 2019 compared to 2012. Notably, the figures highlight the impact of missing data in the elementary indicators of Sud Sardegna province. These missing values notably increase the uncertainty surrounding Sud Sardegna's CI estimate, evident through much wider credibility intervals than in other provinces."

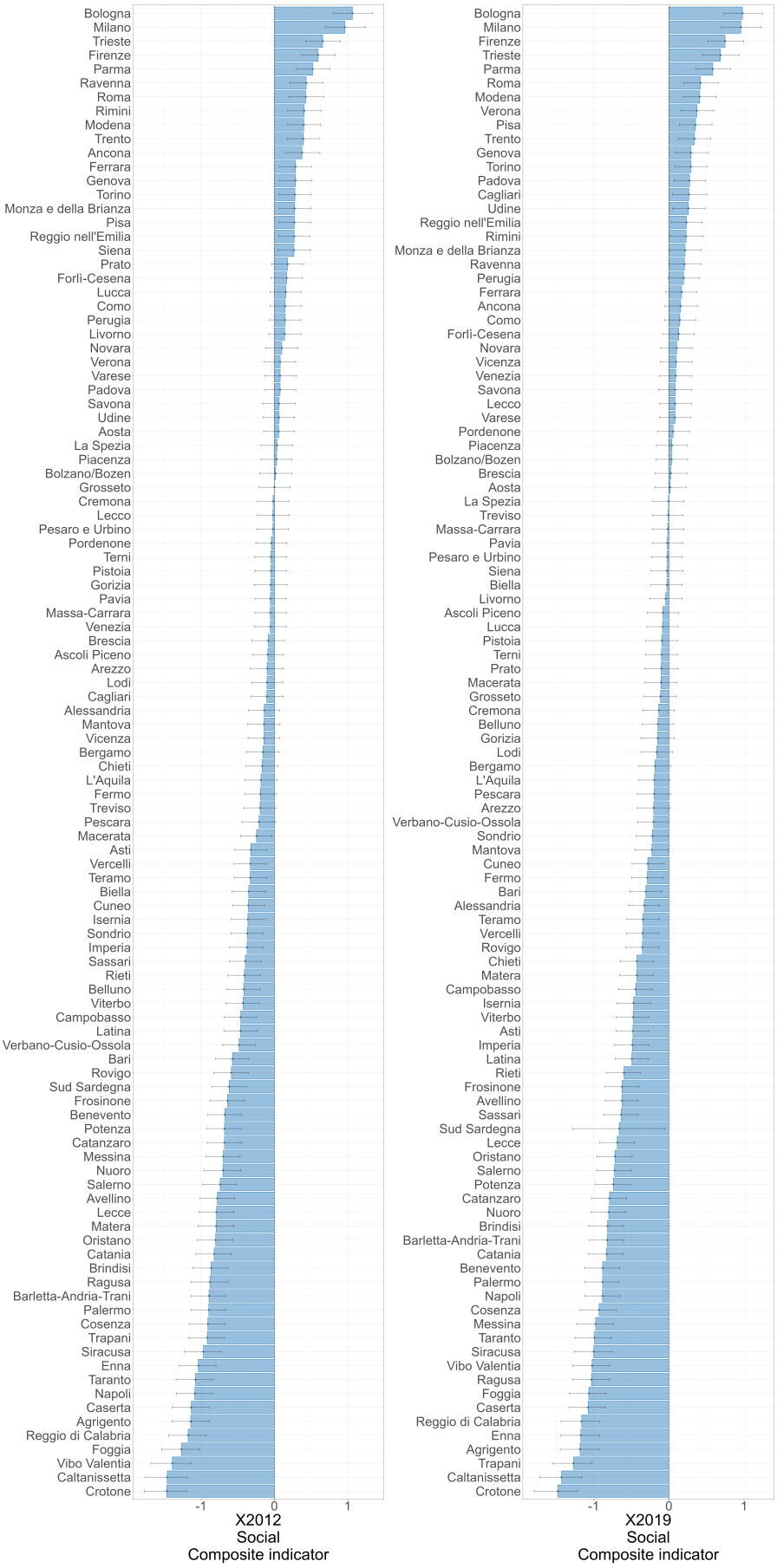
Next, we map the composite indicators' estimates for all Italian provinces at the beginning (the year 2012) and end (the year 2019) of the analysis period. Figure 4.3.5 and Figure 4.3.6 showcase the spatial distributions of social, economic and environmental well-being, respectively. Consistent with our earlier findings, the spatial distribution of social well-being (Figure 4.3.5) remains relatively stable over

Table 4.3.3: Social well-being: factor loadings, residual standard deviations, and squared correlations with 95% credibility intervals, on CAR B model, in 2019

Indicator (d)	Factor Loadings (95% CI)	Residual Standard Deviations (95% CI)	Squared Correlations (95% CI)
Graduates mobility (25–39 years)	1.72 (1.41, 2.04)	0.18 (0.12, 0.26)	0.99 (0.98, 1.00)
People not in education employment or training (neet)	-1.61 (-1.96, -1.30)	0.28 (0.20, 0.39)	0.97 (0.94, 0.99)
People with at least upper secondary education (25–64 years)	1.50 (1.19,1.83)	0.36 (0.27 0.49)	0.94 (0.90, 0.97)
Participation in lifelong learning	1.51 (1.20, 1.86)	0.38 (0.28, 0.50	0.94 (0.89, 0.97)
Irregular electricity services	-1.49 (-1.84,-1.17)	0.39 (0.28, 0.52)	0.93 (0.88, 0.97)
People having completed tertiary education (25–34 years)	1.46 (1.15, 1.81)	0.40 (0.30 0.55)	0.93 (0.87, 0.97)
Children who benefited of early childhood services	1.39 (1.06, 1.75)	0.48 (0.35, 0.64)	0.89 (0.81, 0.95)
Life expectancy at birth	1.31 (0.97, 1.68)	0.52 (0.39 0.70)	0.86 (0.76, 0.94)
Public transportation network	0.97 (0.63, 1.33)	0.76 (0.58 1.01)	0.61 (0.40, 0.79)
Widespread crimes reported	0.95 (0.58, 1.33)	0.77 (0.58 1.03)	0.59 (0.36, 0.78)
Mortality rate in extra-urban road accidents	-0.86 (-1.24, -0.50)	0.82 (0.63, 1.08)	0.52 (0.29, 0.74)
Youth (< 40 years old) political representation	-0.71 (-1.09, -0.35)	0.89 (0.67, 1.17)	0.39 (0.16, 0.64)
Women's political representation in municipalities	0.66 (0.29, 1.02)	0.88 (0.67, 1.15)	0.36 (0.13, 0.61)
Specialized doctors	0.68 (0.32 1.05)	0.91 (0.69, 1.20)	0.36 (0.13, 0.62)
Voluntary murders	-0.65 (-1.02 -0.28)	0.92 (0.69, 1.21)	0.33 (0.12, 0.59)
Health services outflows admittance	-0.63 (-1.03, -0.27)	0.92 (0.70, 1.20)	0.32 (0.10, 0.59)
Hospital beds in high care wards	0.49 0.12 0.86	0.96 (0.73, 1.28)	0.21 (0.03, 0.48)
Roads accidents mortality rate (15–34 years)	-0.38 (-0.76, 0.01)	0.99 (0.76 1.29)	0.15 (0.01, 0.40)
Prison density	0.33 (-0.04, 0.72)	1.00 (0.76 1.31)	0.12 (0.00, 0.35)
Other reported crimes	0.29 (-0.09, 0.67)	1.00 (0.77 1.32)	0.10 (0.00, 0.32)

Note: see Table 4.3.1

Figure 4.3.2: Social well-being: composite indicator estimates for Italian provinces in 2012 (left panel) and 2019 (right panel)

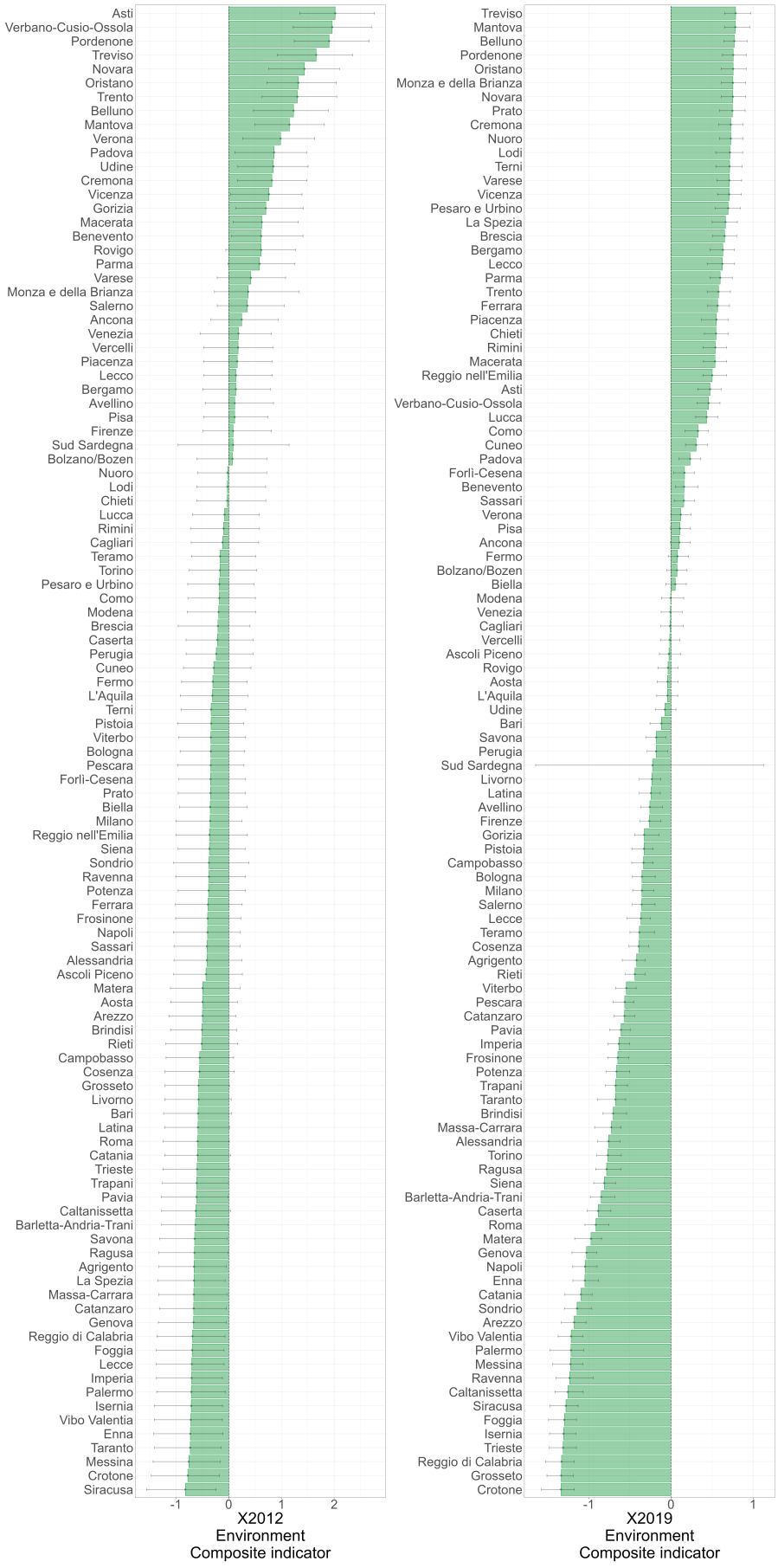


Note: The bars indicate each province's mean posterior composite indicator value in each panel. The horizontal black line corresponds to the 90% posterior credibility interval. The vertical bar at 0 indicates the Italian average for 2012–2019. The wide credibility interval for Sud-Sardegna province is due to the high percentage of missing values. *Source*: Our elaboration of ISTAT "Province BES".

Figure 4.3.3: Economic well-being: composite indicator estimates for Italian provinces in 2012 (left panel) and 2019 (right panel)

Note: see Figure 4.3.2

Figure 4.3.4: Environmental well-being: composite indicator estimates for Italian provinces in 2012 (left panel) and 2019 (right panel)



Note: see Figure 4.3.2

time. Some of the Northern provinces, particularly the regional capitals, exhibit higher levels of social well-being. In the economic domain, a noticeable polarisation persists between the northern and southern provinces, with the former consistently displaying higher levels of well-being. Regarding environmental well-being (Figure 4.3.6), the North-East provinces tend to fare better. However, an interesting trend emerges: the northern provinces show improvement in their environmental well-being levels, whereas the southern provinces experience a marked decline, indicated by the darker shading throughout the years.

Finally, for each well-being domain, we compare the rankings based on our method with rankings based on the Mazziotta-Pareto methodology, which is widely used for policy decision-making in Italy. The Mazziotta-Pareto index (MPI) consists of the arithmetic mean of normalized elementary indicators, incorporating a penalization term for indicator variability, with equal weights assigned to all indicators. The general formula for computing the index entails two steps Mazziotta and Pareto, 2013. First, we calculate the normalized indicator values as follows:

$$z_{id} = 100 + \frac{(y_{id} - \bar{y}_d)}{s_d} 10,$$

where $\bar{y}_d$ and s_d represent the elementary indicator d mean and standard deviation respectively. Then, we estimate the MPI as follows:

$$\text{MPI}_i = M_{z_i} + S_{z_i} \text{cv}_i,$$

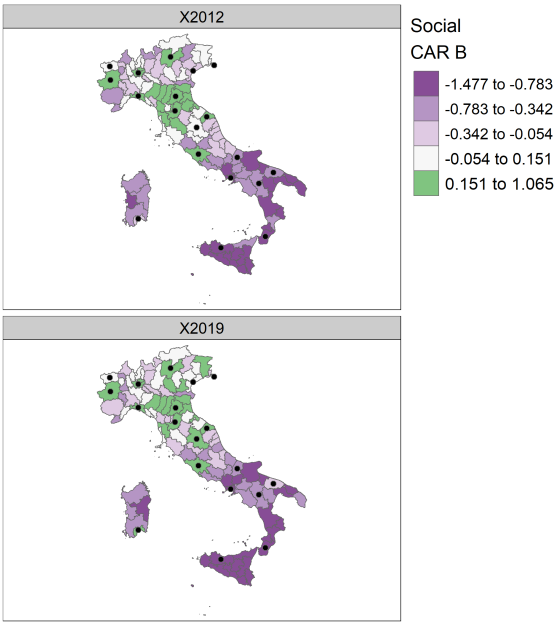
where $M_{z_i} = \frac{\sum_{d=1}^{D} z_{id}}{D}$, $S_{z_i} = \sqrt{\frac{\sum_{d=1}^{D} (z_{id} - M_{z_i})^2}{D}}$, and $\text{cv}_i = \frac{S_{z_i}}{M_{z_i}}$.

Figures 4.3.7 and 4.3.8 show the rankings estimated by our Bayesian model on the x-axes and the corresponding Mazziotta-Pareto rankings on the y-axes. The diagonal line indicates perfect agreement between the Mazziotta-Pareto rank and our mean posterior rank. The farther the provinces are located from this line, the higher the disagreement between the two methodologies.

First, we notice high agreement (Pearson correlation coefficients $(\rho) > 0.8$) between the two methodologies, more pronounced at the top 20% of the rank distribution in all three domains. The economic domain has the highest ranking agreement ($\rho = 0.96$), followed by the social ($\rho = 0.92$) and the environmental domains ($\rho = 0.86$). We observe more disagreement towards the bottom to the middle of the distribution.

These discrepancies in rankings can be attributed to the variation in the weights assigned to the elementary indicators in our model compared to the equally weighted Mazziotta-Pareto indicator. This finding indicates that different evaluation methods for provincial well-being can lead to changes in provincial ranks. However, it is important to note that our results align with the Mazziotta-Pareto ranking for some provinces, suggesting that certain provinces have specific needs that warrant more focused interventions.

Figure 4.3.5: Maps of provincial social well-being composite indicators, for 2012 (top panel) and 2019 (bottom panel)

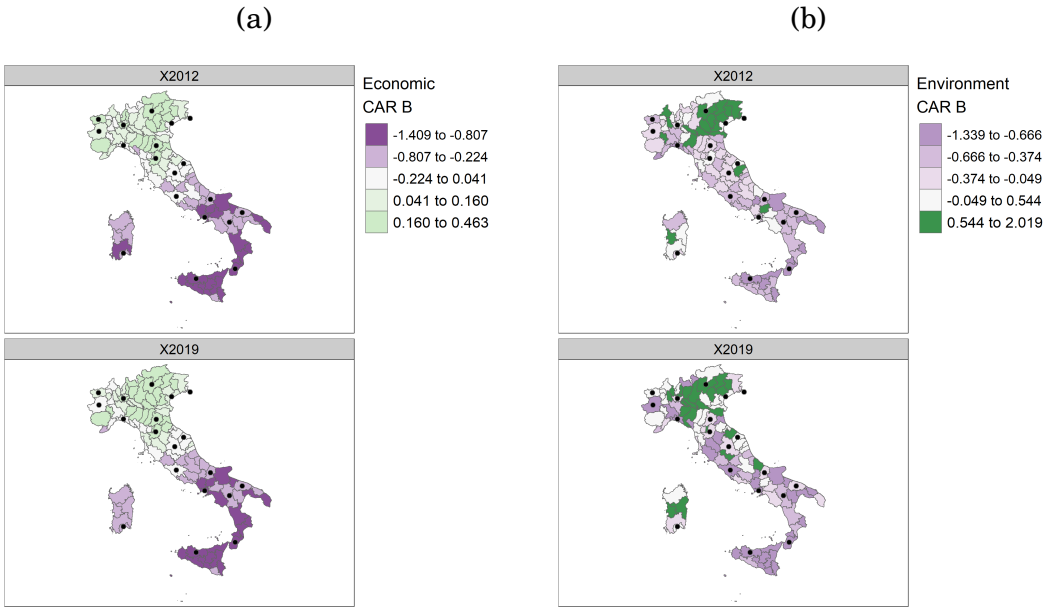


Note: Italian provinces are grouped in well-being quintiles. The more 'purple' colors refer to worse-off provinces, while 'greener' shades indicate better-off provinces. The black dots indicate provincial capitals. Provinces with negative values are below the Italian averages over the entire period of analysis

4.3.2 Overall well-being

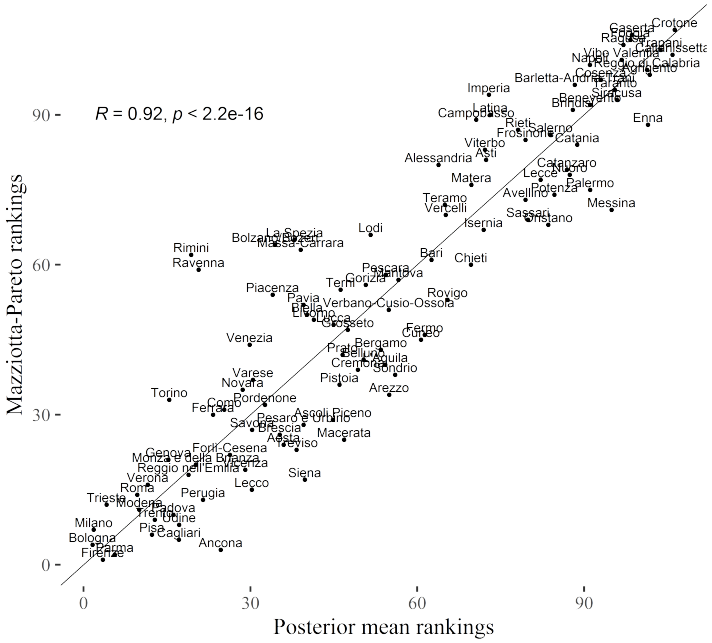
To provide a comprehensive assessment of the overall well-being of each province, we condense the three previously estimated composite indicators into a single composite value. The three composite indicators already account for the spatial correlation among Italian provinces. Here, we employ a spatially independent latent factor model using posterior mean estimates of the three well-being composite indicators as the model's outcomes. Let $\hat{\boldsymbol{\delta}}_i = (\hat{\delta}_{i1}, \hat{\delta}_{i2}, \hat{\delta}_{i3})$ indicating the 3-dimensional vector of composite well-being indicators for each province i. We consider the following Bayesian factor model:

Figure 4.3.6: Maps of provincial economic (left) and environmental (right) well-being composite indicators, for 2012 (top panel) and 2019 (bottom panel)



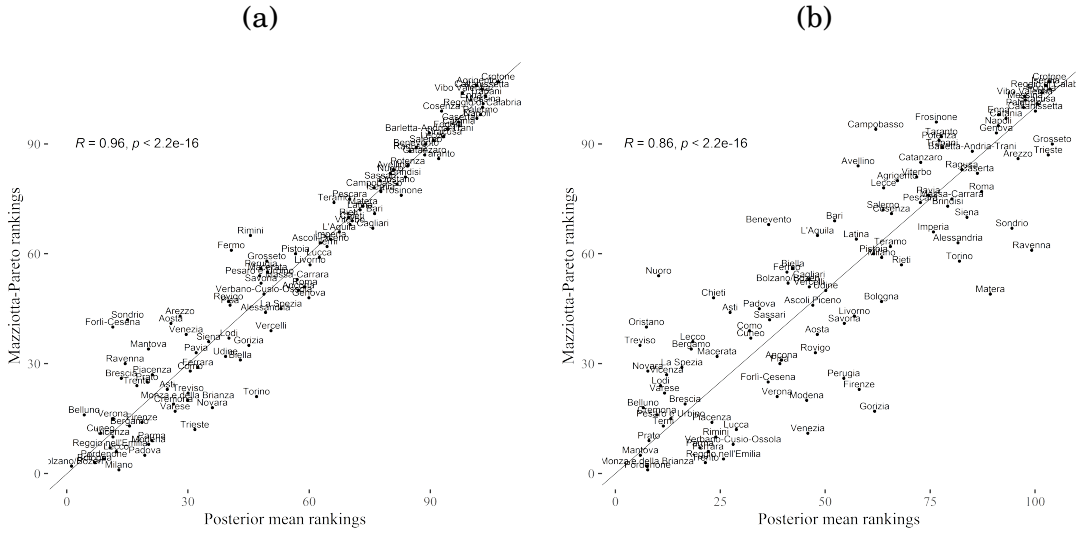
Note: see Figure 4.3.5

Figure 4.3.7: Social well-being posterior mean rankings and Mazziotta-Pareto rankings for 2019



Note: Posterior mean rankings produced by model CAR B. The R in the left corner is the Pearson correlation coefficient between posterior mean ranking and the Mazziotta-Pareto rankings. We remove Sud-Sardegna province from the plot for its many missing values.

Figure 4.3.8: Economic well-being (a) and environmental well-being (b) posterior mean rankings and Mazziotta-Pareto rankings for 2019



Note: see Figure 4.3.7

$$\hat{\boldsymbol{\delta}}_i \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, w_i, \Sigma \sim \text{Multivariate Normal}(\boldsymbol{\mu} + \boldsymbol{\lambda} w_i, \Sigma) \quad \textit{likelihood}$$

$$w_i \sim \text{Normal}(0, 1) \quad \textit{prior}$$

We estimate the posterior distribution of factor loadings, residual standard deviation, and squared correlations, as presented in Table 4.3.4. This table reveals two key insights. Firstly, we observe a strong correlation between the overall well-being composite indicator (w_i) and the economic and social well-being composite indicators. Notably, the economic domain appears to have the highest weight among the well-being domains. This suggests that targeting economic aspects in low-developed provinces would reduce the disparities in overall well-being between Italian provinces.

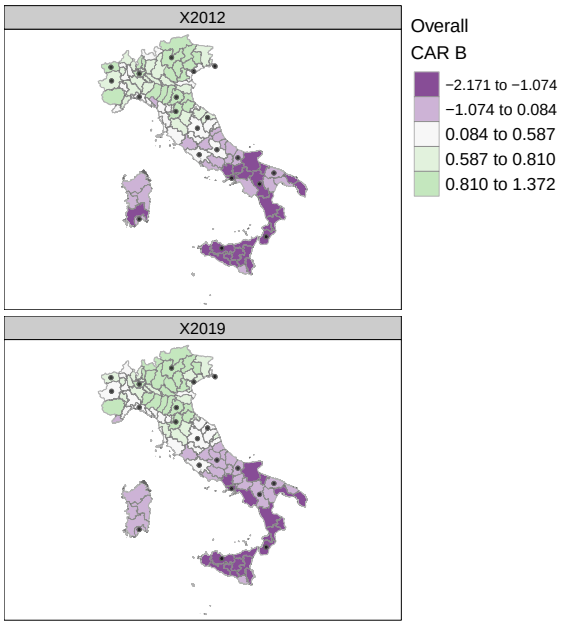
Figure 4.3.9 provides a map representation to illustrate overall well-being visually on the Italian surface. The spatial distribution of overall well-being is non-random, with the northern provinces consistently exhibiting higher levels of well-being while the southern provinces consistently experience lower well-being. Additionally, overall well-being slightly increases over time in some southern and central provinces, revealing a moderate polarisation reduction.

