Essays in quantitative economics: Improvements in measurements and macroeconomic analysis

Doctoral Thesis

in fulfillment of the

requirements for the degree of Dr. rer. pol.

from the Faculty of Economics at

Georg-August-University Göttingen

submitted by

Richard Wolfgang Haarburger

born on May 10th, 1992 in Fulda, Germany

Göttingen, October 2023

1. Referee: Prof. Dr. Tatyana Krivobokova

2. Referee: Prof. Dr. Holger Strulik

3. Referee: Prof. Dr. Sebastian Vollmer

Date of submission: October 23rd, 2023

Overview of co-authors

This dissertation includes four chapters written in joint work with co-authors and a general introduction. This section describes the contributions by each of the co-authors and myself.

1. The chapter *"Factor analysis for data with heterogeneous blocks"* is co-authored with Prof. Dr. Tatyana Krivobokova. The core idea of the chapter, the blockPCA algorithm, was developed by Tatyana Krivobokova. Both authors equally contributed to the coding involved, and to the writing of the draft. The data acquisition and preparation were conducted by Richard Haarburger.

2. The chapter *"Taking over the World? Automation and Market Power"* is co-authored with Dr. Henry Stemmler and Prof. Dr. Florian Unger. Henry Stemmler had the initial research idea. The further conceptualization of the research idea was conducted with equal contributions by Henry Stemmler and Richard Haarburger. Richard Haarburger and Florian Unger developed the theoretical model with equal contributions. The empirical analysis, literature review and the writing of the draft where conducted by Henry Stemmler and Richard Haarburger with equal contributions.

3. The chapter *"Expanding the industrial automation data universe: Prices, Production, Trade Flows"* is co-authored with Néstor Duch-Brown, PhD. Both authors contributed equally to the initial research idea. The further conceptualization, data preparation and analysis, literature review and writing of the draft were conducted by Richard Haarburger.

4. The chapter *"Interviewer Biases in Medical Survey Data: The Example of Blood Pressure Measurements"* is co-authored with Pascal Geldsetzer, PhD., Andrew Young Chang, PhD., Vivek Charu, PhD., Erik Meijer, PhD., Dr. Nikkil Sudharsanan and Dr. Peter Kramlinger. Pascal Geldsetzer and Nikkil Sudharsanan had the initial research idea. Pascal Geldsetzer, Nikkil Sudharsanan, Richard Haarburger and Peter Kramlinger further conceptualized the idea. Richard Haarburger conducted the data preparation for the IFLS and NIDS data sets, Erik Meijer conducted the data preparation for the LASI data set. The development of the methodology and the corresponding coding were conducted mostly by Richard Haarburger and Peter Kramlinger with equal contributions, Erik Meijer, Pascal Geldsetzer and Nikkil Sudharsanan contributed with helpful comments. Vivek Charu provided feedback on the methodology afterwards. Andrew Young Chang and Richard Haarburger provided major contributions to the writing of the draft, Pascal Geldsetzer and Peter Kramlinger provided minor contributions.

Acknowledgements

I would like to dedicate this page to the people who have accompanied and encouraged me on the journey to my doctoral degree. First of all, I would like to thank my principal supervisor, Tatyana Krivobokova. Her lightning-fast feedback and tireless drive made it possible to bridge any geographical distance. Thanks to her optimism and creative mind, we have always found a way to move forward with our research and explore new avenues.

I also would like to thank my second supervisor, Holger Strulik, who has accompanied me throughout all my academic endeavours. His lectures on growth and development were instrumental in shaping my academic career and set the course that led to this doctoral dissertation. I always appreciated his thoughtful feedback, even if it often meant I had to take two steps back before I could take a another step forward.

Thanks are also due to Stephan Klasen. He encouraged me to start a dissertation that would connect different disciplines within economics and I learned an incredible amount from that. Unfortunately, he passed away in 2020.

I am also grateful to my third supervisor, Sebastian Vollmer, for integrating me into his chair, when the logistical circumstances meant that I would have become a satellite. Moreover, I'm grateful to him for believing and investing in a research project that unfortunately did not make it into the thesis. However, only via his network have I met most of the co-authors of the fourth chapter of this thesis, for which I'm very grateful.

I would also like to thank the German Research Foundation, which funded my research as a part of the Research Training Group 1723 - Globalization and Development. In this context I would also like to thank my fellow doctoral students of the research training group, for countless inspiring conversations and joyful moments over the years - Henry Stemmler, Laura Barros, Anna Gasten, Tatiana Orozco García, Claudia Schupp, Lukas Wellner, Johannes Matzat, Anna Reuter, Lisa Rogge, Andrea Cinque, Tobias Korn and Yuanwei Xu.

A special thanks also goes to Krisztina Kis-Katos who did a phenomenal job as the head of the research group in Göttingen. The research training group would not have been functional without Antje Juraschek. Together, we have solved every bureaucratic problem to date, no matter how complicated.

I would also like to thank all the people at the Joint Research Centre of the European Commission, who welcomed me with open arms and taught me most, if not all, I know today about policy work as an economist during my research visit. A special thanks to Néstor Duch-Brown, Daniel Nepelski and Sarah de Nigris for making that time so enriching and exciting.