Table 4.3.4: Overall well-being: factor loadings, residual standard deviations and squared correlations with 95% credibility intervals in 2019

Domain (d)	Factor Loadings (95% CI)	Residual Standard Deviations (95% CI)	Squared Correlations (95% CI)
Social	0.77 (0.68 ,0.85)	0.06 (0.05, 0.08)	0.77 (0.68, 0.85)
Economic	0.98 (0.92, 0.99)	0.006 (0.0006, 0.03)	0.98 (0.92,1.00)
Environmental	0.34 (0.19, 0.49)	0.31 (0.25, 0.4)	0.34 (0.20, 0.49)

Note: see Table 4.3.1

Figure 4.3.9: Maps of provincial overall well-being composite indicator, for 2012 (top panel) and 2019 (bottom panel)



Note: see Figure 4.3.5

4.3.3 Macro region well-being

Finally, we aggregate provinces belonging to the same macro-region m (NUTS1), for $m = 1, \ldots, 5$, i.e. Northwest, Northeast, Center, South, and Islands, to assess the evolution of the Italian macro-regional well-being over time. We consider a

hierarchical model, which requires specifying a prior distribution for the mean ($\alpha_{m[i]}$) for each macro-area m of the latent variable $\hat{\delta}_i$. We also assume the variance ($\sigma_{m[i]}$) to vary across macro-areas. More formally, for the three well-being domains, the model becomes:

$$\begin{aligned}
\hat{\delta}_i \mid \alpha_{m[i]}, \sigma_{m[i]} &\sim \text{Normal}(\alpha_{m[i]}, \sigma_{m[i]}) \quad \textit{likelihood} \\
\alpha_{m[i]} &\sim \text{Normal}(0, 1) \quad \textit{prior} \\
\sigma_{m[i]} \mid \nu, \tau &\sim \text{cauchy}(\nu, \tau) \quad \textit{prior}
\end{aligned}$$

As standard practice, we chose a normal distribution as the prior distribution for the mean ($\alpha_{m[i]}$) and a Cauchy distribution for the standard deviation ($\sigma_{m[i]}$) of the latent factor distribution Gelman et al., 2013, and interpret α_m as the well-being level of macro-region m.

Figures 4.3.10 and 4.3.11 show each macro area time series for social, economic, environmental, and overall well-being. These figures reveal a consistent and enduring macro-territorial division that characterizes the Italian territory throughout the analyzed period. Notably, the South and Islands consistently fall below the average, while the Center, Northwest, and Northeast remain above the average. Moreover, these macro areas intersect in specific years and for particular well-being domains.

The trend in economic well-being remains relatively flat over time, with a consistent ranking of macro areas across the years. On the other hand, social well-being, illustrated on the left in Figure 4.3.10 exhibits more interaction among macro areas over the years. The Center aligns with the Northwest, maintaining a similar trajectory until 2019, while the Northeast shows a slight upward trend. In contrast, the Islands' social well-being experienced a decline over time, reaching a lower level in 2019 compared to the beginning of the series.

Only environmental well-being demonstrates a non-flat trend over time among the four estimated time series. In 2016 the Northwest and Northeast aligned, while the South experienced a steady decline after 2015. The Center exhibits an upward trend after 2016, and the Islands remain relatively stable.

We estimate the hierarchical models above using STAN interfaces in R Carpenter et al., 2017. The code for implementing the Hierarchical models is available on GitHub.

Finally, the overall domain mirrors the evolution of social and economic well-being levels. The environmental domain contributes minimally to determining the overall well-being trend.

Figure 4.3.10: Social (left) and economic (right) well-being composite indicator for Italian macro territorial areas (black dotted line indicates the Italian average)

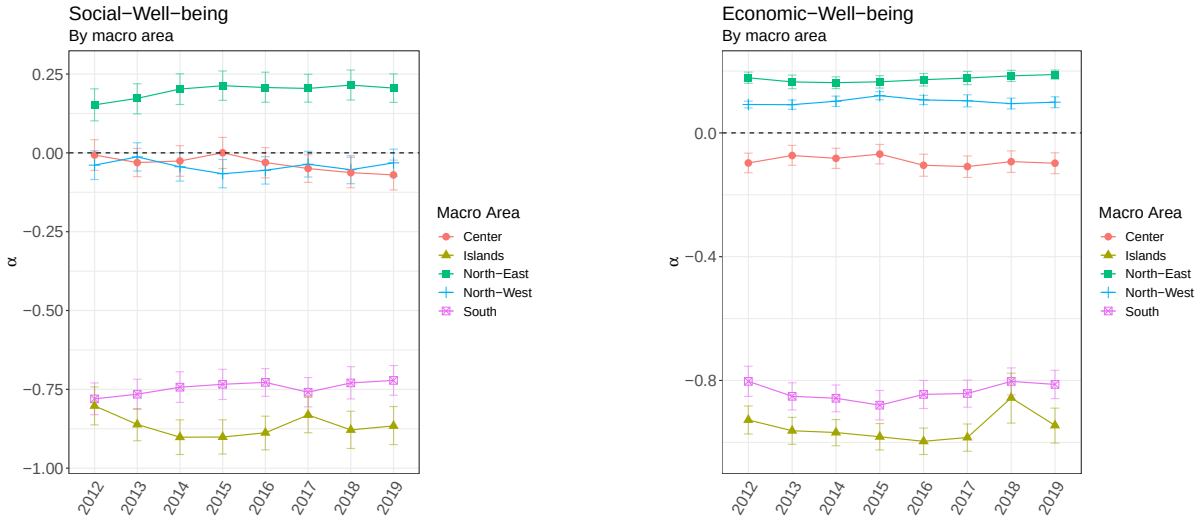
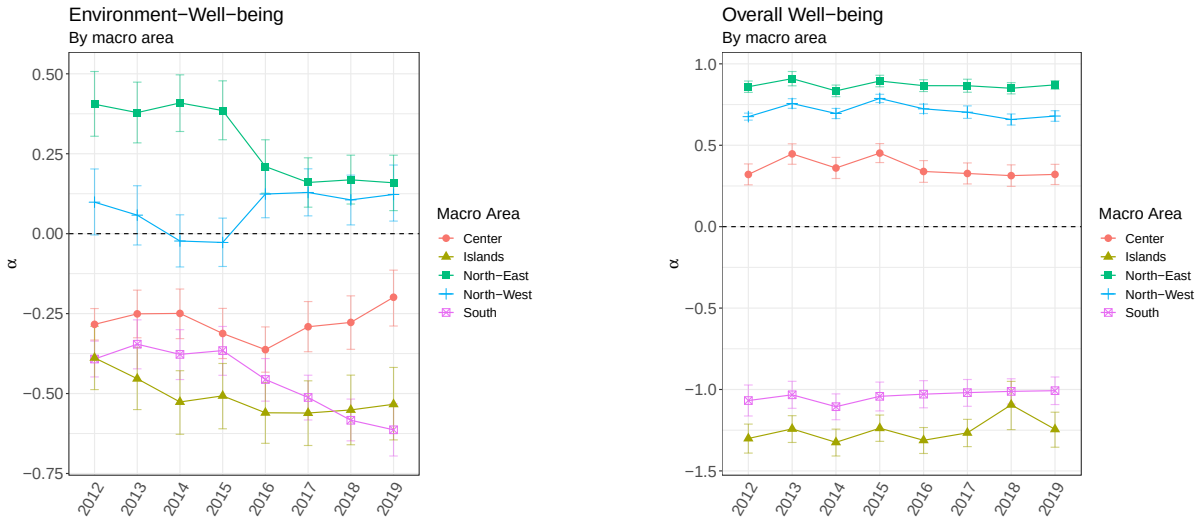


Figure 4.3.11: Environmental (left) and overall(right) well-being composite indicator for Italian macro territorial areas (black dotted line indicates the Italian average)



4.4 Concluding remarks

This paper applies a Bayesian spatial latent factor model to propose well-being composite indicators and rankings for all Italian provinces from 2012 to 2019. Our approach differs from traditional composite indicators methodologies in several

ways. First, we modeled the spatial dependence of elementary indicators, capturing potential socioeconomic spillover effects. Second, we incorporate a measure of composite indicators uncertainty related to missing data. Third, we estimate data-driven weights for elementary indicators, thus avoiding an arbitrary selection of weight exposed to subjective opinion.

Using the ”Province BES” dataset from ISTAT, we examine the assumption of spatial independence in the elementary indicators by conducting global and local tests of spatial association. This initial assessment confirms positive spatial association in the “Province BES” indicators. We then categorize the indicators into three sustainable development well-being domains: social, economic, and environmental. Employing a Bayesian approach, we estimate the posterior distribution of latent variables, with their expected values interpreted as hidden well-being indicators for Italian provinces.

The study reveals significant disparities in social and economic well-being between northern and southern regions, with the northern provinces consistently demonstrating higher levels of well-being. In contrast, the environmental dimension exhibits less persistent polarization, with above-average levels observed in the South. One possible interpretation is that environmental consciousness has gained prominence more recently than socioeconomic aspects. Consequently, northern and southern provinces are experiencing increased climate awareness, with similar provincial investments in environmental well-being rates. Compared to the Mazziotta-Pareto approach, our rankings diverge, particularly at the upper end of the well-being distribution and within the environmental domain. Uncertainty in ranking estimates is also higher for provinces that are better off. These findings suggest that the government could allocate resources more effectively by targeting provinces at the lower end of the well-being ranking.

Subsequently, we reduce the three well-being dimensions into an overall well-being indicator for each Italian province. This composite indicator, driven primarily by the economic domain and with minimal weight given to environmental well-being, remains stable and clustered throughout the analyzed period. These findings emphasize the significance of economic factors in shaping overall well-being and highlight regional disparities within Italy. They also indicate that focused interventions to improve economic conditions can reduce provincial well-being disparities.

Additionally, we extend the analysis to the NUTS-1 level, encompassing Northwest, Northeast, Center, South, and Island macro-regions, in order to provide well-being trends across the analyzed period. The results demonstrate varying degrees of heterogeneity among these macro areas.

The primary limitation of this study is the limited number of indicators available within the environmental dimension compared to the social and economic dimensions, which also suffer from more missing observations. As long as data on environmental aspects remain scarce, it will be challenging for researchers to provide robust evidence in favour of climate policy interventions. In future research, we aim to enrich the environmental analysis by integrating advanced sensor measurements of air pollution, water quality, and soil temperature into national accounts. Additionally, we plan to incorporate a subjective dimension that considers citizens' perceptions of life satisfaction.

Overall, this study contributes to the understanding of well-being dynamics in Italy, offering valuable insights for policymakers in addressing regional disparities and focusing on targeted interventions for improved well-being outcomes.

References

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*, *95*(7-8), 476–487.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, *27*(2), 93–115.

Anselin, L., & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers in Regional Science*, *65*(1), 11–34.

Atkinson, A. B., & Bourguignon, F. (1982). The comparison of multi-dimensioned distributions of economic status. *The Review of Economic Studies*, *49*(2), 183–201.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Bourguignon, F., & Chakravarty, S. R. (2003). The measurement of multidimensional poverty. *The Journal of Economic Inequality*, *1*, 25–49.

Canning, D., French, D., & Moore, M. (2013). Non-parametric estimation of data dimensionality prior to data compression: The case of the human development index. *Journal of Applied Statistics*, *40*(9), 1853–1863.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Chelli, F. M., Ciommi, M., Emili, A., Gigliarano, C., & Taralli, S. (2015). Comparing equitable and sustainable well-being (bes) across the italian provinces. a factor analysis-based approach. *Rivista Italiana di Economia Demografia e Statistica*, *LXIX (3)*, 61–72.

Ciommi, M., Gigliarano, C., Chelli, F. M., Gallegati, M., et al. (2020). It is the total that does [not] make the sum: Nature, economy and society in the equitable and sustainable well-being of the Italian provinces. *Social Indicators Research.*, *https://doi.org/10.1007/s11205-020-02331-w*.

Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., & Chelli, F. M. (2017). A new class of composite indicators for measuring well-being at the local level: An application to the equitable and sustainable well-being (bes) of the Italian provinces. *Ecological indicators*, *76*, 281–296.

Davis, W., Gordan, A., & Tchernis, R. (2021). Measuring the spatial distribution of health rankings in the United States. *Health Economics*, *30*(11), 2921–2936.

De Muro, P., Mazziotta, M., & Pareto, A. (2011). Composite indices of development and poverty: An application to MDGs. *Social indicators research*, *104*, 1–18.

Fusco, E., Vidoli, F., & Sahoo, B. K. (2018). Spatial heterogeneity in composite indicator: A methodological proposal. *Omega*, *77*, 1–14.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Hogan, J. W., & Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, *99*(466), 314–324.

ISTAT. (2021). BES 2021. Il benessere equo e sostenibile in Italia. Rome.

Kasparian, J., & Rolland, A. (2012). OECD's 'Better Life Index': Can any country be well ranked? *Journal of Applied Statistics*, *39*, 2223–2230.

Machado, C., Paulino, C. D., & Nunes, F. (2009). Deprivation analysis based on Bayesian latent class models. *Journal of Applied Statistics*, *36*, 871–891.

Mazziotta, M., & Pareto, A. (2013). Methods for constructing composite indices: One for all or all for one. *Rivista Italiana di Economia Demografia e Statistica*, *LXVII (2)*, 67–80.

Mazziotta, M., & Pareto, A. (2018). Measuring well-being over time: The adjusted Mazziotta–Pareto index versus other non-compensatory indices. *Social Indicators Research*, *136*(3), 967–976.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

Rijpma, A. (2016). *What can't money buy? wellbeing and gdp since 1820* (tech. rep.). Utrecht University, Centre for Global Economic History.

Sarra, A., & Nissi, E. (2020). A spatial composite indicator for human and ecosystem well-being in the Italian urban areas. *Social Indicators Research*, *148*, 353–377.

Scaccabarozzi, A., Mazziotta, M., & Bianchi, A. (2022). Measuring competitiveness: A composite indicator for Italian municipalities. *Social Indicators Research*, 1–30.

Stiglitz, J. E., Sen, A., Fitoussi, J.-P., et al. (2009). Report by the commission on the measurement of economic performance and social progress.

Appendix

4.A Descriptive statistics

Table 4.A.1: Descriptive statistics of selected elementary indicators, all years

Domain	Indicator	mean	median	sd	Unit
Social	Graduates mobility (25-39 years)	-9.18	-6.13	13.04	Ratio
	People not in education, employment, or training (neet)	23.28	21.00	8.35	%
	Participation in lifelong learning	7.45	7.20	2.21	%
	People with at least upper secondary education level (25–64 years)	58.97	60	7.51	%
	Irregular electricity services	2.27	1.88	1.28	Average number for user
	People having completed tertiary education (25-34 years)	23.38	23.00	5.57	%
	Children who benefited from early childhood services	13.22	12.10	7.65	%
	Life expectancy at birth	82.56	82.5	0.83	Years
	Public transportation network	2618.30	2187.45	2037.70	Seat-km per capita
	Widespread crimes reported	190.74	179.40	72.02	For10.000 inhabitants
	Mortality rate in extra-urban road accidents	5.58	5.10	2.84	%
	Youth (¡ 40 years old) political representation	30.81	30.70	5.36	%
	Specialized doctors	27.04	24.7	7.46	For10.000 inhabitants
	Women's political representation in municipalities	28.16	29.00	6.73	%
	Voluntarily murders				
	Health services outflows admittance	7.96	6.30	5.07	%
	Hospital beds in high care wards	2.95	2.7	1.24	For10.000 inhabitants
	Road accidents mortality rate (15–34 years)	0.75	0.70	0.41	For 10.000 inhabitants
	Prison density	128.55	126.70	41.55	%
	Other reported crimes	16.47	15.60	5.03	For 10.000 inhabitants
Economic	Employment rate (20–64 years).	61.54	65.90 9.85	%	
	Non-participation rate	18.36	14.30	10.84	%
	Youth non-participation rate (15–29 years)	33.43	30.45	16.72	%
	Pensioners with low pension	10.97	9.47	3.21	%
	Youth employment rate (15–29 years)	35.97	36.15	10.90	%
	Working days of paid of employees	75.16	76.61	5.52	%
	Average yearly earnings of employee	18302.73	18123.38	3077.61	Euro
	Average yearly per-capita pension income	16028.68	15981.43	1597.06	Euro
	Rate of bank non-performing loans to households	1.30	1.20	0.54	%
Environmental	Waste recycling services	45.05	30.01	77.21	mq for inhabitant
	Separate collection of municipal waste	37.52	37.60	20.17	%
	Collection of urban waste	37.52	37.60	20.17	%
	Density of historical green areas	2.49	1.60	3.26	mq for 100 mq res.areas
	Availability of urban green areas	45.05	30.01	77.21	mq for inhabitant
	Source: our elaboration on "Province BES', ISTAT 2019				

4.B Spatial Exploratory Data Analysis

Table 4.B.1

Domain	Indicator	**Moran I** 2012	**p value**	**Moran I** 2019	**p value**
Soc.	Graduates mobility (25–39 years)	0.59	< 0.001	0.68	< 0.001
	People not in education employment or training (neet)	0.72	< 0.001	0.74	¡ 0.001
	Participation in long life learning	0.21	< 0.001	0.33	< 0.001
	People with at least upper secondary education level (25–64 years)	0.40	< 0.001	0.45	< 0.001
	Irregular electricity services	0.67	< 0.001	0.64	< 0.001
	People having completed tertiary education (25–39 years)	0.26	< 0.001	0.18	< 0.001
	Children who benefited of early childhood services	0.70	< 0.001	0.65	< 0.001
	Life expectancy at birth	0.53	< 0.001	0.60	< 0.001
	Public transport network	-0.06	0.81	-0.03	0.67
	Widespread crimes reported	0.28	< 0.001	0.21	< 0.001
	Mortality rate in extra urban road accidents	0.29	< 0.001	0.35	< 0.001
	Youth (¡40 years old) political representation in municipalities	0.49	< 0.001	0.39	< 0.001
	Specialized doctors	0.04	0.23	0.03	0.26
	Women s political representation in municipalities	0.81	< 0.001	0.64	< 0.001
	Voluntary murders	0.22	< 0.001	0.08	0.07
	Health services outflows admittance's	0.39	< 0.001	0.43	< 0.001
	Hospital beds in high care wards	-0.06	0.77	-0.09	0.90
	Roads accidents mortality rate (15–34 years)	0.10	0.05	0.04	0.23
	Prison density	0.09	0.05	0.18	< 0.001
	Other reported crimes	0.14	0.01	0.10	0.04
Eco.	Employment rate (20–64 years)	0.82	< 0.001	0.82	< 0.001
	Non participation rate	0.82	< 0.001	0.81	< 0.001
	Youth non participation rate (15–29 years)	0.79	< 0.001	0.80	< 0.001
	Pensioners with low pension	0.78	< 0.001	0.80	< 0.001
	Youth employment rate (15–29 years)	0.74	< 0.001	0.77	< 0.001
	Working days of paid employee	0.65	< 0.001	0.62	< 0.001
	Average yearly earnings of employee	0.65	< 0.001	0.68	< 0.001
	Average yearly per-capita pension income	0.54	< 0.001	0.61	< 0.001
	Rate of bank's non performing loans to households	0.35	< 0.001	0.52	< 0.001
Env	Waste recycling services	0.53	< 0.001	0.34	< 0.001
	Separate collection of municipal waste	0.67	0.00	0.47	< 0.001
	Collection of urban waste	0.51	< 0.001	0.57	< 0.001
	Density of historical green areas	-0.05	0.78	-0.04	0.70
	Availability of urban green areas	0.08	0.05	0.03	0.21

Note: Each row corresponds to one of the 34 elementary indicators used in our model. The table reports the results from Moran's test of spatial autocorrelation. The second and fourth columns report the value of the observed Moran's I coefficient in 2012 and 2019. The third and fifth columns reports the p-value of the test. When p-value is < 0.001, we reject the null hypothesis of spatial randomness at 1% significance level.

Table 4.B.2: Proportion of provinces with statistically significant p-value ($p < 0.005$) for the LISA statistic, for each BES elementary indicator, for 2012 and 2019

Dom.	Indicator	2012	2019
Soc.	Prison density	0.14	0.12
	Other reported crimes	0.13	0.12
	Youth (< 40 years old) political representation	0.20	0.19
	Women s political representation in municipalities	0.22	0.25
	Children who benefited from early childhood services	0.24	0.20
	Widespread crimes reported	0.17	0.10
	Regional health services outflows hospital admittances	0.21	0.19
	People not in education employment or training (neet)	0.17	0.18
	Irregular electricity services	0.20	0.15
	People having completed tertiary education (25–39 years)	0.10	0.10
	Graduates mobility (25–39 years)	0.22	0.26
	Roads accidents mortality rate	0.10	0.08
	Mortality rate in extra urban road accidents	0.17	0.19
	Participation in long life learning	0.18	0.19
	People with at least upper secondary education level (25–64 years)	0.18	0.15
	Public transport network	0.07	0.07
	Life expectancy at birth	0.23	0.19
	Specialized doctors	0.11	0.13
	Voluntary murders	0.08	0.06
	Hospital beds in high care wards	0.07	0.03
Eco.	Employment rate (20–64 years)	0.21	0.21
	Non-participation rate	0.19	0.20
	Youth non participation rate (15–29 years)	0.22	0.21
	Pensioners with low pension	0.20	0.20
	Youth employment rate (15–29 years)	0.25	0.25
	Average yearly earnings of employee	0.21	0.20
	Average yearly per capita pension income	0.20	0.21
	Rate of bank's non-performing loans to households	0.07	0.15
	Working days of paid of employees	0.23	0.24
Env	Waste recycling services	0.12	0.19
	Separate collection of municipal waste	0.21	0.20
	Collection of urban waste	0.19	0.21
	Density of historical green areas	0.06	0.04
	Availability of urban green areas	0.07	0.08

Note: These are results from the function that estimates the (non-centred) local indicators of spatial association modified form proposed in Anselin, 1995. The p-value is the permutation two-sided p-value for each observation.

4.C Models' selection criteria

Table 4.C.1: Goodness of fit measures for the three well-being domains, for 2019.

Criteria	Model	Economic	Social	Environment
p	Marginal Correlation	340.12	1554.01	357.97
	CAR A	258.43	1551.40	362.79
	CAR B	**233.95**	**1549.78**	**356.44**
G	Marginal Correlation	515.97	1378.69	329.52
	CAR A	250.95	1381.51	325.43
	CAR B	**206.99**	**1377.46**	**329.63**
C	Marginal Correlation	856.09	2932.71	687.50
	CAR A	509.38	2932.91	688.22
	CAR B	**440.94**	**2927.24**	**686.07**

Note: We assess the goodness of fit (G) and variability (p) of the models following the model selection criterion proposed by Gelfand and Ghosh, 1998. These criteria penalize the lack of fit and high posterior predictive variance due to over and under-parametrization. C is the sum of the two measures. Lower values of C indicate better models' performance.

4.D Factor Loadings across spatial models and years

Figure 4.D.1: Social well-being: factor loadings with 95% credibility intervals, for the three spatial models, in 2012 and 2019

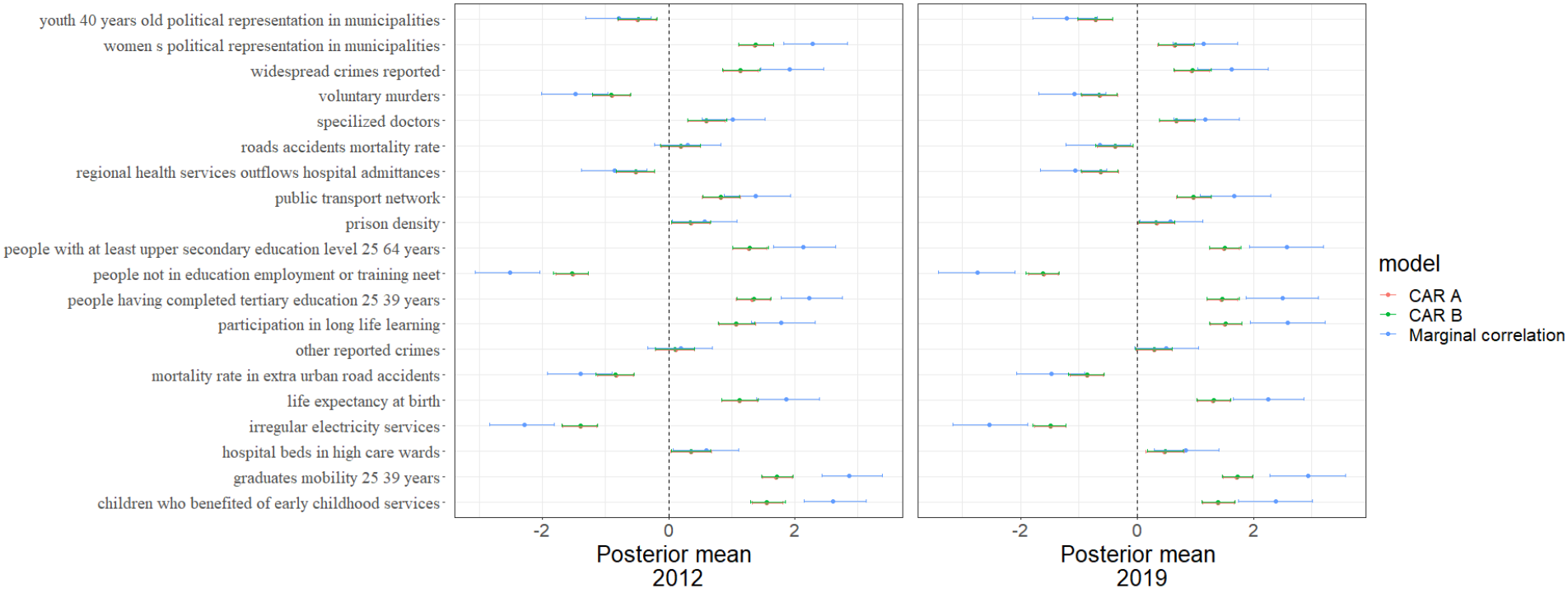


Figure 4.D.2: Economic well-being: factor loadings with 95% credibility intervals, for the three spatial models, in 2012 and 2019

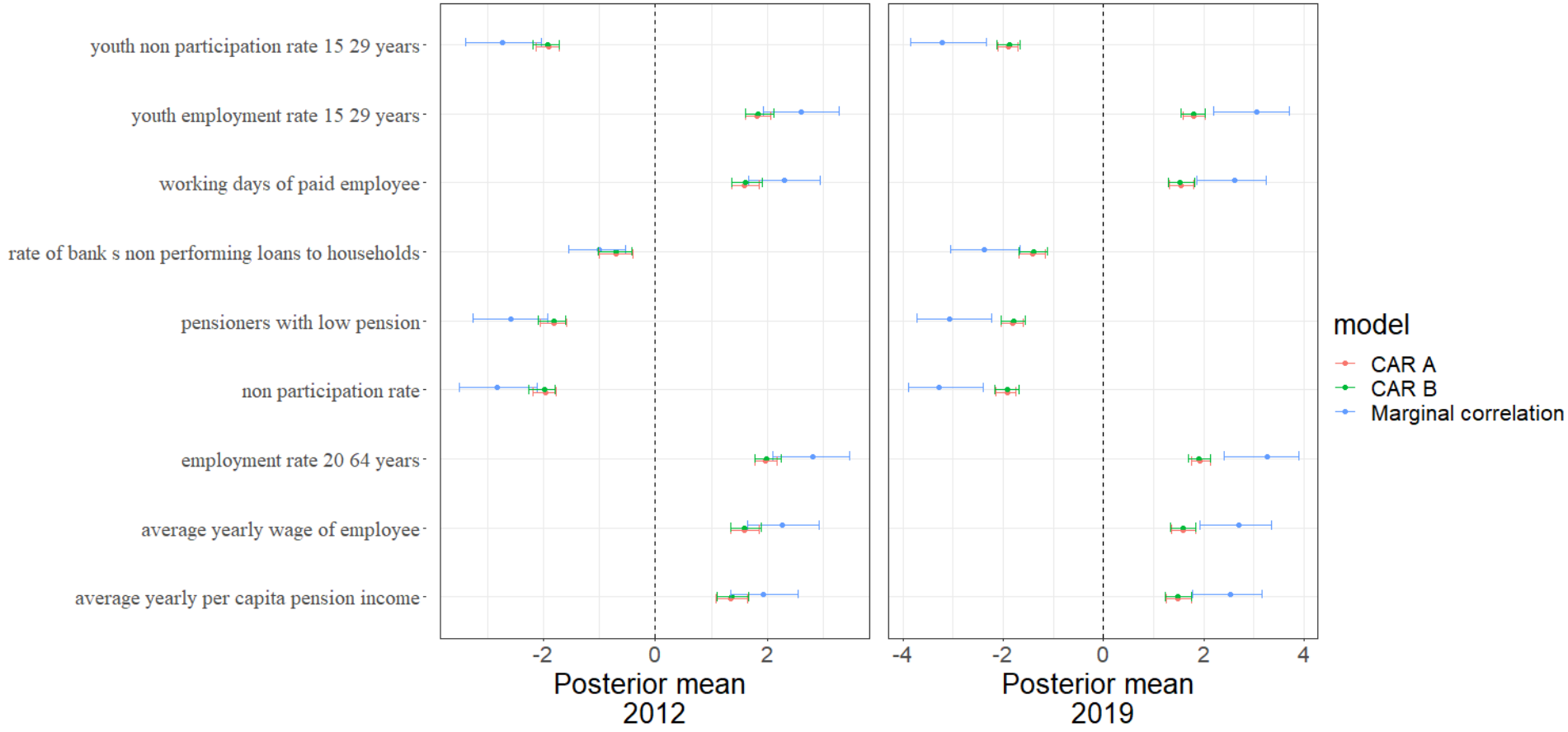
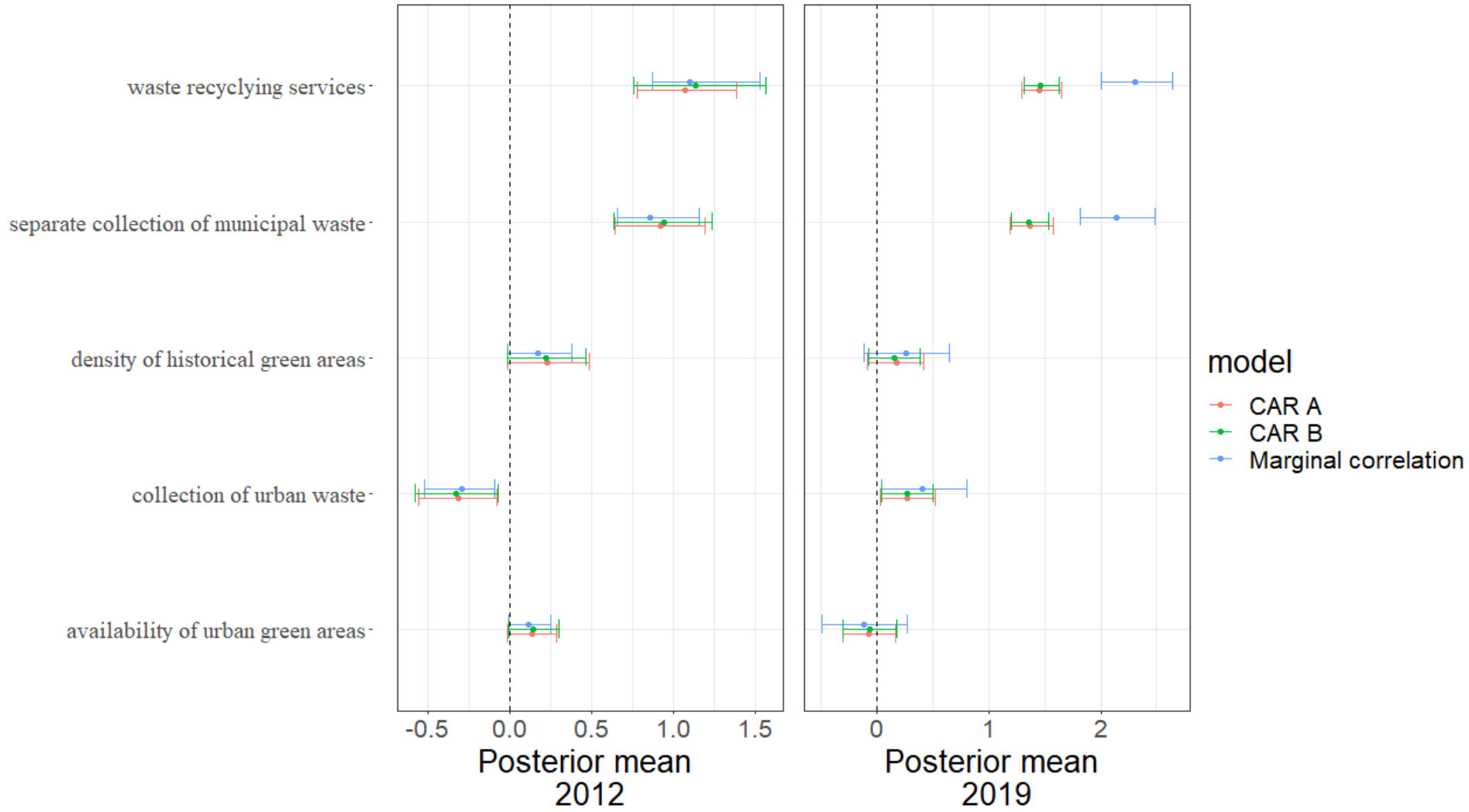


Figure 4.D.3: Environmental well-being: factor loadings with 95% credibility intervals, for the three spatial models, in 2012 and 2019



4.E Full distribution of composite indicators

Table 4.E.1: Summary of posterior distribution of the composite indicator for the social dimension. Model CAR B. Year 2019.

Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR
Agrigento	-1.18	-1.18	-1.50	-0.89	0.21	Foggia	-1.06	-1.06	-1.39	-0.79	0.21	Pescara	-0.20	-0.19	-0.46	0.07	0.17
Alessandria	-0.33	-0.32	-0.58	-0.09	0.17	Forlì-Cesena	0.13	0.13	-0.11	0.39	0.17	Piacenza	0.04	0.04	-0.21	0.28	0.17
Ancona	0.16	0.15	-0.09	0.44	0.18	Frosinone	-0.62	-0.62	-0.90	-0.36	0.18	Pisa	0.36	0.35	0.11	0.62	0.18
Aosta	0.02	0.02	-0.22	0.26	0.17	Genova	0.30	0.29	0.04	0.56	0.18	Pistoia	-0.09	-0.09	-0.35	0.15	0.17
Arezzo	-0.20	-0.20	-0.47	0.06	0.17	Gorizia	-0.15	-0.15	-0.42	0.11	0.17	Pordenone	0.06	0.05	-0.18	0.31	0.17
Ascoli Piceno	-0.08	-0.08	-0.33	0.18	0.17	Grosseto	-0.11	-0.11	-0.36	0.14	0.17	Potenza	-0.74	-0.73	-1.03	-0.47	0.19
Asti	-0.48	-0.47	-0.75	-0.23	0.18	Imperia	-0.49	-0.48	-0.77	-0.23	0.18	Prato	-0.10	-0.10	-0.37	0.15	0.18
Avellino	-0.62	-0.62	-0.89	-0.37	0.18	Isernia	-0.47	-0.47	-0.74	-0.20	0.18	Ragusa	-1.03	-1.02	-1.33	-0.75	0.21
Bari	-0.31	-0.30	-0.57	-0.05	0.18	L'Aquila	-0.19	-0.19	-0.44	0.05	0.18	Ravenna	0.21	0.20	-0.04	0.47	0.17
Barletta-Andria-Trani	-0.82	-0.82	-1.10	-0.56	0.19	La Spezia	0.00	0.00	-0.25	0.24	0.17	Reggio di Calabria	-1.16	-1.15	-1.49	-0.88	0.20
Belluno	-0.14	-0.14	-0.40	0.11	0.17	Latina	-0.49	-0.49	-0.76	-0.23	0.18	Reggio nell'Emilia	0.23	0.23	-0.01	0.48	0.17
Benevento	-0.88	-0.88	-1.16	-0.61	0.19	Lecce	-0.69	-0.69	-0.97	-0.41	0.18	Rieti	-0.60	-0.59	-0.88	-0.34	0.18
Bergamo	-0.18	-0.18	-0.45	0.06	0.18	Lecco	0.08	0.08	-0.16	0.35	0.17	Rimini	0.23	0.23	-0.02	0.49	0.18
Biella	-0.03	-0.03	-0.28	0.22	0.17	Livorno	-0.04	-0.05	-0.30	0.21	0.17	Roma	0.42	0.42	0.16	0.70	0.18
Bologna	0.98	0.97	0.69	1.29	0.21	Lodi	-0.16	-0.16	-0.41	0.09	0.17	Rovigo	-0.35	-0.35	-0.62	-0.09	0.18
Bolzano/Bozen	0.04	0.04	-0.22	0.29	0.17	Lucca	-0.08	-0.09	-0.33	0.16	0.17	Salerno	-0.72	-0.72	-1.01	-0.46	0.19
Brescia	0.03	0.03	-0.22	0.28	0.17	Macerata	-0.10	-0.10	-0.36	0.14	0.17	Sassari	-0.64	-0.63	-0.92	-0.38	0.18
Brindisi	-0.82	-0.81	-1.11	-0.56	0.18	Mantova	-0.22	-0.22	-0.50	0.03	0.17	Savona	0.08	0.08	-0.18	0.35	0.18
Cagliari	0.27	0.26	0.01	0.55	0.18	Massa-Carrara	-0.01	-0.02	-0.26	0.24	0.17	Siena	-0.03	-0.02	-0.27	0.22	0.17
Caltanissetta	-1.43	-1.42	-1.78	-1.12	0.22	Matera	-0.43	-0.42	-0.71	-0.16	0.18	Siracusa	-1.00	-0.99	-1.31	-0.71	0.21
Campobasso	-0.44	-0.44	-0.71	-0.19	0.18	Messina	-0.97	-0.97	-1.27	-0.68	0.19	Sondrio	-0.22	-0.22	-0.47	0.03	0.18
Caserta	-1.07	-1.07	-1.38	-0.79	0.20	Milano	0.96	0.95	0.65	1.30	0.22	Sud Sardegna	-0.66	-0.67	-1.39	0.10	0.48
Catania	-0.83	-0.82	-1.12	-0.57	0.19	Modena	0.41	0.41	0.16	0.68	0.17	Taranto	-0.99	-0.98	-1.30	-0.72	0.19
Catanzaro	-0.79	-0.79	-1.08	-0.52	0.19	Monza e della Brianza	0.22	0.21	-0.03	0.47	0.18	Teramo	-0.35	-0.35	-0.61	-0.08	0.17
Chieti	-0.43	-0.42	-0.69	-0.16	0.18	Napoli	-0.88	-0.88	-1.18	-0.61	0.20	Terni	-0.10	-0.10	-0.35	0.15	0.17
Como	0.15	0.14	-0.09	0.41	0.17	Novara	0.10	0.10	-0.15	0.36	0.17	Torino	0.29	0.29	0.04	0.56	0.18
Cosenza	-0.93	-0.92	-1.24	-0.66	0.20	Nuoro	-0.80	-0.80	-1.08	-0.53	0.19	Trapani	-1.27	-1.26	-1.61	-0.97	0.22
Cremona	-0.13	-0.14	-0.39	0.11	0.17	Oristano	-0.72	-0.72	-1.01	-0.46	0.19	Trento	0.34	0.34	0.09	0.61	0.17
Crotone	-1.48	-1.47	-1.84	-1.16	0.24	Padova	0.28	0.27	0.03	0.53	0.18	Treviso	-0.01	-0.01	-0.26	0.23	0.17
Cuneo	-0.28	-0.28	-0.53	-0.03	0.17	Palermo	-0.88	-0.88	-1.18	-0.62	0.19	Trieste	0.69	0.68	0.41	1.00	0.20
Enna	-1.17	-1.16	-1.49	-0.88	0.21	Parma	0.58	0.58	0.33	0.87	0.18	Udine	0.26	0.26	0.01	0.53	0.17
Fermo	-0.29	-0.29	-0.55	-0.02	0.17	Pavia	-0.02	-0.02	-0.27	0.23	0.17	Varese	0.08	0.08	-0.17	0.33	0.18
Ferrara	0.17	0.17	-0.08	0.42	0.17	Perugia	0.20	0.20	-0.05	0.45	0.17	Venezia	0.09	0.09	-0.16	0.35	0.18
Firenze	0.75	0.74	0.48	1.05	0.19	Pesaro e Urbino	-0.02	-0.02	-0.27	0.22	0.17	Verbano-Cusio-Ossola	-0.20	-0.20	-0.46	0.04	0.18
												Vercelli	-0.35	-0.35	-0.61	-0.09	0.17
												Verona	0.37	0.37	0.12	0.64	0.18

Table 4.E.2: Summary of posterior distribution of the latent variable (composite indicator) for the economic dimension. CAR model B. Year 2019.

Province	mean	median	25%	75%	IQR
Agrigento	-1.25	-1.25	-1.46	-1.07	0.14
Alessandria	0.01	0.01	-0.08	0.10	0.06
Ancona	-0.08	-0.08	-0.17	0.02	0.06
Aosta	0.16	0.16	0.06	0.25	0.06
Arezzo	0.14	0.14	0.05	0.24	0.06
Ascoli Piceno	-0.18	-0.18	-0.27	-0.09	0.06
Asti	0.16	0.16	0.07	0.26	0.06
Avellino	-0.68	-0.68	-0.82	-0.55	0.09
Bari	-0.56	-0.56	-0.69	-0.45	0.08
Barletta-Andria-Trani	-0.93	-0.93	-1.10	-0.77	0.11
Belluno	0.31	0.31	0.21	0.41	0.07
Benevento	-0.83	-0.83	-0.99	-0.68	0.10
Bergamo	0.21	0.21	0.12	0.31	0.07
Biella	0.05	0.05	-0.04	0.15	0.06
Bologna	0.28	0.28	0.18	0.38	0.07
Bolzano/Bozen	0.45	0.45	0.35	0.55	0.07
Brescia	0.23	0.23	0.14	0.32	0.06
Brindisi	-0.76	-0.75	-0.90	-0.62	0.10
Cagliari	-0.55	-0.55	-0.68	-0.44	0.08
Caltanissetta	-1.27	-1.27	-1.48	-1.08	0.14
Campobasso	-0.56	-0.56	-0.68	-0.45	0.08
Caserta	-1.16	-1.16	-1.35	-0.98	0.13
Catania	-1.16	-1.16	-1.35	-0.98	0.13
Catanzaro	-0.90	-0.90	-1.06	-0.75	0.11
Chieti	-0.41	-0.41	-0.52	-0.31	0.07
Como	0.13	0.13	0.04	0.22	0.06
Cosenza	-1.02	-1.01	-1.20	-0.85	0.12
Cremona	0.15	0.15	0.06	0.25	0.06
Crotone	-1.41	-1.41	-1.64	-1.20	0.16
Cuneo	0.26	0.26	0.17	0.36	0.07
Enna	-1.23	-1.22	-1.43	-1.04	0.14
Fermo	0.07	0.07	-0.03	0.16	0.06
Ferrara	0.12	0.12	0.03	0.21	0.06
Firenze	0.20	0.19	0.10	0.29	0.06
Foggia	-1.07	-1.07	-1.26	-0.90	0.12
Forlì-Cesena	0.24	0.24	0.14	0.34	0.07
Frosinone	-0.73	-0.73	-0.88	-0.59	0.10
Genova	-0.12	-0.12	-0.21	-0.02	0.06
Gorizia	0.04	0.04	-0.05	0.13	0.06
Grosseto	0.01	0.01	-0.09	0.10	0.06
Imperia	-0.24	-0.24	-0.34	-0.15	0.07
Isernia	-0.60	-0.60	-0.74	-0.48	0.09
L'Aquila	-0.31	-0.31	-0.42	-0.21	0.07
La Spezia	-0.03	-0.03	-0.12	0.07	0.06
Latina	-0.46	-0.46	-0.58	-0.36	0.07
Lecce	-0.96	-0.96	-1.13	-0.80	0.11
Lecco	0.23	0.23	0.15	0.33	0.06
Livorno	-0.12	-0.12	-0.22	-0.03	0.07
Lodi	0.07	0.07	-0.02	0.16	0.06
Lucca	-0.17	-0.17	-0.27	-0.07	0.06
Macerata	0.00	0.00	-0.09	0.09	0.06
Mantova	0.19	0.18	0.10	0.28	0.06
Massa-Carrara	-0.07	-0.07	0.02	-0.16	0.06
Matera	-0.49	-0.48	-0.61	-0.37	0.08
Messina	-1.28	-1.28	-1.49	-1.08	0.14
Milano	0.23	0.23	0.14	0.33	0.07
Modena	0.19	0.18	0.09	0.28	0.06
Monza e della Brianza	0.13	0.13	0.04	0.23	0.07
Napoli	-1.25	-1.25	-1.46	-1.07	0.14
Novara	0.10	0.10	0.01	0.19	0.06
Nuoro	-0.66	-0.66	-0.80	-0.54	0.09
Oristano	-0.70	-0.70	-0.84	-0.58	0.09
Padova	0.19	0.19	0.10	0.29	0.06
Palermo	-1.27	-1.26	-1.48	-1.08	0.14
Parma	0.18	0.18	0.09	0.28	0.06
Pavia	0.12	0.12	0.03	0.21	0.06
Perugia	0.02	0.02	-0.07	0.10	0.06
Pesaro e Urbino	0.02	0.02	-0.07	0.11	0.06
Pescara	-0.39	-0.39	-0.50	-0.28	0.07
Piacenza	0.18	0.18	0.09	0.28	0.06
Pisa	0.07	0.07	-0.02	0.16	0.06
Pistoia	-0.07	-0.07	-0.16	0.02	0.06
Pordenone	0.26	0.25	0.16	0.35	0.06
Potenza	-0.77	-0.77	-0.91	-0.64	0.10
Prato	0.19	0.19	0.09	0.29	0.07
Ragusa	-0.78	-0.78	-0.94	-0.64	0.10
Ravenna	0.22	0.22	0.13	0.31	0.06
Reggio di Calabria	-1.28	-1.27	-1.49	-1.08	0.14
Reggio nell'Emilia	0.24	0.24	0.15	0.34	0.06
Rieti	-0.39	-0.38	-0.50	-0.28	0.07
Rimini	0.04	0.04	-0.06	0.13	0.06
Roma	-0.10	-0.10	-0.19	0.00	0.06
Rovigo	0.07	0.07	-0.02	0.17	0.06
Salerno	-0.91	-0.90	-1.06	-0.76	0.11
Sassari	-0.60	-0.60	-0.73	-0.47	0.09
Savona	0.02	0.02	-0.07	0.11	0.06
Siena	0.10	0.10	0.19	0.01	0.06
Siracusa	-1.03	-1.03	-1.21	-0.87	0.12
Sondrio	0.22	0.22	0.12	0.31	0.07
Sud Sardegna	-0.30	-0.31	-1.44	0.84	0.93
Taranto	-1.00	-0.99	-1.17	-0.83	0.12
Teramo	-0.27	-0.27	-0.37	-0.17	0.07
Terni	-0.22	-0.22	-0.32	-0.12	0.07
Torino	0.03	0.03	-0.07	0.12	0.06
Trapani	-1.30	-1.29	-1.52	-1.09	0.15
Trento	0.20	0.20	0.11	0.30	0.06
Treviso	0.13	0.13	0.04	0.23	0.06
Trieste	0.12	0.12	0.03	0.22	0.06
Udine	0.08	0.08	-0.01	0.17	0.06
Varese	0.15	0.15	0.06	0.24	0.06
Venezia	0.14	0.14	0.04	0.23	0.07
Verbano-Cusio-Ossola	0.01	0.01	-0.08	0.10	0.06
Vercelli	0.00	0.00	-0.10	0.09	0.06
Verona	0.24	0.24	0.15	0.34	0.06
Vibo Valentia	-1.18	-1.18	-1.38	-1.00	0.13
Vicenza	0.24	0.24	0.15	0.34	0.07
Viterbo	-0.40	-0.39	-0.51	-0.29	0.08

Table 4.E.3: Summary of posterior distribution of the environmental dimension's latent variable (composite indicator). CAR model B. Year 2019

Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR
Agrigento	-0.77	-0.90	-0.44	-1.23	0.79	Foggia	-0.32	-0.26	0.26	-0.84	1.10	Pescara	-0.43	-0.49	-0.05	-0.80	0.76
Alessandria	-0.82	-0.87	-0.42	-1.22	0.80	Forlì-Cesena	-0.21	-0.24	0.12	-0.54	0.66	Piacenza	-0.03	0.00	0.30	-0.36	0.66
Ancona	0.21	0.17	0.57	-0.16	0.73	Frosinone	-0.83	-0.92	-0.41	-1.32	0.91	Pisa	-0.75	-0.80	-0.32	-1.20	0.89
Aosta	-0.43	-0.45	-0.10	-0.77	0.67	Genova	0.29	0.44	1.05	-0.28	1.33	Pistoia	0.12	0.16	0.51	-0.25	0.76
Arezzo	-0.73	-0.80	-0.35	-1.13	0.78	Gorizia	0.17	0.08	0.59	-0.35	0.94	Pordenone	1.04	1.01	1.40	0.67	0.73
Ascoli Piceno	-0.32	-0.30	0.06	-0.68	0.75	Grosseto	-0.36	-0.30	0.04	-0.75	0.79	Potenza	-0.14	-0.28	0.22	-0.63	0.86
Asti	0.07	0.04	0.45	-0.35	0.80	Imperia	-0.84	-0.80	-0.48	-1.21	0.73	Prato	0.35	0.29	0.72	-0.04	0.76
Avellino	0.15	0.22	0.56	-0.26	0.81	Isernia	-1.41	-1.46	-1.01	-1.80	0.79	Ragusa	-0.42	-0.35	-0.05	-0.79	0.74
Bari	-0.84	-0.83	-0.49	-1.20	0.71	L'Aquila	-0.05	-0.06	0.27	-0.36	0.63	Ravenna	-0.25	-0.31	0.13	-0.62	0.75
Barletta-Andria-Trani	0.11	0.25	0.71	-0.35	1.05	La Spezia	-0.11	-0.15	0.28	-0.50	0.78	Reggio di Calabria	-0.70	-0.83	-0.33	-1.21	0.88
Belluno	0.68	0.66	1.04	0.29	0.75	Latina	-0.73	-0.76	-0.40	-1.07	0.67	Reggio nell'Emilia	0.82	0.80	1.15	0.50	0.66
Benevento	-0.24	-0.28	0.16	-0.69	0.85	Lecce	-0.27	-0.21	0.29	-0.79	1.07	Rieti	-0.68	-0.64	-0.35	-1.01	0.66
Bergamo	0.35	0.32	0.70	-0.01	0.72	Lecco	-0.06	-0.10	0.31	-0.44	0.75	Rimini	-0.09	-0.14	0.25	-0.47	0.72
Biella	0.26	0.28	0.63	-0.11	0.74	Livorno	-0.69	-0.70	-0.35	-1.02	0.67	Roma	1.36	1.48	2.16	0.67	1.49
Bologna	0.18	0.19	0.51	-0.15	0.66	Lodi	1.05	1.07	1.41	0.72	0.69	Rovigo	-0.07	-0.10	0.29	-0.42	0.71
Bolzano Bozen	0.93	1.00	1.43	0.47	0.96	Lucca	0.01	-0.05	0.37	-0.37	0.74	Salerno	-0.31	-0.37	0.05	-0.69	0.74
Brescia	1.55	1.66	2.16	1.02	1.14	Macerata	0.02	-0.01	0.41	-0.42	0.84	Sassari	-0.42	-0.46	-0.07	-0.76	0.68
Brindisi	-0.23	-0.19	0.13	-0.62	0.75	Mantova	0.74	0.69	1.13	0.32	0.80	Savona	-0.81	-0.82	-0.42	-1.19	0.78
Cagliari	-0.32	-0.40	0.10	-0.78	0.89	Massa-Carrara	-0.68	-0.66	-0.33	-1.05	0.72	Siena	-0.30	-0.29	0.08	-0.66	0.74
Caltanissetta	-0.54	-0.49	-0.14	-0.95	0.81	Matera	0.15	0.02	0.47	-0.35	0.82	Siracusa	-1.18	-1.14	-0.79	-1.58	0.78
Campobasso	-0.90	-0.86	-0.56	-1.22	0.66	Messina	-0.99	-0.96	-0.67	-1.33	0.66	Sondrio	-0.11	-0.23	0.23	-0.59	0.82
Caserta	0.20	0.18	0.53	-0.12	0.65	Milano	1.19	1.31	1.81	0.68	1.13	Sud Sardegna	-0.99	-1.11	-0.66	-1.44	0.78
Catania	-1.28	-1.33	-0.90	-1.65	0.75	Modena	1.12	1.16	1.49	0.75	0.74	Taranto	-1.44	-1.48	-1.03	-1.84	0.81
Catanzaro	-0.39	-0.48	0.00	-0.80	0.81	Monza e della Brianza	0.78	0.75	1.13	0.42	0.71	Teramo	0.18	0.16	0.53	-0.18	0.71
Chieti	-0.84	-0.89	-0.40	-1.27	0.87	Napoli	-0.68	-0.74	-0.35	-1.03	0.68	Terni	0.39	0.30	0.86	-0.16	1.02
Como	1.91	1.97	2.50	1.34	1.16	Novara	-0.13	-0.18	0.22	-0.49	0.72	Torino	0.34	0.38	0.72	-0.06	0.77
Cosenza	0.52	0.58	0.99	0.06	0.94	Nuoro	0.88	0.91	1.25	0.50	0.75	Trapani	-1.18	-1.11	-0.76	-1.59	0.83
Cremona	0.10	0.06	0.47	-0.28	0.75	Oristano	0.21	0.17	0.64	-0.27	0.92	Trento	0.63	0.52	1.04	0.15	0.90
Crotone	-2.45	-2.46	-2.11	-2.80	0.69	Padova	0.22	0.20	0.56	-0.11	0.67	Treviso	0.80	0.78	1.21	0.43	0.78
Cuneo	0.23	0.20	0.58	-0.12	0.69	Palermo	-1.17	-1.12	-0.84	-1.51	0.67	Trieste	0.35	0.41	0.74	-0.02	0.76
Enna	-1.34	-1.30	-0.97	-1.69	0.73	Parma	0.39	0.45	0.82	0.05	0.77	Udine	0.15	0.10	0.49	-0.19	0.67
Fermo	-0.19	-0.24	0.23	-0.60	0.83	Pavia	-0.24	-0.17	0.11	-0.58	0.69	Varese	0.10	0.03	0.48	-0.29	0.77
Ferrara	0.15	0.09	0.55	-0.24	0.80	Perugia	-0.01	-0.08	0.40	-0.47	0.86	Venezia	0.91	0.92	1.24	0.58	0.67
Firenze	0.29	0.28	0.65	-0.07	0.72	Pesaro e Urbino	-0.36	-0.40	-0.01	-0.74	0.73	Verbano-Cusio-Ossola	0.07	-0.02	0.49	-0.39	0.88
												Vercelli	0.11	0.08	0.50	-0.24	0.74
												Verona	-0.04	-0.07	0.30	-0.39	0.69
												Vibo Valentia	-0.49	-0.49	-0.14	-0.83	0.69
												Vicenza	0.67	0.67	1.04	0.29	0.75

Table 4.E.4: Summary of posterior distribution of the latent variable (composite indicator) for the overall well-being dimension. CAR model B. Year 2019

Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR	Province	mean	median	25%	75%	IQR
Agrigento	-1.84	-1.86	-1.70	-1.98	0.28	Foggia	-1.52	-1.53	-1.39	-1.66	0.27	Pescara	-0.21	-0.22	-0.11	-0.33	0.22
Alessandria	0.46	0.47	0.57	0.36	0.21	Forlì-Cesena	0.95	0.95	1.05	0.85	0.21	Piacenza	0.85	0.84	0.95	0.74	0.21
Ancona	0.40	0.39	0.50	0.29	0.21	Frosinone	-0.86	-0.87	-0.74	-0.98	0.24	Pisa	0.68	0.67	0.78	0.57	0.21
Aosta	0.78	0.78	0.89	0.68	0.21	Genova	0.32	0.31	0.42	0.21	0.21	Pistoia	0.37	0.37	0.47	0.27	0.20
Arezzo	0.70	0.72	0.81	0.60	0.21	Gorizia	0.55	0.56	0.66	0.45	0.21	Pordenone	0.98	0.98	1.08	0.88	0.21
Ascoli Piceno	0.20	0.19	0.29	0.09	0.21	Grosseto	0.48	0.48	0.58	0.38	0.20	Potenza	-0.95	-0.96	-0.83	-1.07	0.24
Asti	0.73	0.75	0.85	0.63	0.22	Imperia	0.00	0.01	0.11	-0.10	0.21	Prato	0.84	0.84	0.95	0.74	0.21
Avellino	-0.76	-0.78	-0.65	-0.88	0.23	Isernia	-0.63	-0.63	-0.53	-0.74	0.21	Ragusa	-1.02	-1.01	-0.89	-1.13	0.23
Bari	-0.51	-0.53	-0.40	-0.63	0.23	L'Aquila	-0.06	-0.07	0.03	-0.17	0.20	Ravenna	0.89	0.89	1.00	0.79	0.20
Barletta-Andria-Trani	-1.24	-1.25	-1.12	-1.37	0.25	La Spezia	0.48	0.47	0.58	0.38	0.20	Reggio di Calabria	-1.90	-1.91	-1.77	-2.04	0.27
Belluno	1.04	1.05	1.15	0.94	0.21	Latina	-0.37	-0.38	-0.27	-0.48	0.21	Reggio nell'Emilia	0.98	0.97	1.08	0.87	0.21
Benevento	-1.05	-1.06	-0.93	-1.17	0.24	Lecce	-1.26	-1.28	-1.14	-1.40	0.26	Rieti	-0.25	-0.26	-0.16	-0.36	0.20
Bergamo	0.87	0.87	0.98	0.76	0.22	Lecco	0.94	0.94	1.04	0.84	0.21	Rimini	0.62	0.61	0.72	0.51	0.21
Biella	0.60	0.60	0.70	0.50	0.20	Livorno	0.29	0.28	0.39	0.18	0.20	Roma	0.38	0.36	0.47	0.26	0.21
Bologna	1.13	1.10	1.23	0.99	0.24	Lodi	0.63	0.63	0.73	0.53	0.20	Rovigo	0.59	0.60	0.70	0.49	0.20
Bolzano Bozen	1.29	1.30	1.42	1.18	0.23	Lucca	0.21	0.20	0.31	0.10	0.20	Salerno	-1.17	-1.18	-1.05	-1.30	0.25
Brescia	0.92	0.92	1.03	0.82	0.21	Macerata	0.52	0.51	0.61	0.41	0.20	Sassari	-0.61	-0.62	-0.51	-0.73	0.22
Brindisi	-0.93	-0.94	-0.82	-1.05	0.23	Mantova	0.82	0.82	0.93	0.72	0.20	Savona	0.55	0.54	0.65	0.44	0.20
Cagliari	-0.40	-0.44	-0.29	-0.56	0.27	Massa-Carrara	0.37	0.36	0.46	0.27	0.20	Siena	0.66	0.67	0.77	0.56	0.21
Caltanissetta	-1.94	-1.94	-1.79	-2.08	0.29	Matera	-0.41	-0.42	-0.31	-0.53	0.22	Siracusa	-1.45	-1.46	-1.32	-1.58	0.26
Campobasso	-0.54	-0.55	-0.44	-0.65	0.21	Messina	-1.88	-1.88	-1.74	-2.03	0.29	Sondrio	0.83	0.84	0.95	0.72	0.22
Caserta	-1.67	-1.68	-1.53	-1.81	0.28	Milano	1.04	1.01	1.14	0.90	0.24	Sud Sardegna	-0.10	-0.10	0.00	-0.20	0.20
Catania	-1.65	-1.66	-1.50	-1.79	0.28	Modena	0.89	0.88	1.00	0.78	0.22	Taranto	-1.38	-1.39	-1.25	-1.50	0.25
Catanzaro	-1.17	-1.18	-1.05	-1.30	0.25	Monza e della Brianza	0.79	0.77	0.89	0.68	0.21	Teramo	-0.01	-0.02	0.08	-0.12	0.20
Chieti	-0.25	-0.25	-0.15	-0.36	0.21	Napoli	-1.81	-1.82	-1.67	-1.96	0.29	Terni	0.14	0.12	0.24	0.03	0.21
Como	0.77	0.75	0.86	0.66	0.20	Novara	0.71	0.70	0.81	0.61	0.21	Torino	0.58	0.57	0.67	0.47	0.20
Cosenza	-1.39	-1.41	-1.26	-1.53	0.27	Nuoro	-0.73	-0.74	-0.62	-0.85	0.23	Trapani	-1.94	-1.94	-1.79	-2.08	0.29
Cremona	0.77	0.77	0.87	0.67	0.20	Oristano	-0.79	-0.81	-0.68	-0.92	0.24	Trento	0.93	0.91	1.02	0.81	0.21
Crotone	-2.17	-2.17	-2.01	-2.32	0.30	Padova	0.89	0.88	0.99	0.78	0.21	Treviso	0.76	0.75	0.86	0.65	0.21
Cuneo	0.93	0.95	1.05	0.83	0.22	Palermo	-1.85	-1.86	-1.70	-1.99	0.29	Trieste	0.79	0.78	0.89	0.68	0.21
Enna	-1.81	-1.82	-1.66	-1.95	0.29	Parma	0.92	0.90	1.03	0.80	0.23	Udine	0.68	0.67	0.78	0.57	0.21
Fermo	0.59	0.60	0.70	0.49	0.21	Pavia	0.71	0.71	0.81	0.61	0.20	Varese	0.80	0.79	0.90	0.70	0.20
Ferrara	0.75	0.74	0.86	0.65	0.21	Perugia	0.56	0.55	0.66	0.45	0.21	Venezia	0.76	0.76	0.86	0.65	0.21
Firenze	0.95	0.93	1.05	0.83	0.22	Pesaro e Urbino	0.56	0.55	0.65	0.46	0.20	Verbano-Cusio-Ossola	0.51	0.51	0.61	0.41	0.20
												Vercelli	0.46	0.46	0.56	0.36	0.20
												Verona	0.99	0.98	1.09	0.88	0.21
												Vibo Valentia	-1.72	-1.72	-1.57	-1.85	0.28
												Vicenza	0.96	0.95	1.06	0.85	0.21

Chapter 5

Quality of Government for Environmental Well-Being? Subnational Evidence from European Regions

5.1 Introduction

Building effective, transparent, accountable, and uncorrupted public institutions is a core target of the United Nations' Sustainable Development Goal 16. Besides being a globally agreed standalone policy target, institutional quality is also widely understood to be a prerequisite to achieving the broader goals of the sustainable development agenda (e.g. DESA, 2019; UN, 2012). The basic argument is that successfully reaching any policy objective becomes more complicated under dysfunctional public institutions. Along these lines of thought, increasing citizens' well-being necessitates sound public institutions, often referred to as 'quality of government.' And indeed, this view is supported by abundant scholarly research (e.g. Charron et al., 2015; Evans and Rauch, 1999; Holmberg and Rothstein, 2011)

A nation's well-being is commonly conceived as a combination of economic, social, and environmental dimensions, and each of them as a positive polarity with respect to the overall well-being (e.g., Ciommi et al., 2022; Giovannini, 2015; Michalos, 1997). However, the vast majority of past studies on the relationship between quality of government and nations' wellbeing have focused on the first two dimensions. Less research attention has been paid to the link between the quality of government and the natural environment, even if environmental degradation is one of the biggest global concerns of our time.

This paper aims to shed light on the association between quality of government and environmental well-being by providing a theoretically and statistically more rigorous approach than in previous studies. Ultimately, our goal is to investigate whether quality of government —defined as "the extent to which states perform their required activities and administer public services in an impartial and uncorrupted manner" (Charron et al., 2015: 316)— is a key predictor of environmental wellbeing.

Current literature points to inconclusive results. Some suggest that quality of government increases environmental wellbeing (Povitkina, 2018), but others find no evidence of any significant effect (Peiró-Palomino et al., 2020), and still, others find evidence of an inverse link between the quality of government and environmental wellbeing (Holmberg et al., 2009).

Besides the scarce research attention and inconclusiveness, we identify three additional shortcomings that might have affected research results. First, there is a

lack of cross-country quantitative studies at the subnational level. Second, there is inadequate consideration of the multidimensionality of environmental well-being. Third, spatial correlation in environmental well-being is seldom taken into account.

On the first shortcoming, the lion's share of cross-country studies on the topic is focused on the country-level, disregarding subnational variation within and across countries. Yet experts have recently demonstrated that both quality of government (Charron et al., 2019) and wellbeing (Iammarino et al., 2019) vary significantly from one country to another and within them. These diversities have notably increased in Europe in the last decade (Iammarino et al., 2019), which makes the European context particularly relevant for our investigation.

We argue that a comprehensive picture of the relationship between the quality of government and environmental well-being requires investigating subnational dynamics. Our study fills this gap by investigating European regions' institutions-environment nexus at the subnational level (NUTS-2). The downside of focusing on European regions is that our findings are confined to a specific context and may not be generalisable to other parts of the world.

Concerning the second shortcoming, we contend that current knowledge of the institutions-environment focuses on an excessively narrow empirical understanding of the environment, mainly in terms of air pollution. Both national (Azimi et al., 2023) and sub-national (Peiró-Palomino et al., 2020) cross-country studies on the topic tend to measure environmental well-being with exposure to a specific air pollutant like carbon dioxide or a combination of multiple pollutants. It is self-evident, however, that air pollution does not represent environmental well-being in its entirety. We address this problem by using a multidimensional approach to environmental well-being.

We identify four core dimensions of environmental well-being (see ISTAT, 2021) and measure them with multiple representative indicators. Specifically, we look at the dimensions of (1) air quality, (2) water quality, (3) soil quality, and (4) energy and climate change. Instead of using single variables as proxies, we construct a set of composite indicators to represent each of them as comprehensively as possible. We take a spatial Bayesian latent variable approach to composite indicator construction (Hogan and Tchernis, 2004, Davis et al., 2021). Compared to frequentist methods, our approach results in more precise estimates and provides information on their

uncertainty.

Third, most studies on environmental wellbeing ignore spatial characteristics and interlinkages among neighboring countries or regions, causing potentially biased results. This is highly problematic, because recent empirical evidence suggests that wellbeing tends to be spatially interdependent, at least in Europe (Peiró-Palomino et al. 2020). In linear regressions, the presence of spatial correlation in the dependent variable creates duplicate information, inflating the variance of the statistical model and damaging the validity of the estimated standard errors (Moran, 1950).

We assess the magnitude of spatial correlation in our environmental data, finding significant spatial patterns on the European subnational surface. Thus, we first model this spatial dependence to increase precision in the environmental composite indicators estimates. Second, we use them as dependent variables in subsequent spatial regressions of environmental well-being on the quality of government. Our spatial regression models provide robust evidence that well-functioning and effective public institutions are strongly related to environmental well-being —especially the quality of air and soil— even when accounting for the data's spatial characteristics

The findings of our study have significant relevance to the policy debate. Even though improving economic, social, and environmental wellbeing is a globally agreed policy objective, combining economic performance with environmental sustainability has proved challenging. According to experts, environmental concerns will increase in the coming years and climate change has been named as one of the major threats for humanity (MacAskill, 2022). Understanding how to advance environmental well-being must be thus one of the main priorities of policymakers in Europe and worldwide. The findings of our study indicate that by strengthening the regional quality of government, European policymakers can contribute to protecting their societies from environmental degradation.

Our paper proceeds as follows. Section 5.2 summarises the literature on quality of government and environmental well-being and presents our theoretical argument. Section 5.3 delves into the data and methods. Section 5.4 presents the empirical results and discusses the findings. Section 5.5 summarises our main findings and briefly reflects on their implications for policymakers and future studies on the topic.

5.2 Background

5.2.1 Literature review

An important body of literature at the crossroads between economics and political science underscores the relevance of well-functioning and effective public institutions for positive development outcomes (e.g. Acemoglu and Robinson, 2012; Evans, 1995; Rothstein, 2011).

Findings show that at the national level well-functioning public institutions foster economic growth (e.g. Evans and Rauch, 1999), reduce poverty (e.g. Chong and Calderon, 2000) and income inequality (e.g., Panaro and Vaccaro, 2022), strengthen food security (e.g., Sachs, 2015), increase life satisfaction (e.g., Helliwell and Huang, 2008), and improve public health outcomes (e.g. Holmberg and Rothstein, 2011) among many other indicators of human wellbeing.

Only a handful of cross-country studies suggest the opposing view that institutional quality does not affect or negatively affects economic and social well-being. For instance, Kraay (2004) finds that poverty-increasing distributional change occurs especially in countries with high institutional quality, and Huang (2016) finds that in many Asian countries, impartial institutions do not affect economic growth, except for South Korea, where corruption seems to foster economic growth. The conventional view, supported by the majority of empirical findings, is then that quality of government advances economic and social wellbeing. Yet, as previously noted, this body of research focuses on economic and social well-being, overlooking environmental dimension.

The few country-level studies that have focused on the link between the quality of government and environmental aspects of well-being are contradicting. On the one hand, well-functioning public institutions appear to reduce carbon dioxide emissions (Azimi et al., 2023), improve drinking water quality (Povitkina & Bolkvadze, 2019), reduce deforestation (Meyer et al., 2003), generate more stringent environmental policies (Pellegrini & Gerlagh, 2006). On the other hand, institutional quality could be related to higher carbon dioxide emissions (Holmberg et al., 2009), and the magnitude and sign of the net effect of corruption on pollution might depend on the level of national income (Cole, 2007).

While the above studies have increased our understanding of the relationship

between quality of government and well-being, their results are based on countries as the units of analysis, ignoring subnational disparities. Only recently have experts begun to stress the importance of considering subnational variation regarding both the quality of government and the well-being. The broader well-being literature has shifted towards higher territorial level disaggregation (Mazziotta et al., 2021), and a similar trend has emerged in the literature on institutional quality (see, e.g., Charron et al., 2019; Iddawela et al., 2021). Environmental problems, in particular, tend to differ among subnational territories (Halkos et al., 2015) and quality of government matters for national and regional development (Rodríguez-Pose, 2013).

Only in the last few years has a substantial body of studies explored the subnational variation in the institutions' quality and well-being nexus, especially in Europe. These studies generally support the view that a higher quality of government is linked to higher economic and social well-being.

Charron et al. (2015) show that regional quality of government is positively correlated with income, social trust, and education and inversely correlated with infant mortality, unemployment, and economic and gender inequality. Other studies push forward by controlling for confounding factors. Peiró-Palomino et al. (2020) show that subnational quality of government positively affects the most common aspects of well-being: education, jobs, income, safety, civic engagement, access to services, housing, and community support. Other scholars analyse more specific aspects of economic and social wellbeing and show that subnational institutional quality fosters entrepreneurship (Nistotskaya et al., 2015), boosts economic resilience (Ezcurra & Rios, 2019), increases trade flows((Barbero et al., 2021), curbs social exclusion (Di Cataldo & Rodríguez-Pose, 2017), reduces income inequality (Parente, 2019), strengthens innovation capacity (Rodríguez-Pose & Di Cataldo, 2015), deepens citizens' trust in public administration (Van de Walle & Migchelbrink, 2020) and in other people (Lombardo & Ricotta, 2021), and facilitates convergence among regions (Charron et al., 2019) across subnational territories in Europe.

Yet again, even if these studies highlight the importance of considering subnational variation, they focus only on economic and social dimension of well-being, neglecting the environment. One of the few studies that considers the subnational relationship between environmental well-being and institutional quality in Europe finds no evidence of any significant association (Peiró-Palomino et al. 2020). In

this study, the authors measure environmental well-being as the estimated average exposure to air pollution. Another subnational study on the topic finds that environmental performance in NUTS-1 regions in France, Germany, and the UK is curvilinearly related to institutional quality: the association is positive in regions with ineffective and dysfunctional institutions, but as subnational quality of government increases the link between institutions and environmental performance becomes unexpectedly inverse (Halkos et al., 2015). Therefore, the relationship between quality of government and environmental well-being at the subnational level in Europe is thus unclear and understudied.

5.2.2 Theoretical framework

In this study, we hypothesize that citizens in regions with poorer government quality suffer from lower environmental well-being. The precise mechanisms linking higher quality of government to higher environmental well-being are likely manifold, but here we discuss three of the most important ones.

First, we expect quality of government to foster environmental well-being through higher effectiveness. Corruption, which is at odds with quality of government (Rothstein & Teorell, 2008) reduces the effectiveness of environmental policies because "if corrupt officials accept bribes for looking the other way, individuals and organizations who pollute or destroy natural resources can avoid fines without changing their behaviour" (Aklin et al., 2014). Regions with high-quality of government should be thus more effective in enforcing environmental regulations than regions with low-quality.

Second, we expect quality of government to improve environmental well-being through more competent civil servants. Given that quality of government, understood as impartiality, entails meritocratic recruitment in the civil service (Rothstein, 2011), regions with higher quality of government should have highly skilled bureaucrats compared to regions with poor quality of government. We can quite confidently expect —ceteris paribus— competent civil servants to be better at implementing policies than incompetent civil servants. This reasoning applies to any policies, including those targeting environmental well-being.

Third, public institutions in which civil servants are not only recruited based on meritocracy but also have predictable career structures —instead of institutions

in which civil servants need to rely on political connections or other personal ties to advance on the career ladder— incentivise to work harder and are conducive to an organization that is in general more concerned about long-term objectives rather than short-term private gains (Cornell et al., 2020). We expect this last point to be especially important for facilitating environmental well-being, which often requires long-term commitment by policymakers and civil servants who are ultimately responsible for successful policy implementation. Moreover, the prospect of a predictable career could further boost civil servants' average level of competence by increasing the attractiveness of working in the public sector.

5.3 Empirical approach

5.3.1 Data

Measuring environmental well-being in European regions is tricky due to the absence of multidimensional subnational cross-country data. At the time of this writing, no comprehensive measure of environmental well-being at the subnational European level exists. To cope with the lack of data, scholars often use single proxy measures that capture only specific parts of environmental well-being. Yet, as already discussed, such proxies cannot represent environmental well-being precisely and, at best, have weak content validity. For instance, one of the most well-known cross-national datasets on subnational wellbeing —OECD's Regional Wellbeing Dataset— equates environmental wellbeing simplistically with air pollution by particulate matter (OECD, 2020).

To tackle this problem and to better measure multidimensional subnational environmental wellbeing in Europe, we first scrutinise and collect a battery of subnational indicators related to air quality, water quality, soil quality, and energy and climate change —the four main aspects of environmental wellbeing— and then develop a set of composite indicators to capture these four aspects as comprehensively as possible. The next section presents the methods used to construct our composite indicators. Here, we present the elementary indicators that we have collected from four different sources (Joint Research Centre, European Observation Network for Territorial Development and Cohesion, European Environment Agency, and Eurostat).

After meticulously reviewing publicly available subnational data on the environment in Europe, we collected 16 indicators related to our four dimensions of environmental well-being. Ultimately, our measurement framework consists of six elementary indicators for air quality, four for water quality, four for soil quality, and two for energy and climate change. Appendix 5.A provides a detailed description of the elementary indicators and their sources.

We measure Air quality as a composite index that synthesises the following six elementary indicators: nitrogen dioxide (NO2) removal capacity by urban vegetation (measured in the year 2020), urban population exposed to particle matter (PM) of size $10\,\mu m$ (2020), air concentration of PM of size $2.5\,\mu m$ (2016), air concentration of PM of size $10\,\mu m$ (2016), air concentration of ozone (2017), and air concentration of NO2 (2017). The composite indicator of water quality consists of the following four elementary indicators: water productivity (2020), quality of drinking water (2020), sewage treatment (2014/2016), and freshwater consumption (2020). Soil quality has four elementary indicators: capacity of ecosystems to avoid soil erosion (2020), severe soil erosion by water (2016), artificial surfaces inside Natura 2000 protected areas (2018), and organic farming (2016). Energy and climate change synthesize two elementary indicators: energy recovery capacity (2018) and potential vulnerability to climate change (2071-2100 prediction).

To measure subnational quality of government in European regions, we use arguably the most widely used and well-constructed dataset on the topic: the European Quality of Government Index Survey Dataset Charron et al., 2019. The dataset, published by University of Gothenburg's Quality of Government Institute, provides subnational data for European Union countries in four different years —2010, 2013, 2017, and 2021. The European Quality of Government Index (EQI) entails first aggregating individual survey question scores into three dimensions of quality of government and then synthesizing these three indicator's components —Quality, Impartiality, and Corruption— into a composite indicator. EQI captures institutional quality in Europe at the NUTS-2 level in 238 subnational territories and runs from low to high on a z-score scale (mean of 0; standard deviation of 1). In line with our theoretical expectations on the direction of the link between quality of government and environmental wellbeing, in our main analysis, we use the 2017 measure of quality of government because most indicators of environmental

wellbeing refer to years 2017-2020.

5.3.2 Estimation methods

One of the shortcomings in past subnational studies on the relationship between quality of government and environmental wellbeing is the lack of a comprehensive and synthetic measure of environmental wellbeing and its core dimensions. Hence, through a data-driven approach based on Bayesian latent factor models, we construct four environmental composite indicators, one for each of the four environmental domains —air quality, water quality, soil quality, and energy and climate change— summarising the information of the above discussed 16 elementary environmental indicators. Then, we run a series of spatial lag regression models to shed light on the link between quality of government and environmental well-being in 233 European regions.

We hypothesise the existence of spatial spillovers, so that environmental conditions in each region are partially determined by the environmental conditions of its neighbouring regions. To verify this initial assumption, we test for spatial autocorrelation in the 16 environmental elementary indicators through the Global Moran I test (Moran, 1950), which provides significant results for all the indicators except the Urban population exposed to PM10 (Table 5.B.1, Appendix 5.B). We also compute spatial correlograms to assess the degree at which spatial autocorrelations change as a function of distance (Figures A1-A4, Appendix 5.B). Based on these results, we follow Hogan and Tchernis (2004) and estimate a Bayesian latent factor model for spatially correlated data. The Bayesian approach naturally adapts to the hierarchical structure of the latent factor model. Moreover, through priors' distribution specification, the Bayesian approach allows embedding information on the spatial structure of the data at the latent factor level, resulting in more precise latent factors' estimates (Hogan and Tchernis, 2004). Finally, the Bayesian approach also has the advantage of quantifying the uncertainty about the latent factor scores through the posterior parameters' distribution information.

We apply the spatial Bayesian latent factor model by Hogan and Tchernis, 2004 separately for each of the four environmental domains. In particular, for each European region i, where $i = 1, \ldots, N$, with $N = 233$, let Y_{ip} denote the elementary environmental indicator p in region i and $p = 1, \ldots, P$, with $P = 6, 4, 4, 2$, for air

quality, water quality, soil quality, energy and climate change, respectively. Hence, $Y_i = (Y_{i1}, \ldots, Y_{iP})^T$ is the vector of the observed outcome variables for region i. For each of the four environmental domains, we assume the existence of a latent variable δ_i that fully characterizes the environmental well-being level, manifesting itself through Y_i. We represent the model in a hierarchical form. At the first level, we have:

$$Y_i \mid \mu_i, \delta_i, \Sigma \sim \text{Multivariate Normal}(\mu_i + \lambda\delta_i, \Sigma), \tag{5.1}$$

where μ_i is a $P \times 1$ mean vector, λ is a $P \times 1$ vector of factor loadings, and $\Sigma = \text{Diag}(\sigma_1^2, \ldots, \sigma_P^2)$ is a diagonal matrix measuring residual variation in Y_i. Assuming Σ diagonal implies independence among the elements of Y_i conditionally on δ_i.

Let $\delta = (\delta_1, \ldots, \delta_N)^T$ be the vector of regions' latent environmental wellbeing. We add spatial information to the latent factor prior distribution by assuming:

$$\delta \sim \text{Multivariate-Normal}(0_N, \Psi), \tag{5.2}$$

where Ψ is a $N \times N$ spatial variance-covariance matrix having 1's on the diagonal and $\psi_{i,j} = \text{corr}(\delta_i, \delta_j)$ on the off-diagonal. When $\Psi = I_N$ the model assumes spatial independence across regions' environmental wellbeing levels. The literature proposes several alternatives to introduce spatial correlation based on a marginal or conditional specification of spatial dependency. Here, we consider a conditional specification, which defines spatial dependence in the latent environmental factor as a conditional autoregressive process in which a region's environmental well-being is determined by the average environmental well-being of its neighbours (Besag et al., 1991; Cressie and Chan, 1989) so that:

$$\delta_i \mid \delta_j : j \in \mathcal{R}_i \sim \text{Normal}(\Sigma_{j \in R_i} \omega (n_j/n_i)^{1/2} \delta_j, 1/n_i), \tag{5.3}$$

where ω is a spatial correlation parameter to be estimated, n_i is the number of region i's neighbouring regions and R_i is the set of indices for regions that are neighbours of region i. This conditional autoregressive specification implies a spatial variance-covariance matrix $\Psi = (I_n - \omega(n_i * n_j)^{1/2} R)^{-1}$, where R is the neighborhood adjacency matrix with elements $r_{ij} = r_{ji} = 1$ if region i and region j share a common boundary, and 0 otherwise.

Finally, a characteristic of the Bayesian framework is the introduction of prior dis-

tributions on all the model parameters. In our case, we have set $\lambda_p \sim \text{Normal}(g, G)I(\lambda_1 > 0)$, $\sigma_p^2 \sim$ Inverse-Gamma $(\alpha/2, \beta/2)$, $\mu_p \sim \text{Normal}(0, V_\mu)$. The primary scope of prior distributions is to include subjective opinions on the parameters of interest. Yet, to let the data 'speak for themselves', we use diffuse priors by choosing $g = 0$, $G = 1000$, $\alpha = 1/1000$, $\beta = 1/1000$, and $V_\mu = 1000$.

We estimate the model posterior distribution using Markov Chain Monte Carlo methods. Specifically, we use a Gibbs sampling algorithm that includes Metropolis Hasting steps to estimate spatial parameter ω. At each step of the sampling algorithm, we obtain draws from the conditional posterior distribution of the model parameters (i.e. λ, μ, Σ and the latent environmental factor δ). We use these draws to build the posterior distributions of all model parameters after accounting for a burn-in period before convergence. We simulate 6000 draws and burn 3000 of them. Another key advantage of this model is that it can handle missing values through a posterior imputation procedure. The procedure replaces missing elementary indicator values with draws from the first level equation conditional on current iterations' draws of the latent factor and the other models' parameters (see also Davis et al., 2021).

Next, we retrieve the mean from the estimated environmental composite indicators' posterior distributions and use it as the dependent variable in spatial lag models to analyse the linear dependence between environmental wellbeing and quality of government. Using spatial regression models instead of simple ordinary least square regressions, we address the bias arising from the strong spatial association in the environmental indicators illustrated in Figure 5.4.1. Indeed, spatial correlation in the outcome violates assumptions of homoskedasticity and independence. We perform Lagrange Multiplier tests to select the most appropriate spatial model. Results from this test show that the spatial lag model is the most appropriate for the data at hand (Table 5.B.2, Appendix 5.B).

In the Spatial Lag model, the dependent variable in a region i is affected by the independent variables in the same region i and those in its neighbouring regions (LeSage & Pace, 2009). For each environmental domain, set $\hat{\delta}_i = E[\delta_i \mid Y, \mu, \lambda, \Sigma]$, where i indicates the region. QoG_i is quality of government indicator in region i.

Then, the spatial lag model takes the following form:

$$\hat{\delta}_i = \rho \sum_{j \in \mathcal{R}_i} (w_{ij} \hat{\delta}_j) + \beta \mathbf{QoG}_i + \sum_d (\gamma_d x_{id}) + \epsilon, \tag{5.4}$$

In this model, ρ reflects the strength of spatial dependence among neighbouring regions. When $\rho \neq 0$, the coefficient of interest β captures the impact of quality of government on environmental well-being as a combination of direct and indirect spatial effects (LeSage & Pace, 2009). $\mathcal{R}_i$ represents the set of neighbouring regions of i, as in the previous model. w_{ij} is a weighted neighborhood indicator, with $0 < w_{ij} \leq 1$ if region i and region j share a common boundary, and $w_{ii} = 0$. For the error term, we assume $\epsilon \sim N(0, \sigma)$. We add region-specific controls in x_{id}, namely GDP/capita, population density, region area, unemployment rate, the share of employment in agriculture and manufacturing, and expenditure in research and development. We selected these control variables based on the literature and data availability. We also run 'baseline' models without control variables and models in which we control for the effect of the 'remaining' aspects of environmental wellbeing. Additionally, we also analyse the effect of quality of government on an overall index of environmental wellbeing and test the robustness of our results to quality of government measured in different years.

Therefore, we disentangle the average direct and indirect spatial effect from the estimated β following the methodology of LeSage and Pace (2009). The direct effect is simply the magnitude of the link between quality of government and environmental well-being in a given region, excluding any effects via neighbouring regions. Its value derives from the mean of the diagonal terms of the matrix of partial derivatives. The indirect association instead reflects the magnitude of the impact on environmental wellbeing in a given region, rendered by a change in quality of government in the neighboring regions. This term derives from the difference between total and direct effects (Golgher & Voss, 2016).

5.4 Results and Discussion

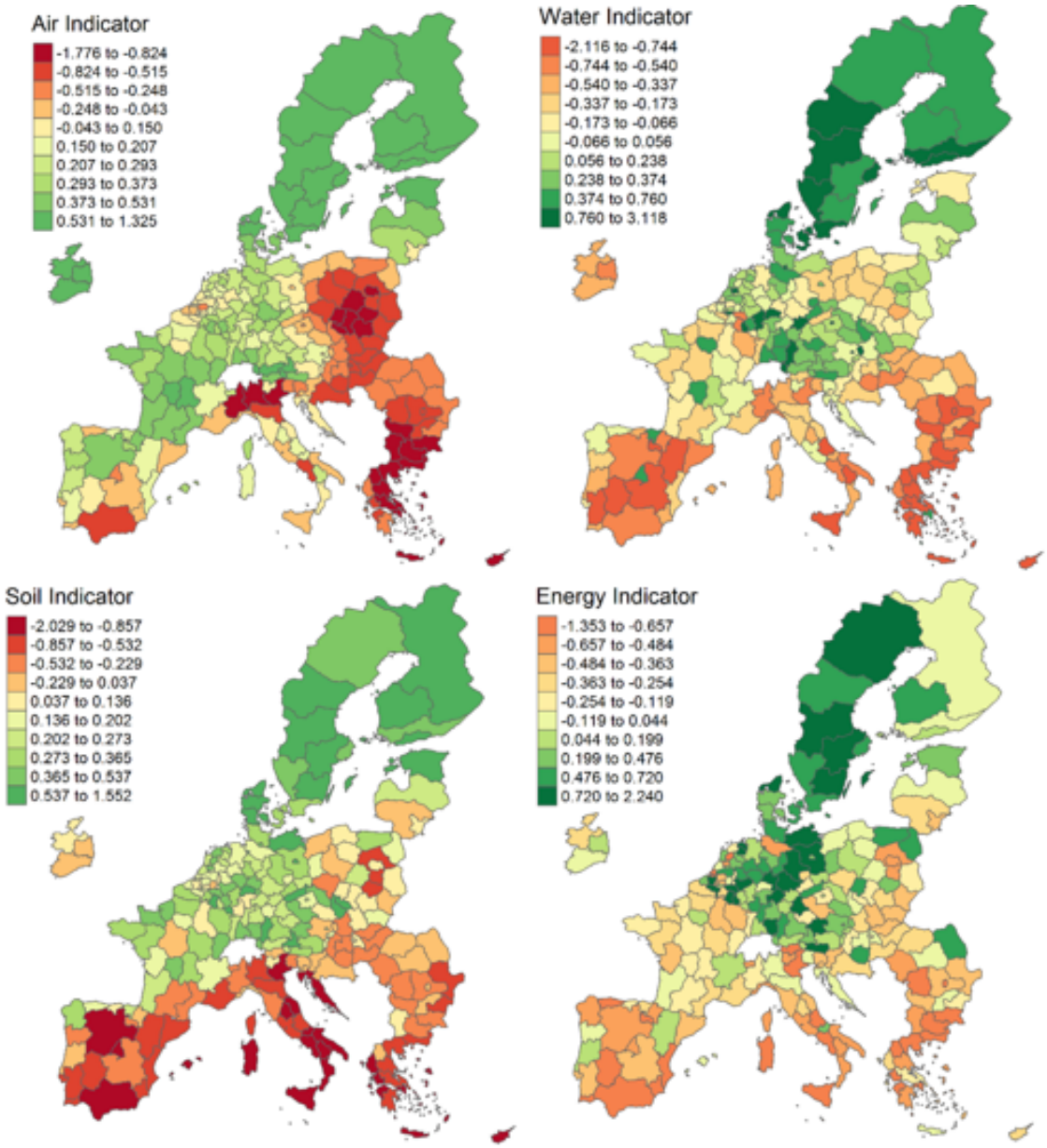
We begin the empirical part by drawing a map of the level of quality of government in European regions in 2017 in Figure 5.4.1. The map leaves little doubt that, in general, Northern and Western European countries have a higher quality of government than

Southern and Eastern European countries. And the map also confirms substantial differences among regions within many countries. To give an example of the nuances that would be missed in a national-level approach, consider the case of Italy. At the national level, according to EQI, Italy has a lower quality of government than any other country except Bulgaria, Croatia, Greece, and Romania. At the regional level, however, the Southern Italian region of Calabria has the second lowest level of subnational institutional quality in Europe. In contrast, the Northern Italian autonomous provinces of Trento and Bolzano have higher subnational institutional quality than relatively successful regions such as Catalonia (Spain) and Warsaw (Poland). Studies that do not dig deeper into within-country differences neglect these 'details'.

Figure 5.4.2 shows that environmental well-being follows a relatively similar geographical division to the quality of government, with some variation among dimensions. Generally, citizens living in Northern and Western Europe enjoy greater environmental wellbeing than citizens living in Southern and Eastern Europe. The regions with the poorest environmental wellbeing are Attica (Greece) for air, Thessalia (Greece) for water, Sicily (Italy) for soil, and Algarve (Portugal) for energy and climate change. The regions with the best environmental wellbeing instead are Upper Norrland (Sweden) for air, Copenhagen (Denmark) for water, Salzburg (Austria) for soil, and Saxony-Anhalt (Germany) for energy and climate change. Nonetheless, compared to quality of government, we find even larger within-country differences and exceptions to the general pattern.

As exemplified by the box plots in Figures C5-C8 (Appendix 5.C), there is substantial variation in environmental well-being within many countries. Outlier regions such as Berlin (Germany), Lombardy (Italy), and Moravia-Silesia (Czech Republic) have significantly poorer air quality than other regions in their respective countries. The same applies for regions such as Liège (Belgium), Northwest Bohemia (Czech Republic), and Lower Austria for water quality, Corsica (France), Ionian Islands (Greece), and Swietokrzyskie (Poland) for soil quality, and Algarve (Portugal) for energy and climate change. Some outlier regions instead perform much better than expected. Regions such as Stockholm (Sweden), Brussels (Belgium), and Prague (Czech Republic) have much higher water quality, regions such as Bremen (Germany), Budapest (Hungary), and Åland (Finland) have much higher soil quality,

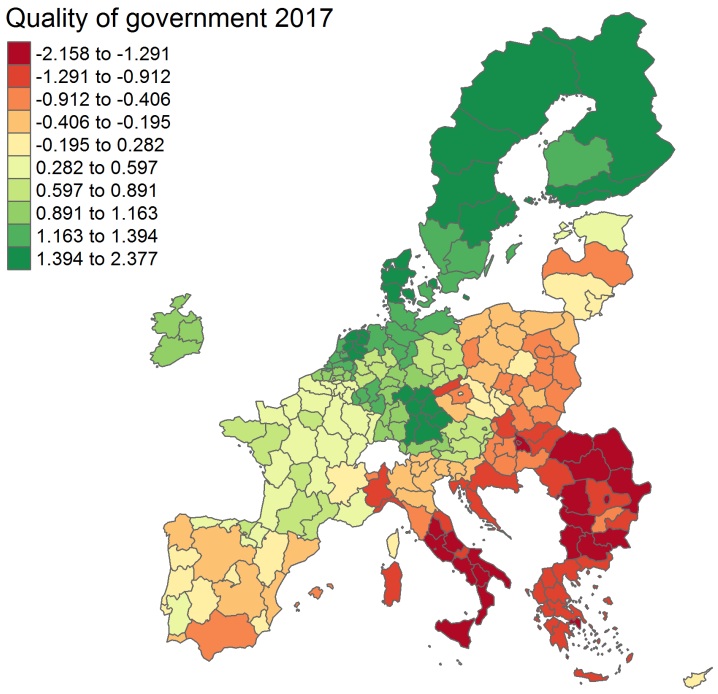
Figure 5.4.1: Map of environmental wellbeing in European regions, by decile



and regions such as Bratislava (Slovakia), East-Central Sweden, and Saxony-Anhalt have much more wellbeing in terms of energy and climate change compared to other regions in their respective countries.

By computing the standard deviation (sd) of regional environmental wellbeing scores in a given country, we can classify countries according to the amount of within-country variation. As for air quality, the largest within-country variation occurs in Croatia (sd = 0.52), Italy (sd = 0.51), and Greece (sd = 0.46). As for water quality, the largest within-country variation occurs in Belgium (sd = 1.03), Denmark (sd = 1.02), and Greece (sd = 0.62). As for soil quality, the largest within-country variation occurs in Croatia (sd = 0.60), Spain (sd = 0.52), and Italy (sd = 0.50).

Figure 5.4.2: Map of quality of government in European regions, by decile



As for energy and climate change, the largest within-country variation occurs in Belgium (sd = 0.79), Netherlands (sd = 0.63), and Germany (sd = 0.61). These results show that while environmental wellbeing tends to be higher in Northern and Western Europe, within-country differences do not follow any clear geographical pattern. A complete picture of the relationship between quality of government and environmental wellbeing requires then taking into account subnational variation.

Table 5.4.1 reports the estimated factor loadings of the spatial Bayesian latent factor models. Factor loadings with negative signs imply an inverse association between the elementary indicators and the latent dimension of environmental well-being. Conversely, factor loadings with positive signs imply a positive association between the elementary indicators and the latent dimension of environmental wellbeing. When the factor loading distribution is highly centered around zero, we consider the associated indicator insignificant for improving well-being. The posterior means of all our factor loadings have the expected signs.

Some elementary indicators are more strongly related to their respective latent dimensions of environmental well-being than others. The indicators of PM10 and PM2.5-based air pollution have the strongest relationship with the dimension of air. Ozone and NO2-based air pollution and urban vegetation's capacity to remove NO2 are moderately related to air, whereas urban exposure to PM10 is weakly

related to air. Water productivity is relatively strongly related to the dimension of water, whereas the quality of drinking water, sewage treatment, and freshwater consumption are moderately related to water. The capacity of ecosystems to avoid soil erosion has the strongest covariance with the dimension of soil. Severe soil erosion by water, organic farming, and artificial surfaces inside protected areas have a weaker relationship with soil. Energy recovery capacity represents energy and climate change better than climate change vulnerability.

Table 5.4.1: Elementary indicators of wellbeing: posterior mean and 95% credibility interval of factor loadings for each environmental dimension

Elementary indicator	**Air**	**Water**	**Soil**	**Energy**
NO2 Removal capacity by urban vegetation	0.739 (0.657, 0.979)			
Urban population exposed to PM10	-0.262 (-0.341, -0.039)			
Air pollution – PM2.5	-1.608 (-1.661, -1.450)			
Air pollution – PM10	-1.674 (-1.725, -1.513)			
Air pollution – Ozone	-0.626 (-0.699, -0.410)			
Air pollution – NO2	-0.729 (-0.801, -0.510)			
Water productivity or use efficiency		0.756 (0.678, 0.985)		
Drinking water quality		0.491 (0.424, 0.686)		
Sewage treatment		0.397 (0.329, 0.591)		
Freshwater consumption per capita		-0.430 (-0.498, -0.231)		
Capacity of ecosystems to avoid soil erosion			0.558 (0.476, 0.875)	
Severe soil erosion by water			-0.340 (-0.443, -0.035)	
Artificial surfaces inside N2000 in km²			-0.220 (-0.280, 0.049)	
Organic farming			0.318 (0.239, 0.547)	
Energy recovery (R1) capacity per capita				0.619 (0.520, 0.949)
Potential vulnerability to climate change				-0.436 (-0.531, -0.149)

Note: Each row corresponds to one of the elementary indicators used in the composite indicator's construction, for each environmental pillar. Factor loadings represent the posterior mean of each λ in our statistical model. The numbers in square brackets are the left and right bounds of the 95% credibility intervals. In a Bayesian framework, these values represent the boundary within which rely 95% of the λ posterior probability.

With our new composite indicators of the core dimensions of environmental wellbeing, we are now ready to assess the institutions-environment nexus through regression analysis. Table 5.4.2 summarises the results of our main regressions, where quality of government is measured in 2017. In the baseline models, we do not include any controls in the regression equation. In the intermediate models, we control for potential confounders, including GDP/capita, population density, total area, employment in agriculture and manufacturing, unemployment rate, and research and development expenditure. In the full models, to exclude that the associations are driven by other aspects of environmental wellbeing, we also control different dimensions of environmental wellbeing.

The baseline models (1-4) without controls show that quality of government is strongly correlated with each of our four dimensions of environmental well-being. The positive sign of the slope coefficients indicates that a higher level of

quality of government goes together with higher environmental well-being. In each environmental dimension the result is statistically significant at the 1% level. Nevertheless, to get more robust evidence on the link between quality of government and environmental well-being we must control for potential confounding factors. The results remain substantially unaltered in the intermediate models (5-8), where we include the previously discussed set of controls. Quality of government is a positive and statistically significant predictor of air ($\beta = 0.213$), water ($\beta = 0.183$), soil ($\beta = 0.237$), and energy ($\beta = 0.212$) at the 1% level. None of the other independent variables seems to be an equally important determinant of all our four dimensions of environmental well-being.

In the full models (9-12), we also control for the main dimensions of environmental well-being. At least in theory, these different dimensions are likely to be interrelated. The regression results confirm our theoretical expectations in part. The predictive power of quality of government on air ($\beta = 0.137$), soil ($\beta = 0.165$), and energy ($\beta = 0.136$) decreases compared to the previous sets of models but the slope coefficients remain statistically significant at least at the 1% level. The relationship between quality of government and water instead becomes considerably weaker and the slope coefficient becomes non-significant ($\beta = 0.041$). These results show that quality of government has, in general, a positive impact on the quality of air, soil, and energy and climate change. Although these full models do not suggest a clear association between quality of government and quality of water, the link between the two could work in a more indirect way through the other dimensions of environmental wellbeing —especially soil quality, which is strongly related to water quality according to the regression estimates.

The coefficients discussed above reflect the total effect of government quality on environmental well-being, a combination of direct and indirect effects. Table 5.4.3 reports disaggregated estimates of these two types of effects for the models discussed above.

In the baseline models, both the direct and indirect associations between quality of government and environmental wellbeing are positive and statistically significant, at least at the 1% level across models. The direct impact is consistently stronger than the indirect impact. The pattern emerges from the intermediate models with controls, except that the indirect relationship between quality of government and air

Table 5.4.2: Environmental wellbeing and quality of government (2017): main regression results

Dependent variable:	Air	Water	Soil	Energy
Baseline models	(1)	(2)	(3)	(4)
Quality of government	0.219***	0.361***	0.221***	0.270***
	(0.025)	(0.040)	(0.025)	(0.035)
N	233	233	233	233
Wald test (df = 1)	167.557***	5.100**	126.309***	8.148***
LR test (df = 1)	93.169***	6.509***	73.929***	7.457***
Intermediate models	(5)	(6)	(7)	(8)
Quality of government	0.213***	0.183***	0.237***	0.211***
	(0.036)	(0.050)	(0.049)	(0.051)
Ln(GDP/capita)	0.076	0.195	-0.110	-0.303*
	(0.114)	(0.161)	(0.165)	(0.163)
Population density	-0.0001***	0.0003***	0.0001***	0.0000
	(0.00004)	(0.0001)	(0.0000)	(0.0001)
Total area	0.0000	0.0000	0.0000	0.0000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.007	-0.006	-0.007	-0.004
	(0.005)	(0.007)	(0.007)	(0.007)
Employment in manufacturing	0.002	0.000	0.002	0.003
	(0.004)	(0.005)	(0.004)	(0.005)
Unemployment (15-75)	-0.002	-0.026***	-0.019***	-0.005
	(0.006)	(0.008)	(0.006)	(0.009)
R&D expenditure	0.0000	0.0002***	0.0000	0.0002***
	(0.00005)	(0.00008)	(0.00005)	(0.00008)
N	221	221	221	221
Wald test (df = 1)	85.832***	1.951	93.202***	6.949***
LR test (df = 1)	53.166***	2.383	54.503***	6.040**
Full models	(9)	(10)	(11)	(12)
Quality of government	0.137***	0.041	0.165***	0.136***
	(0.042)	(0.059)	(0.051)	(0.062)
Air		0.045		0.045
		(0.080)		(0.084)
Water	0.045		0.131***	
	(0.048)		(0.043)	
Soil		0.288***		0.122
		(0.087)		(0.094)
Energy	0.032	0.068	0.039	
	(0.045)	(0.063)	(0.041)	
Ln(GDP/capita)	0.115	0.226	-0.146	-0.308*
	(0.114)	(0.159)	(0.146)	(0.162)
Population density	0.00004*	0.0003***	0.0001**	0.0001
	(0.00004)	(0.0001)	(0.00004)	(0.0001)
Total area	0.0000	0.0000	0.0000	0.0000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.007	-0.006	-0.007	-0.004
	(0.005)	(0.007)	(0.007)	(0.007)
Employment in manufacturing	0.002	0.000	0.002	0.003
	(0.004)	(0.005)	(0.004)	(0.005)
Unemployment (15-75)	-0.002	-0.026***	-0.019***	-0.005
	(0.006)	(0.008)	(0.006)	(0.009)
R&D expenditure	0.0000	0.0002***	0.0000	0.0002***
	(0.00005)	(0.00008)	(0.00005)	(0.00008)
N	221	221	221	221
Wald test (df = 1)	85.832***	1.951	93.202***	6.949***
LR test (df = 1)	53.166***	2.383	54.503***	6.040**

Note: Robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Quality of government refers to the year 2017. LR test indicates the significance of the spatial autoregressive parameter.

quality becomes non-significant. Finally, results from the full models, where we also include the other dimensions of environmental well-being, show that both the direct and indirect links are positive and significant for water and soil quality. Energy and climate change are only directly but not indirectly associated with the quality of government. Conversely, in the full models, there seems to be no statistically significant relationship between water quality and quality of government.

We test the robustness of our results with a battery of alternative regression models. First, instead of measuring quality of government in 2017, we use the 2013 measurement (Table 5.4.4) and 2010 one (Table 5.4.5) to investigate if longer time lags between our main independent and dependent variables affect the institutions-

Table 5.4.3: Environmental wellbeing and quality of government (2017): direct and indirect effects

	Air		Water		Soil		Energy	
	Direct	Indirect	Direct	Indirect	Direct	Indirect	Direct	Indirect
Baseline models								
Quality of government	0.257***	0.161***	0.363***	0.043**	0.252***	0.141***	0.273***	0.041***
	(0.026)	(0.020)	(0.038)	(0.020)	(0.033)	(0.019)	(0.034)	(0.014)
Intermediate models								
Quality of government	0.235***	0.114***	0.184***	0.012	0.261***	0.126***	0.214***	0.030**
	(0.044)	(0.023)	(0.045)	(0.009)	(0.035)	(0.020)	(0.058)	(0.014)
Full models								
Quality of government	0.149***	0.077***	0.041	0.001	0.183***	0.081***	0.137***	0.016
	(0.042)	(0.023)	(0.057)	(0.004)	(0.037)	(0.018)	(0.058)	(0.009)

Note: Direct and indirect effects are averaged over all N regions/observations. The direct effect provides a summary measure of the impact of quality of government in region i. It considers feedback effects that arise from the change in the i region's quality of government on the environmental quality of neighbouring regions in the system of spatially dependent regions. The indirect effect measures the impact of an increase in quality of government in all other regions on the environmental quality of a given region.

environment nexus.

We find that quality of government in both 2013 and 2010 is a strong and statistically significant predictor of each of our four dimensions of environmental wellbeing in the baseline and intermediate models. This mirrors the results of the previous set of regressions. The full models instead show interesting differences among our four dimensions of environmental wellbeing. When we add our environmental controls into the models, the effect of quality of government on air remains more or less the same over time, regardless of whether quality of government is measured in 2017, 2013, or 2010. The effect of quality of government on energy and climate change decreases considerably over time. It does not retain its statistical significance in the full models when the quality of government is measured in 2013 or 2010. The relationship between quality of government and water quality remains non-significant across full models, like in our previous regressions.

Then, instead of disaggregating environmental wellbeing into its underlying dimensions, we create a composite index of 'overall' environmental wellbeing, synthesising all our 16 elementary indicators. We use this overall index in our spatial lag models as a dependent variable and test the link between overall environmental quality well-being and institutions (Table 5.D.1 in Appendix 5.D). Regardless of confounding factors or time lags, well-functioning, effective institutions correlate positively with overall environmental well-being at the highest statistical signifi-

cance level. The year of measurement of quality of government essentially does not affect the magnitude of the association, indicating that environmental well-being as a whole can be advanced by improving regional institutional quality. The estimates in Table 5.D.2 (Appendix 5.D) show that the effect of the quality of government on overall environmental well-being occurs both in direct and indirect ways. Still, the direct effect is stronger than the indirect effect.

Table 5.4.4: Environmental wellbeing and quality of government (2013): regression results

	Air	Water	Soil	Energy
Baseline models	(1)	(2)	(3)	(4)
Quality of government	0.236***	0.334***	0.188***	0.223***
	(0.025)	(0.040)	(0.025)	(0.036)
N	233	233	233	233
Wald test (df = 1)	166.028***	6.887***	165.421***	14.080***
LR test (df = 1)	94.957**	8.417***	88.566***	12.466***
Intermediate models	(5)	(6)	(7)	(8)
Quality of government	0.230***	0.154***	0.192***	0.144***
	(0.035)	(0.049)	(0.033)	(0.050)
Ln(GDP/capita)	0.048	0.212	-0.083	-0.325*
	(0.113)	(0.163)	(0.108)	(0.169)
Population density	-0.0001***	0.0003***	0.0001***	0.00003
	(0.00004)	(0.0001)	(0.00004)	(0.00003)
Total area	0.00000***	0.00000	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.006	-0.006	0.004	-0.004
	(0.005)	(0.007)	(0.004)	(0.007)
Employment in manufacturing	-0.006	0.001	-0.007*	0.005
	(0.004)	(0.005)	(0.004)	(0.006)
Unemployment (15-75)	0.002	-0.030***	-0.023***	-0.036***
	(0.006)	(0.008)	(0.005)	(0.009)
R&D expenditure	-0.00001	0.0002***	0.00005	0.0002***
	(0.00005)	(0.00001)	(0.00005)	(0.00002)
N	221	221	221	221
Wald test (df = 1)	87.570***	2.311	112.496***	10.061***
LR test (df = 1)	55.014***	2.761**	62.399***	8.600***
Full models	(9)	(10)	(11)	(12)
Quality of government	0.170***	0.028	0.117***	0.047
	(0.038)	(0.056)	(0.035)	(0.059)
Air		0.113		0.147***
		(0.081)		(0.053)
Water	0.043		0.141***	
	(0.047)		(0.043)	
Soil		0.165***		0.181***
		(0.060)		(0.091)
Energy	0.041	0.074	0.062	
	(0.044)	(0.063)	(0.041)	
Ln(GDP/capita)	0.083	0.234	-0.120	-0.259
	(0.112)	(0.158)	(0.108)	(0.168)
Population density	-0.0001***	0.0003***	0.0001**	0.0001
	(0.00004)	(0.0001)	(0.00004)	(0.00004)
Total area	0.00000***	0.00000	-0.00000	-0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.007	-0.007	0.006	0.001
	(0.005)	(0.006)	(0.004)	(0.007)
Employment in manufacturing	-0.007*	0.003	-0.003	-0.005
	(0.004)	(0.005)	(0.003)	(0.006)
Unemployment (15-75)	0.008	-0.016*	-0.017***	-0.027***
	(0.006)	(0.009)	(0.006)	(0.009)
R&D expenditure	0.00003	0.0002***	-0.00005	0.0002***
	(0.00005)	(0.00008)	(0.00005)	(0.00008)
N	221	221	221	221
Wald test (df = 1)	77.557***	2.262	92.961***	5.547***
LR test (df = 1)	50.899***	0.314	55.925***	4.717**

Note: Robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Quality of government refers to year 2013. LR test indicates the significance of the spatial autoregressive parameter ρ.

Table 5.4.5: Environmental wellbeing and quality of government (2010): regression results

Dependent variable:	Air	Water	Soil	Energy
Baseline models	(1)	(2)	(3)	(4)
Quality of government	0.241***	0.321***	0.174***	0.227***
	(0.025)	(0.041)	(0.025)	(0.036)
N	233	233	233	233
Wald test (df = 1)	158.904***	8.536***	189.342***	14.597***
LR test (df = 1)	90.653***	10.439***	98.244***	13.058***
Intermediate models	(5)	(6)	(7)	(8)
Quality of government	0.239***	0.171***	0.207***	0.275***
	(0.035)	(0.049)	(0.033)	(0.050)
Ln(GDP/capita)	0.009	0.175	-0.125	-0.350*
	(0.113)	(0.164)	(0.108)	(0.169)
Population density	-0.0001*	0.0003***	0.0001***	0.00004
	(0.00004)	(0.0001)	(0.00004)	(0.00004)
Total area	0.00000***	0.00000	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.006	-0.006	0.004	0.001
	(0.005)	(0.007)	(0.004)	(0.007)
Employment in manufacturing	-0.006	0.001	-0.007*	0.004
	(0.004)	(0.005)	(0.004)	(0.006)
Unemployment (15-75)	-0.006	-0.032***	-0.026***	-0.038***
	(0.006)	(0.008)	(0.005)	(0.008)
R&D expenditure	-0.00001	0.0002***	0.00005	0.0002***
	(0.00005)	(0.00001)	(0.00005)	(0.00002)
N	221	221	221	221
Wald test (df = 1)	90.293***	2.298	120.949***	9.592***
LR test (df = 1)	55.751***	2.744**	66.878***	8.529***
Full models	(9)	(10)	(11)	(12)
Quality of government	0.178***	0.044	0.132***	0.088
	(0.039)	(0.057)	(0.035)	(0.059)
Air		0.105		0.183***
		(0.082)		(0.053)
Water	0.038		0.136***	
	(0.047)		(0.043)	
Soil		0.163***		0.162*
		(0.060)		(0.091)
Energy	0.032	0.074	0.054	
	(0.044)	(0.063)	(0.041)	
Ln(GDP/capita)	0.052	0.218	-0.145	-0.294*
	(0.112)	(0.160)	(0.108)	(0.162)
Population density	-0.0001***	0.0003***	0.0001**	0.00004
	(0.00004)	(0.0001)	(0.00004)	(0.00004)
Total area	0.00000***	0.00000	-0.00000	-0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.007	-0.007	0.006	0.001
	(0.005)	(0.006)	(0.004)	(0.007)
Employment in manufacturing	-0.007*	0.003	-0.003	-0.005
	(0.004)	(0.005)	(0.003)	(0.006)
Unemployment (15-75)	0.004	-0.017*	-0.019***	-0.028***
	(0.006)	(0.009)	(0.006)	(0.009)
R&D expenditure	0.00003	0.0002***	-0.00005	0.0002***
	(0.00005)	(0.00008)	(0.00005)	(0.00008)
N	221	221	221	221
Wald test (df = 1)	80.360***	2.266	98.266***	5.419***
LR test (df = 1)	51.190***	0.319	49.244***	4.193**

Robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Quality of government refers to year 2010. LR test indicates the significance of the spatial autoregressive parameter ρ.

5.5 Conclusion

The study at hand has investigated the relationship between quality of government and environmental wellbeing in Europe through a multidimensional, comparative, and regional approach. The main contributions of our study are manifold. First, we detected the presence of spatial spillovers in environmental wellbeing in European regions. Second, accounting for this spatial correlation, we constructed a set of composite indicators, capturing four main dimensions of environmental wellbeing: air quality, water quality, soil quality, and energy and climate change. Third, through a battery of spatial regression models, we showed that institutional

quality is a significant and positive predictor of environmental wellbeing. This is particularly true for the dimensions of air and soil, and to a lesser extent for the dimension of energy and climate change. The effect of quality of government on water quality instead seems to be primarily indirect, occurring through other aspects of environmental wellbeing.

The existence of spatial correlation in environmental wellbeing is after all not so surprising. Even if the extent of environmental degradation varies extensively from one region to another in many countries, neighbouring regions tend to have more similar scores than distant regions. The negative externalities of environmental hazards do not follow regional boundaries but can spread to neighbouring regions. For instance, poor air quality caused by Europe's largest coal-fired power plant in Lodz, does not only increase air pollution in Lodz but depending on the winds can affect the quality of air also in neighbouring regions. Our study thus provides strong evidence of the existence of spatial correlation in data on environmental wellbeing. This means that scholars studying the environment should seriously consider the spatial characteristics of their data. Otherwise, their results are likely to be biased.

By developing a set of novel composite indices of environmental wellbeing, we have found that in general Northern and Western European regions have better environmental wellbeing than Southern and Eastern European regions. Yet, our subnational approach has shown that there are many exceptions too —environmental wellbeing varies significantly not only across countries but also within countries and depends up to a certain extent on the dimension of wellbeing. Just to give a couple of examples, in Bolzano (Italy), air quality is higher than in Copenhagen (Denmark), but in Lombardia (Italy) air quality is worse than in almost any other European region. Water quality in Copenhagen however is better than in any other European region. In Lower Austria instead soil quality is poorer than in most Polish regions but Salzburg (Austria) is one of the regions with the best soil quality. These nuances can be captured only with a subnational and multidimensional approach to environmental wellbeing.

Finally, through a series of spatial regression models and robustness tests, we find strong evidence of a positive overall relationship between environmental wellbeing and quality of government. The institutions-environment nexus however is not equal across the core dimensions of environmental wellbeing. Quality of

government matters especially for air and soil quality, and to a lesser extent for energy and climate change. On the contrary, we find only weak support for the direct relationship between quality of government and water quality. The quality of water seems to be more affected by other dimensions of environmental wellbeing rather than institutions. Improvements in quality of government however seem to be positively associated water quality via other dimensions of environmental wellbeing —in particular soil quality. By considering these differences among dimensions of environmental wellbeing, our study shows that equating environmental wellbeing simplistically with air pollution is misleading. Researchers and policymakers must take into account other aspects of environmental wellbeing too.

Our findings do not provide any support for an inverse relationship between institutional quality and the environment. Hence, given that quality of government matters for environmental wellbeing, our study shows clearly that by strengthening regional institutional quality, policymakers can significantly reduce the burden caused by environmental problems and improve the living conditions of their citizens. While the study at hand has focused on European regions, there seems to be no reason to believe that the hypothesised association would not apply to other contexts as well. We call on future research to verify whether institutions matter for environmental wellbeing also in other regions of the world, once more abundant subnational cross-country data on environmental wellbeing and quality of government becomes available.

References

Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Publishers.

Aklin, M., Bayer, P., Harish, S. P., & Urpelainen, J. (2014). Who blames corruption for the poor enforcement of environmental laws? survey evidence from brazil. *Environmental Economics and Policy Studies*, *16*(3), 241–262.

Azimi, M. N., Rahman, M. M., & Nghiem, S. (2023). Linking governance with environmental quality: A global perspective. *Scientific Reports*, *13*(1), 1–18.

Barbero, J., Mandras, G., Rodríguez-Crespo, E., & Rodríguez-Pose, A. (2021). Quality of government and regional trade: Evidence from european union regions. *Regional Studies*, *55*(7), 1240–1251.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Charron, N., Dijkstra, L., & Lapuente, V. (2015). Mapping the regional divide in europe: A measure for assessing quality of government in 206 european regions. *Social Indicators Research*, *122*(2), 315–346.

Charron, N., Lapuente, V., & Annoni, P. (2019). Measuring quality of government in eu regions across space and time. *Papers in Regional Science*, *98*(5), 1925–1953.

Chong, A., & Calderon, C. (2000). Institutional quality and poverty measures in a cross-section of countries. *Economics of Governance*, *1*(2), 123–135.

Ciommi, M., Gigliarano, C., Chelli, F. M., & Gallegati, M. (2022). It is the total that does [not] make the sum: Nature, economy and society in the equitable and sustainable well-being of the italian provinces. *Social Indicators Research*, *161*(2–3), 491–522.

Cole, M. A. (2007). Corruption, income and the environment: An empirical analysis. *Ecological Economics*, *62*(3–4), 637–647.

Cornell, A., Knutsen, C. H., & Teorell, J. (2020). Bureaucracy and growth. *Comparative Political Studies*, *53*(14), 2246–2282.

Cressie, N., & Chan, N. H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, *84*(406), 393–401.

Davis, W., Gordan, A., & Tchernis, R. (2021). Measuring the spatial distribution of health rankings in the United States. *Health Economics*, *30*(11), 2921–2936.

DESA, U. (2019). *United nations world public sector report 2019. sustainable development goal 16: Focus on public institutions*. United Nations.

Di Cataldo, M., & Rodríguez-Pose, A. (2017). What drives employment growth and social inclusion in the regions of the european union? *Regional Studies*, *51*(12), 1840–1859.

Evans, P. (1995). *Embedded autonomy: States and industrial transformation*. Princeton University Press.

Evans, P., & Rauch, J. E. (1999). Bureaucracy and growth: A cross-national analysis of the effects of 'weberian' state structures on economic growth. *American Sociological Review*, *64*(5), 748–765.

Ezcurra, R., & Rios, V. (2019). Quality of government and regional resilience in the european union. evidence from the great recession. *Papers in Regional Science*, *98*(3), 1267–1290.

Giovannini, E. (2015). 'beyond gdp' ten years after the first oecd world forum where do we stand? *Rivista Internazionale di Scienze Sociali*, *1*, 3–15.

Golgher, A. B., & Voss, P. R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography*, *4*, 175–205.

Halkos, G. E., Sundström, A., & Tzeremes, N. G. (2015). Regional environmental performance and governance quality: A nonparametric analysis. *Environmental Economics and Policy Studies*, *17*(4), 621–644.

Helliwell, J. F., & Huang, H. (2008). How's your government? international evidence linking good government and well-being. *British Journal of Political Science*, *38*(4), 595–619.

Hogan, J. W., & Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, *99*(466), 314–324.

Holmberg, S., & Rothstein, B. (2011). Dying of corruption. *Health Economics, Policy, and Law*, *6*(4), 529–547.

Holmberg, S., Rothstein, B., & Nasiritousi, N. (2009). Quality of government: What you get. *Annual Review of Political Science*, *12*(1), 135–161.

Huang, C. J. (2016). Is corruption bad for economic growth? evidence from asia-pacific countries. *North American Journal of Economics and Finance*, *35*, 247–256.

Iammarino, S., Rodríguez-Pose, A., & Storper, M. (2019). Regional inequality in europe: Evidence, theory and policy implications. *Journal of Economic Geography*, *19*(2), 273–298.

Iddawela, Y., Lee, N., & Rodríguez-Pose, A. (2021). Quality of sub-national government and regional development in africa. *Journal of Development Studies*, *57*(8), 1282–1302.

ISTAT. (2021). BES 2021. Il benessere equo e sostenibile in Italia. Rome.

Kraay, A. (2004). When is growth pro-poor? cross-country evidence. *IMF Working Papers*, *WP/04/47*.

LeSage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman; Hall/CRC.

Lombardo, R., & Ricotta, F. (2021). Individual trust and quality of regional government. *Journal of Institutional Economics*, 1–22.

MacAskill, W. (2022). The beginning of history: Surviving the era of catastrophic risk. *Foreign Affairs*, *101*(5), 10–24.

Mazziotta, M., Gigliarano, C., & Rimoldi, S. (2021). Well-being and territory: Methods and strategies. *Social Indicators Research*.

Meyer, A. L., Van Kooten, G. C., & Wang, S. (2003). Institutional, social and economic roots of deforestation: A cross-country comparison. *International Forestry Review*, *5*(1), 29–37.

Michalos, A. C. (1997). Combining social, economic and environmental indicators to measure sustainable human well-being. *Social Indicators Research*, *40*, 221–258.

Moran, P. (1950). A test for the serial independence of residuals. *Biometrika*, *37*(1–2), 178–181.

Nistotskaya, M., Charron, N., & Lapuente, V. (2015). The wealth of regions: Quality of government and smes in 172 european regions. *Environment and Planning C: Government and Policy*, *33*(5), 1125–1155.

OECD. (2020). *How's life? 2020: Measuring well-being, oecd publishing, paris, 2020* (tech. rep.).

Panaro, A. V., & Vaccaro, A. (2022). Income inequality in authoritarian regimes: The role of political institutions and state capacity. *Italian Political Science Review*.

Parente, F. (2019). Inequality and social capital in the eu regions: A multidimensional analysis. *Regional Studies, Regional Science*, *6*(1), 1–24.

Peiró-Palomino, J., Picazo-Tadeo, A. J., & Rios, V. (2020). Well-being in european regions: Does government quality matter? *Papers in Regional Science*, *99*(3), 555–582.

Pellegrini, L., & Gerlagh, R. (2006). Corruption, democracy, and environmental policy: An empirical contribution to the debate. *Environment & Development*, *15*(3), 332–354.

Povitkina, M. (2018). The limits of democracy in tackling climate change. *Environmental Politics*, *27*(3), 411–432.

Povitkina, M., & Bolkvadze, K. (2019). Fresh pipes with dirty water: How quality of government shapes the provision of public goods in democracies. *European Journal of Political Research*, *58*(4), 1191–1212.

Rodríguez-Pose, A. (2013). Do institutions matter for regional development? *Regional Studies*, *47*(7), 1034–1047.

Rodríguez-Pose, A., & Di Cataldo, M. (2015). Quality of government and innovative performance in the regions of europe. *Journal of Economic Geography*, *15*(4), 673–706.

Rothstein, B. (2011). *The quality of government: Corruption, social trust, and inequality in international perspective*. University of Chicago Press.

Rothstein, B., & Teorell, J. (2008). What is quality of government? a theory of impartial government institutions. *Governance*, *21*(2), 165–190.

Sachs, J. D. (2015). *The age of sustainable development*. Columbia University Press.

UN. (2012). *Resolution adopted by the general assembly on 27 july 2012: 66/288 the future we want*. United Nations.

Van de Walle, S., & Migchelbrink, K. (2020). Institutional quality, corruption, and impartiality: The role of process and outcome for citizen trust in public administration in 173 european regions. *Journal of Economic Policy Reform*.

Appendix

5.A Environmental elementary Indicators

Table 5.A.1: Description and sources of elementary indicators in the energy and climate change dimension

Elementary indicator	Year	Polarity	Description	Dim.	Source
Energy recovery (R1) capacity per capita	2018	+	On the basis of the treatment operations defined in Directive 2008/98/EC a distinction is made in treatment types: Recovery - Recycling and backfilling (excluding energy recovery) (RCVRB): operations R2 to R11; Energy recovery (RCVE): Operation R1	Energy	EEA
Potential vulnerability to climate change	2071-2100	–	Potential regional vulnerability to climate change (combination of regional potential impacts and regional capacity to adapt to climate change)	Energy	ESPON

Table 5.A.2: Description and sources of elementary indicators in the air dimension

Elementary indicator	Year	Polarity	Description	Dim.	Source
NO2 Removal capacity by urban vegetation	2020	+	Removal capacity is calculated as the product of dry deposition velocity and pollutant concentration, derived on the context of the LUISA modelling platform.	Air	JRC
Urban population exposed to PM10	2020	–	The EU urban population exposed to PM10 concentrations exceeding the daily limit value on more than 35 days in a year measures the percentage of population in urban areas exposed to PM10 concentrations exceeding the daily limit value (50 µg/m³) established by the Air Quality Directive (2008/50/EC) on more than 35 days in a calendar year.	Air	JRC
Air pollution - PM2.5	2016	–	Population weighted average of a 10 by 10 km of air concentration (µg/m³) of particle matter of size 2.5 micrometers (small particles) interpolated on a grid created by the EEA. Capped to 25 µg/m³ = limit yearly value of the EU Ambient Air Quality Directive.	Air	ESPON
Air pollution - PM10	2016	–	Population weighted average of a 10 by 10 km of air concentration (µg/m³) of particle matter of size 10 micrometers (big particles) interpolated on a grid created by the EEA.	Air	ESPON
Air pollution - Ozone	2017	–	Population weighted average of a 10 by 10 km of air Ozone O3 concentration (µg/m³) interpolated on a grid created by the EEA.	Air	ESPON
Air pollution - NO2	2017	–	Population weighted average of annual average concentration of NO2 in µg/m³, interpolated at 1 km² grid cell level and combined with GEOSTAT 1 km² grid population data, set by the EU Ambient Air Quality Directive.	Air	ESPON

Table 5.A.3: Description and sources of elementary indicators in the water and soil dimension

Elementary indicator	**Year**	**Polarity**	**Description**	**Dim.**	**Source**
Water productivity or use efficiency	2020	+	The indicator reflects productivity in terms of water use, so gives a measure of a country's water use efficiency.	Water	JRC
Drinking water quality	2020	+	Share of people who declared being satisfied with water quality, %.	Water	ESPON
Sewage treatment	2014/2016	+	Percentage of urban wastewater with more stringent treatment in collected wastewater.	Water	ESPON
Freshwater consumption per capita	2020	–	The indicator is the result of the water use model, which allocates sectorial statistical data on freshwater consumption.	Water	JRC
Capacity of ecosystems to avoid soil erosion	2020	+	The indicator measures the capacity of ecosystems to avoid soil erosion assigning values ranging from 0 to 1 at pixel level, covering the EU-28	Soil	JRC
Severe soil erosion by water	2016	–	Severe soil erosion by water is defined as the estimated share of non-artificial areas under risk of being subject to soil erosion by water (from more than 10 tonnes per hectare and year). Non-artificial areas are agricultural areas, forest and semi-natural areas (excluding beaches, dunes, sand plains, bare rock, glaciers and perpetual snow cover).	Soil	JRC
Artificial surfaces inside N2000 in km²	2018	–	Artificial surfaces include urban fabric, industrial, commercial and transport units, mine, dump, and construction sites, and artificial, non-agricultural vegetated areas.	Soil	EEA
Organic farming	2016	+	Share of total organic area in total utilised agricultural area	Soil	ESPON

5.B Elementary Indicator Spatial Autocorrelation

Table 5.B.1: Moran's test for spatial autocorrelation

Elementary Indicator	**Global Moran Index**	**p-value**
NO2 Removal capacity by urban vegetation	0.376	<0.001
Urban population exposed to PM10	0.034	0.191
Air pollution - PM2.5	0.685	<0.001
Air pollution - PM10	0.617	<0.001
Air pollution - Ozone	0.668	<0.001
Air pollution - NO2	0.826	<0.001
Water productivity or use efficiency	0.253	<0.001
Drinking water quality	0.530	<0.001
Sewage treatment	0.434	<0.001
Freshwater consumption per capita	0.486	<0.001
Capacity of ecosystems to avoid soil erosion	0.565	<0.001
Severe soil erosion by water	0.547	<0.001
Artificial surfaces inside N2000 in km²	0.293	<0.001
Organic farming	0.412	<0.001
Energy recovery (R1) capacity per capita	0.260	<0.001
Potential vulnerability to climate change	0.627	<0.001

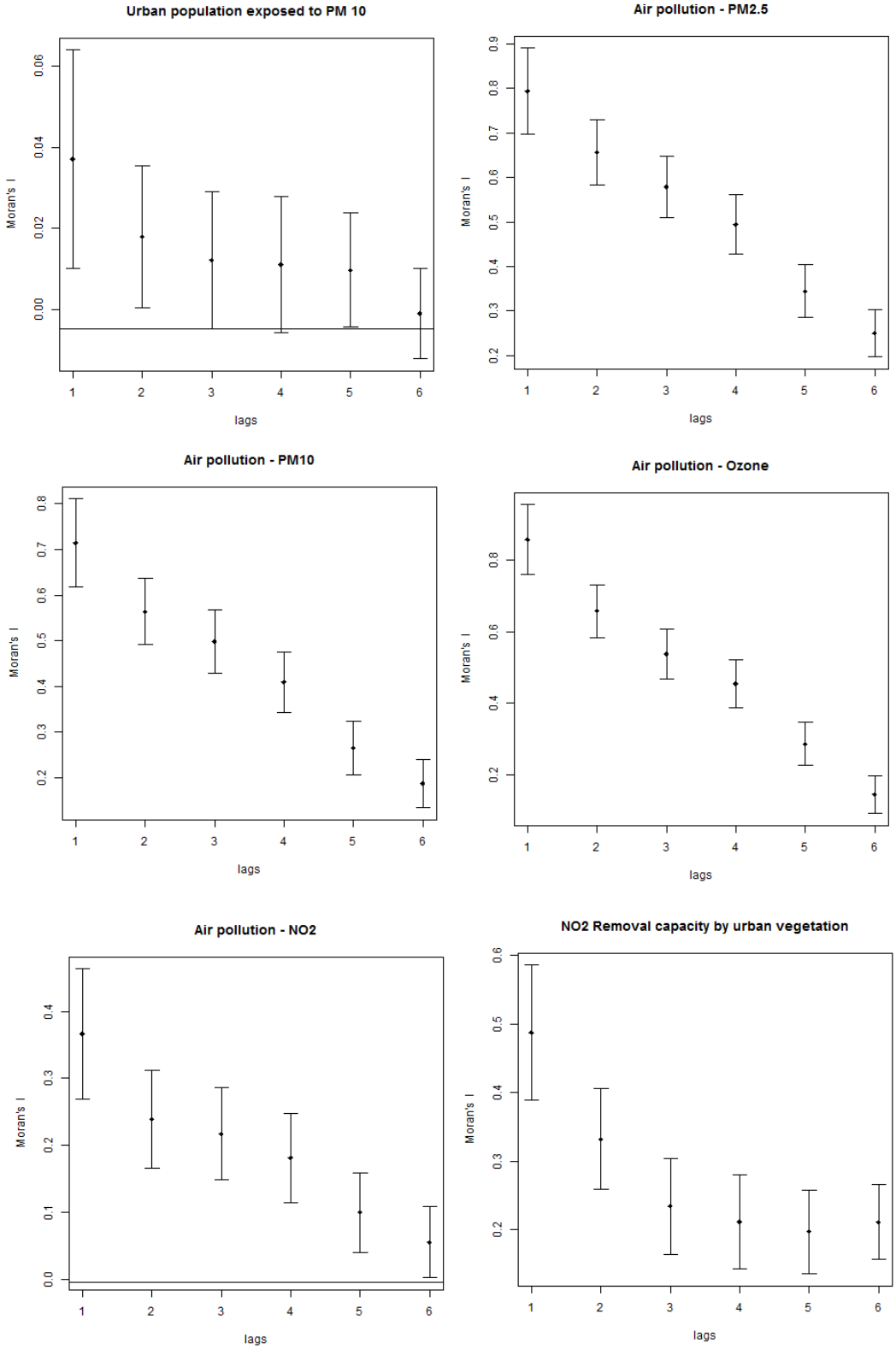
Note: Each row corresponds to one of the 16 elementary indicators used in our model. The second column reports the value of the observed Moran's I coefficient (Moran, 1950). The third column reports the p-value of the test. For all elementary indicators, except "Urban population exposed to PM10", we reject the null hypothesis of spatial randomness at 1% significance level.

Table 5.B.2: Lagrange multiplier tests

	Test			
	LMerr	LMlag	RLMerr	RLMlag
Air	1.55e-11	1.121e-14	0.1454	5.353e-05
Water	0.003587	0.004122	0.5764	0.8068
Soil	6.004e-11	2.91e-12	0.9381	0.01484
Energy	0.02365	0.009059	0.2448	0.08106
Overall	1.079e-11	8.438e-15	0.2286	8.307e-05

Note: The table reports the p-values from the different Lagrange Multiplier tests. *LMerr* is the simple test for spatial dependence in the error terms; *LMlag* is the test for omitted spatially lagged dependent variable. *RLMerr* and *RLMlag* are the robust versions of the two tests. For all the outcomes apart from the water quality indicator, the test does not reject the hypothesis of zero spatial correlation. Results indicate that the best model is the Spatial lag model.

Figure 5.B.1: Spatial correlograms for elementary indicators in the air dimension



Note: The spatial autocorrelation functions show the estimated Moran's I correlations versus spatial lag. The vertical bar corresponds to +/- twice the square root of the Moran I variance. Each lag defines higher-order neighbor sets, i.e., the first lag includes regions with contiguous boundaries, the second lag includes regions with a boundary in common with the boundaries' regions, etc. The horizontal bar indicates 0. Values greater than 0 indicate positive spatial autocorrelation or clustering in the elementary indicator; values less than 0 indicate negative spatial autocorrelation or dispersion in the elementary indicator

Figure 5.B.2: Spatial correlograms for elementary indicators in the soil dimension

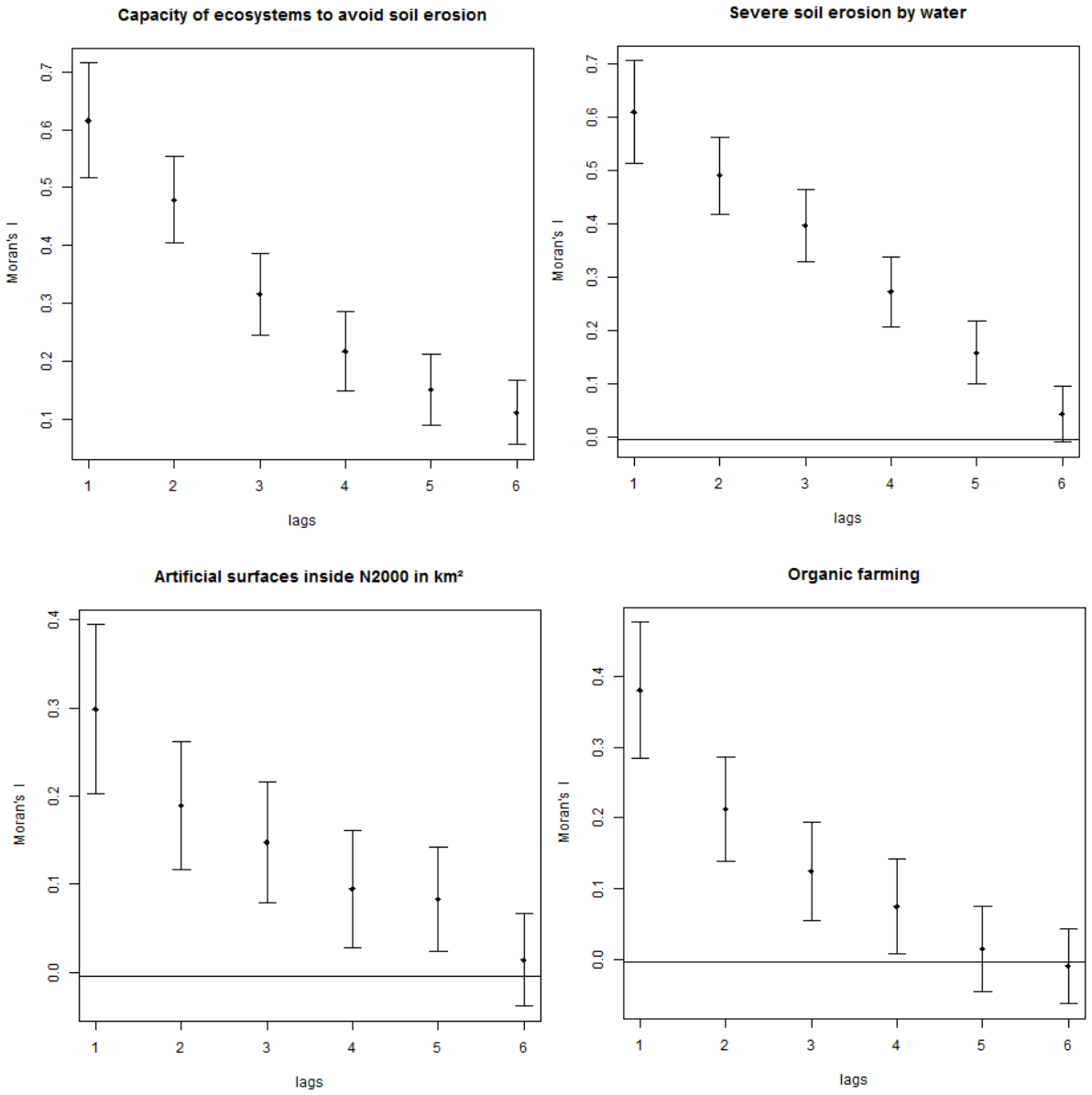


Figure 5.B.3: Spatial correlograms for elementary indicators in the water dimension

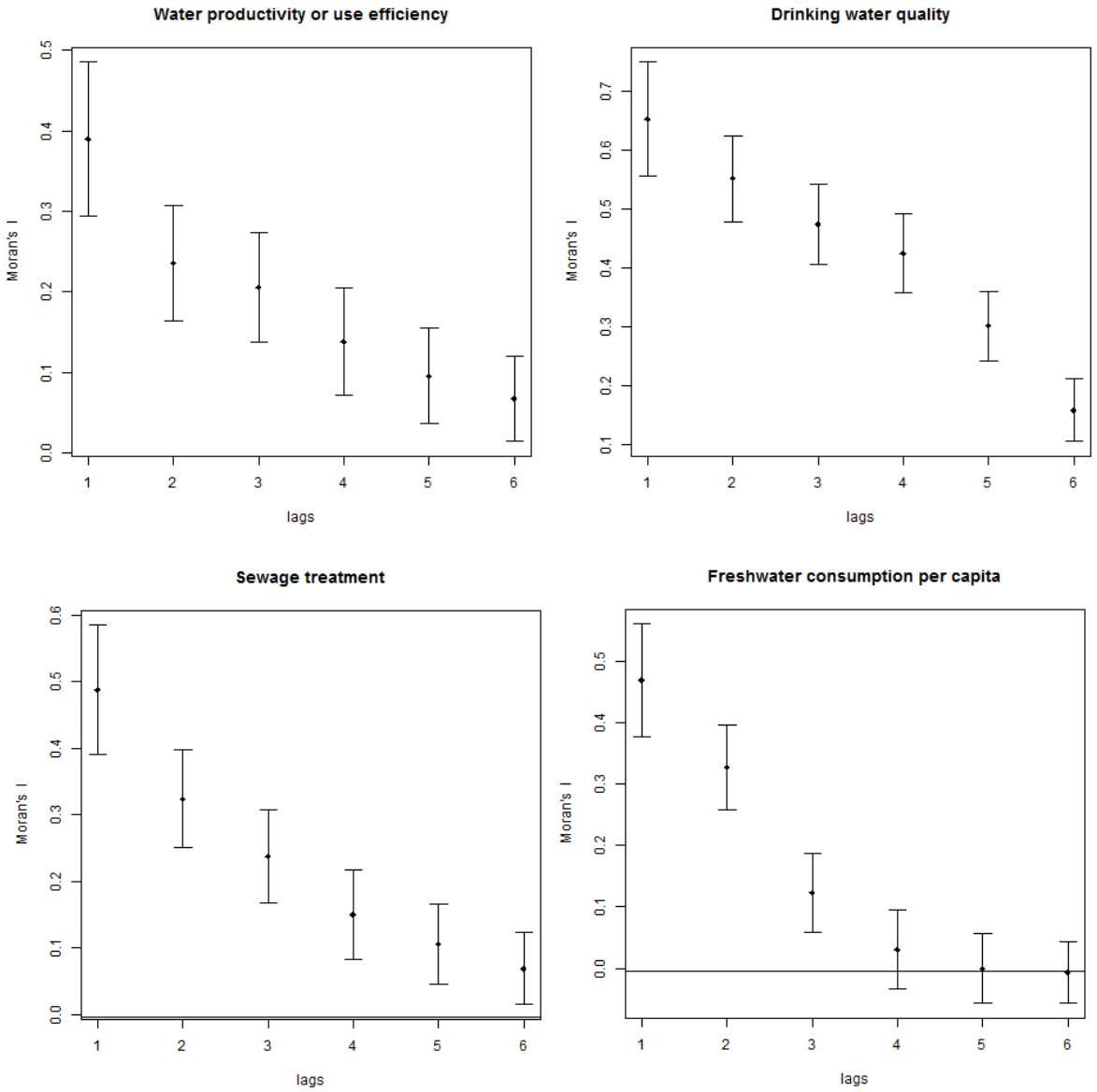
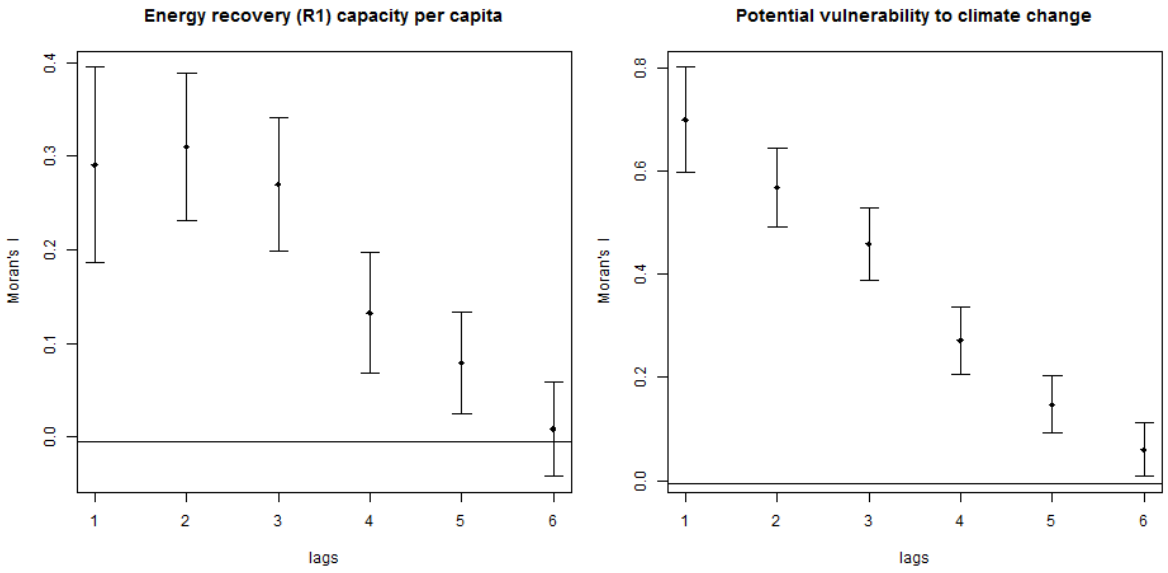


Figure 5.B.4: Spatial correlograms for elementary indicators in the energy dimension



5.C Sub-national environmental well-being

Figure 5.C.1: Boxplots of sub-national air scores by country

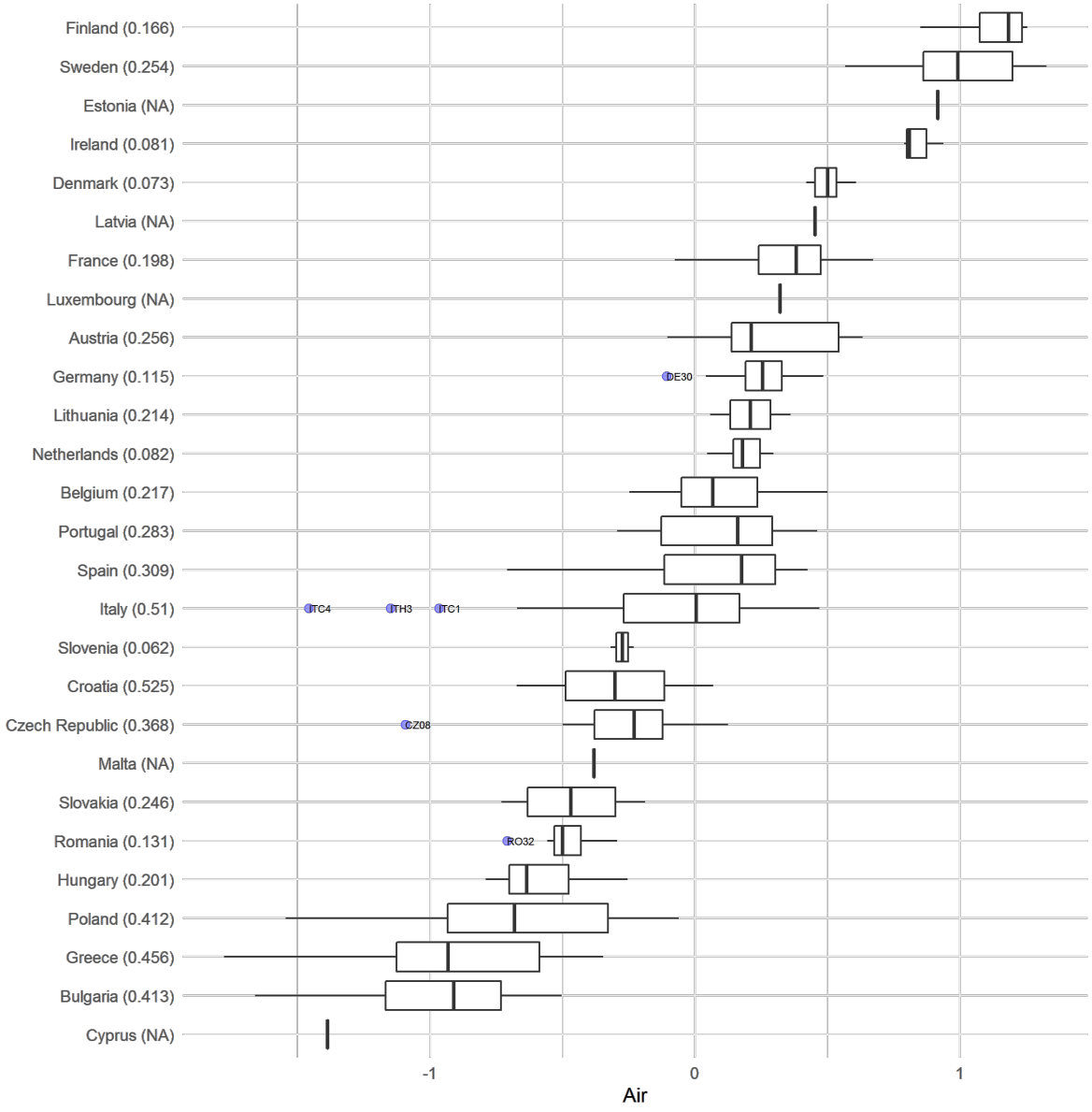


Figure 5.C.2: Boxplots of sub-national soil scores by country

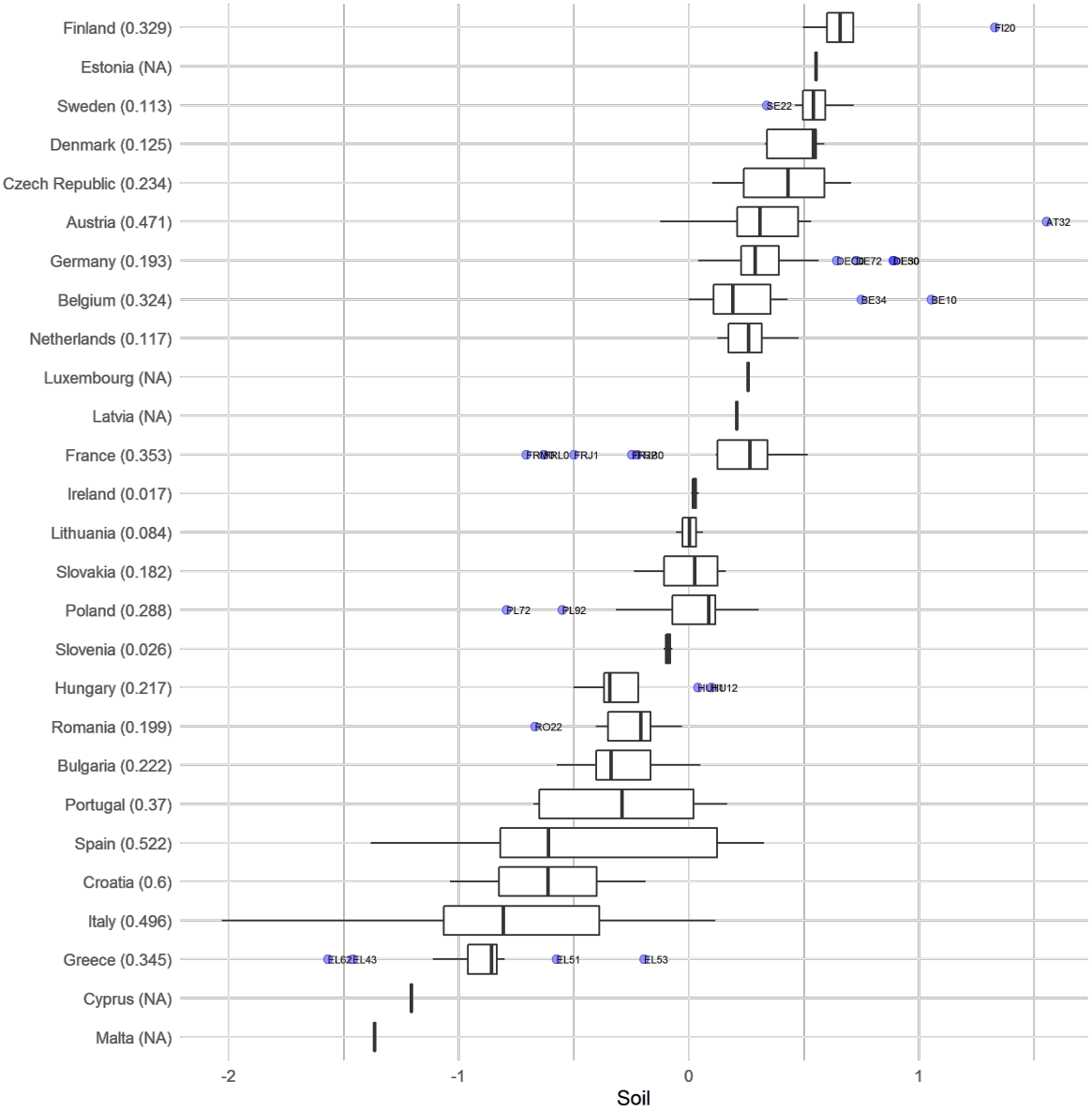


Figure 5.C.3: Boxplots of sub-national water scores by country

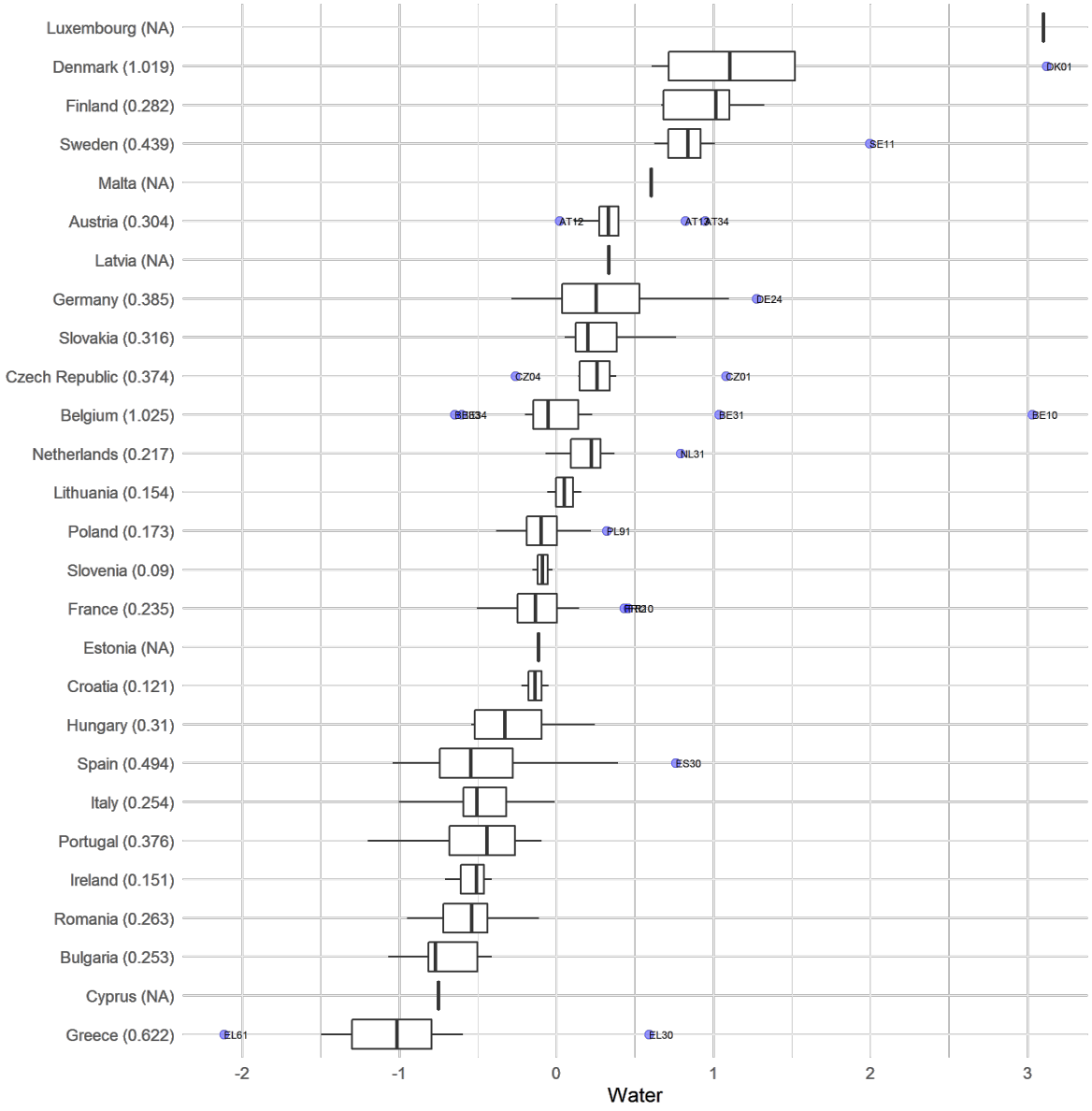
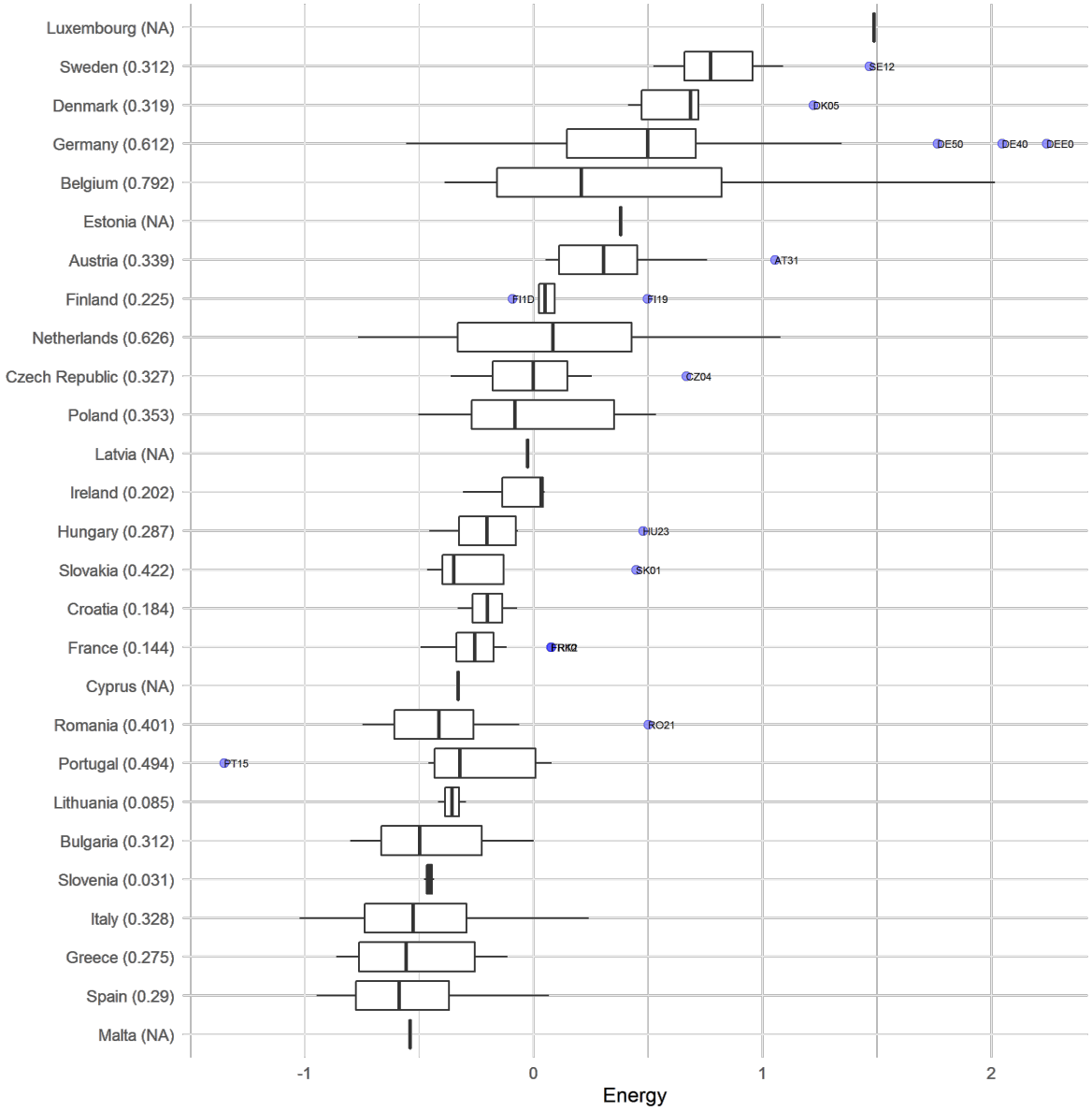


Figure 5.C.4: Boxplots of sub-national energy scores by country



5.D Overall environmental well-being

Table 5.D.1: Robustness checks: overall environmental wellbeing and quality of government

	Dependent variable: overall environmental wellbeing		
Baseline models	(1)	(2)	(3)
Quality of government	0.215***	0.224***	0.225***
	(0.024)	(0.024)	(0.024)
N	233	233	233
Wald test (df = 1)	181.297***	185.988***	182.413***
LR test (df = 1)	96.967**	100.124***	96.919***
Quality of government measured in years	2017	2013	2010
Full models	(4)	(5)	(6)
Quality of government	0.206***	0.214***	0.218***
	(0.034)	(0.033)	(0.034)
log(GDP/capita)	0.095	0.075	0.041
	(0.108)	(0.107)	(0.108)
Population density	-0.0001***	-0.0001***	-0.0001*
	(0.00004)	(0.00004)	(0.00004)
Total area	0.00000***	0.00000***	0.00000***
	(0.00000)	(0.00000)	(0.00000)
Employment in agriculture	-0.005	-0.004	-0.003
	(0.004)	(0.004)	(0.004)
Employment in manufacturing	-0.004	-0.003	-0.003
	(0.004)	(0.004)	(0.004)
Unemployment (15-75)	-0.002	-0.004	-0.008
	(0.006)	(0.005)	(0.005)
R&D expenditure	-0.00000	-0.00000	-0.00000
	(0.00005)	(0.00005)	(0.00005)
N	221	221	221
Wald test (df = 1)	96.419***	99.693***	104.059***
LR test (df = 1)	56.953***	58.968***	60.367***
Quality of government measured in year	2017	2013	2010

Note: Robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. LR test indicates the significance of the spatial autoregressive parameter ρ.

Table 5.D.2: Overall environmental wellbeing and quality of government: direct and indirect effects

	Dependent variable: overall environmental wellbeing	
	Direct	Indirect
Baseline models		
Quality of government (2017)	0.255***	0.164***
	(0.022)	(0.020)
Intermediate models		
Quality of government (2017)	0.229***	0.117***
	(0.037)	(0.023)

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

General Conclusion and Future Research

This dissertation explores the topic of well-being through diverse research questions, methodologies, and contexts. In particular, it contributes to and expands the knowledge in three main areas: predictive analytics and machine learning for well-being research, inter-generational and spillover effects, and the impact of policies and institutional quality on well-being.

Findings from each Chapter contribute to improving policy decision-making for well-being in different ways. Chapter 1 shows the potential of socio-economic life course information for predicting the risk of depression in old age. Irrespective of the specific target, these findings show that machine learning models can solve prediction problems in public policy-making and identify population groups at risk of ill-being (Berryhill et al., 2019). Chapter 2 highlights inter-generational spillovers of parental retirement and pension reforms. This finding affirms the need to evaluate public intervention spillover effects and their critical distributional consequences. Chapter 3 highlights significant ethnic disparities in market outcomes among Airbnb service providers and finds an adverse effect of anti-discrimination design interventions. This finding implies that while designing anti-discrimination interventions is necessary, they must be carefully tested and monitored to avoid adverse consequences that may exacerbate the disparities they aim to reduce. Chapter 4 applies an alternative method to construct composite indicators of well-being. This approach provides policymakers with a more holistic and precise tool for assessing well-being across different macro-regions. This approach can also help identify specific areas of need, allowing for the allocation of resources to maximize overall societal well-being. Finally, Chapter 5 estimates a strong link between the

quality of governments and environmental well-being, highlighting the role of the well-functioning institution in combating environmental degradation.

Overall, these chapters' contributions push the boundaries of well-being research while underscoring the critical role of data-driven insights in enabling effective policymaking. However, many windows remain open for further advancements. The following section outlines two research projects on my agenda inspired by some of this dissertation's findings and limitations.

Future Research Developments

Transformer-based inference on Italian Electronic Health and Labour Records

This project follows Chapter 1 of this dissertation and addresses one of its main limitations: using retrospective data to construct life biographies.

The inspiring idea is the recent publication in the machine learning literature by Savcisens et al. (2024). This paper proposes a procedure to encode individual administrative labour and health records into "sentences". The individual sentences are collections of health or job events (like specific diagnosis or job type) converted into concept tokens and ordered chronologically. The data transformed in this sentence-sequenced structure represents the training dataset for a Transformer architecture that predicts early mortality and personality nuances. Similarly to large language models (LLM), this model learns the association between life sentences and individual outcomes.

My project idea draws from this recent discovery. It explores the potential of Transformer-based models when tailored to predict individual-level labour and health outcomes based on Italian administrative yearly records, namely the WHIP and Health dataset (Bena et al., 2012). These data consist of linked administrative Italian individual health and work histories, and detailed episodes along these two life trajectories.

I want to focus on predicting three paramount health and labor variables: occupational injuries, cardiovascular diseases, and retirement timing. The significance of achieving accurate and precise predictions is particularly relevant when target-

"Health Ministry" for hospitalization events, ISTAT for mortality records, INAIL for occupational injuries, INPS for work episodes

ing such sensitive outcomes, as wrong predictions may propagate into negative socio-economic consequences (Black et al., 2011; Chaudhry et al., 2006; Miller & Galbraith, 1995).

The model will predict these outcomes as a supervised and an unsupervised prediction task. As a supervised task, I will predict the three life outcomes using a similar training pipeline as in Chapter 1. As an unsupervised task, I will study the resulting concept of embedding space on the pre-trained and fine-tuned Transformer-based model, which reveals non-trivial relationships between past life events and targeted outcomes.

The embedding is the very first main outcome generated by the model. It defines the numerical representations of all life events and their associations in the data, from diagnoses to job types and income levels. This concept space forms the foundation for the predictions.

As a final contribution, I will use the SHAP framework adopted in Chapter 1 (Lundberg & Lee, 2017) to unveil the negative or positive impact of significant life events on the predicted outcome.

Technological Change and Labour Market Discrimination – TechnoDiscrimination

Another fascinating research field I want to explore addresses the impact of artificial intelligence (AI) tools on hiring discrimination. This project relates to Chapter 3 of this dissertation and, in general, to my exploration of machine-learning techniques throughout my PhD years.

It is pretty safe to say that AI systems are changing and complementing human resources (HR) practices in the labour market. Their significant ability to speed up some HR procedures, which are costly and divert employees from other productive work, justified the widespread acceptance and deployment in this area. The transformation involves private and public institutions (Broecke, 2023). AI systems search job candidates, screen their resumes, create job ads, and assess candidate suitability for each job posting (Von Krogh, 2018). Recently, AI tools have also begun to conduct and analyze job interviews, evaluating candidates based on verbal and non-verbal cues to determine Big Five personality traits (Hickman et al., 2022).

Despite the great relevance of the phenomenon, only a paucity of data exists on

AI adoption rate in hiring practices (Broecke, 2023). This data shortage justifies the lack of empirical studies in observational settings. Some studies have appeared on the experimental side, with a general conclusion that these technologies do not automatically correct human biases and, in some cases, might even amplify them (Lippens, 2024). However, there is also some evidence of their positive effects (Avery et al., 2023; Pisanelli, 2022). What emerges is that these algorithms are, in most cases, innerly biased.

These biases derive from the dataset's characteristics used to train the algorithm and the objective function assigned. AI recruitment systems are trained using companies' historical information, including candidate resumes and their demographics. By learning from the data, the algorithm discerns patterns and predicts candidates' suitability for specific job positions (C. Li et al., 2020). Candidates' suitability increases as much as their characteristics are similar to current employees. As such, the algorithm usually selects from groups with proven track records rather than taking risks on non-traditional applicants, raising concerns about equality of opportunities (D. Li et al., 2020).

These experimental studies take the algorithm design as given, assuming its predictions are the ground truth. In other words, they fail to test if fairness-aware AI recommendations might be used to de-bias human decision-making. For example, in Avery et al., 2023 paper, they use a popular AI-assisted recruitment tool that provides applicant screening software. In the second, they used the standard version of the chatbot generative AI mode (ChatGPT). Hence, none of these studies has attempted to control the bias in the algorithm's predictive behavior or evaluated how varying such sources of biases might impact human final decisions.

In this project, I want to create an algorithm that recommends suitable candidates to recruiters but corrects the bias derived from the data (*fair algorithm*). Therefore, I plan to explore hiring outcomes from the interaction between humans and *fair algorithm* in an experimental setting.

References

Avery, M., Leibbrandt, A., & Vecci, J. (2023). Does artificial intelligence help or hurt gender diversity? evidence from two field experiments on recruitment in tech. *Evidence from Two Field Experiments on Recruitment in Tech (February 14, 2023)*.

Bena, A., Leombruni, R., Giraudo, M., & Costa, G. (2012). A new italian surveillance system for occupational injuries: Characteristics and initial results. *American Journal of Industrial medicine*, *55*(7), 584–592.

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019). Hello, world: Artificial intelligence and its use in the public sector.

Black, A. D., Car, J., Pagliari, C., Anandan, C., Cresswell, K., Bokun, T., McKinstry, B., Procter, R., Majeed, A., & Sheikh, A. (2011). The impact of ehealth on the quality and safety of health care: A systematic overview. *PLOS Medicine*, *8*(1), 1–16.

Broecke, S. (2023). Artificial intelligence and labour market matching. *OECD Social, Employment and Migration Working Papers*, (284).

Chaudhry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S. C., & Shekelle, P. G. (2006). Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*, *144*(10), 742–752.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323.

Li, C., Fisher, E., Thomas, R., Pittard, S., Hertzberg, V., & Choi, J. D. (2020). Competence-level prediction and resume & job description matching using context-aware transformer models.

Li, D., Raymond, L. R., & Bergman, P. (2020). *Hiring as exploration* (tech. rep.). National Bureau of Economic Research.

Lippens, L. (2024). Computer says 'no': Exploring systemic bias in chatgpt using an audit approach. *Computers in Human Behavior: Artificial Humans*, 100054.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

Miller, T. R., & Galbraith, M. (1995). Estimating the costs of occupational injury in the united states. *Accident Analysis & Prevention*, *27*(6), 741–747.

Pisanelli, E. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. *Economics Letters*, *221*, 110892.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2024). Using sequences of life-events to predict human lives. *Nature Computational Science*, *4*(1), 43–56.

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, *4*(4), 404–409.

List of Figures

List of Tables