I would also like to thank some of my co-authors not, or not sufficiently, mentioned so far. Henry Stemmler, for the countless hours spent honing in on research ideas in our joint office, that I will always remember fondly. Florian Unger, for his guidance in economic modelling. Pascal Geldsetzer, for the many opportunities and all the inspiration he provided. Peter Kramlinger, for the lively exchange and our shared adventures in RStudio.

A big thank you to my family, my parents, Kathrin and Tobias, who always gave me total freedom, blind trust and support in all my endeavours, my grandmother Shirley, Klaus, with whom I will have in common in the future, being a former, nostalgic student of Göttingen University.

A special thank you also to the friends outside of my immediate circle of colleagues who I got to know and appreciate during this time. My former flatmate, Alexander Lange, for great company in the Covid induced months of home office. Felix Turbanisch, for the work breaks spent walking across the often pitch-dark north campus, whatever the weather, discussing research and life. And all the other companions who were by my side before this episode and will continue to be.

Contents

List of Tables

List of Figures

List of Acronyms

2SLS	2 Stage Least Squares
AI	Artificial Intelligence
ADF	Augmented Dickey-Fuller
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Predictor
BMI	Body Mass Index
bPCA	Block Principal Component Analysis
CAPI	Computer-Assisted Personal Interview System
CEBs	Census Enumeration Blocks
DHS	Demographic and Health Surveys
EA	Enumeration Area
EM	Expectation-Maximization
EMX	Edmond-Midrigan-Xu
GDP	Gross Domestic Product
GMM	Generalized Method of Moments
HHI	Herfindahl-Hirschman-Index
HS	Harmonized System
IFLS	Indonesia Family Life Survey
IFR	International Federation of Robotics
IIPS	International Institute for Population Sciences
ILO	International Labor Organization
ISIC	International Standard Industrial Classification of All Economic Activities
IV	Instrumental Variable
JRC	Joint Research Centre
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LASI	Longitudinal Aging Study in India
LMICs	Low- and Middle-Income Countries
LMM	Linear Mixed Model
LRT	Likelihood Ratio Test
MC	Marginal Cost
MFN	Most Favoured Nation
mm Hg	Millimeters of Mercury
MSE	Mean Squared Error
NIDS	National Income Dynamics Study of South Africa
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCR	Principal Component Regression
PPI	Producer Price Index
PLS	Partial Least Squares
PSUs	Primary Sampling Units
RCA	Revealed Comparative Advantage
TFP	Total Factor Productivity
TRAINS	Trade Analysis Information System
UK	United Kingdom
UN	United Nations
UNCDAT	United Nations Conference on Trade and Development
USD	United States Dollar

Chapter 1

General Introduction

1.1. Research Objectives and Contributions

The importance of data is undoubtedly on the rise. In an increasingly globalized, digitally interconntected world, the amount of data generated every day is constantly growing (Manyika et al., 2011). The interpretation of data is the primary tool we have at our disposal to calibrate economic models and test the hypotheses we derive from them (Sims, 1980; Angrist & Pischke, 2009). In academic research, data serves as an intermediary between theory and reality, and thus the growing availability of data offers an abundance of opportunities to answer more and more research questions with more and more precision. This dissertation explores various facets of the use of data in modern economic research and sheds light on some of the economic implications of technological change.

This dissertation was written as part of the Research Training Group Globalisation and Development. The specific project in which I participated as a member of this training group is called "Measurements and Methods". While the project was designed to focus mainly on the methodology in dealing with data, the thematic influence of the research group is apparent. It is noticeable in the research questions and applications pursued throughout the thesis that highlight a range of opportunities and obstacles quantitative researchers face in their work.

The fact that challenges with methods and the measurement of data are relevant to all empirical sub-fields of economic research is reflected in the diversity of the thematic areas of this thesis. This thesis aims to contribute in the field of applied statistics, macroeconomics with a focus on competition economics under technological change, as well as to epidemiological research in the field of global health.

Increasingly complex data structures and larger data sets bring new challenges alongside

the aforementioned opportunities. The so called curse of dimensionality poses a problem in empirical research that is growing increasingly prevalent with data sets becoming more complex (Hastie et al., 2009). One objective of this dissertation is to demonstrate the challenges related to the curse of dimensionality using a concrete example, and to propose a method for this specific case that aims to minimize the downsides of complexity while preserving the upsides.

However, this dissertation focuses not only on data and methods but also on the perpetual advancement of technology, which is largely driven by the automation of processes (Brynjolfsson & McAfee, 2014). The increasing use of robots and artificial intelligence have been controversially discussed in the literature, but also outside of it, for several years. The impact on labor markets and the centralization of power among individual entities such as technology firms who are global pioneers in these fields are the subject of ongoing debates.[1] This gives rise to another question that this dissertation explores. Namely, to what extent firms that adapt new technologies particularly quickly and efficiently gain market power and displace less technologically advanced competitors.

Pursuing this research question led to the emergence of another critical, re-occurring aspect of empirical work, namely issues with data quality. Deficiencies in data quality often pose challenges to researchers. This dissertation addresses two of these deficiencies. The first one involves the presence of many missing entries, which makes the data almost impossible to use with conventional methods (Little & Rubin, 2019). One objective of this dissertation is to show, with a concrete example, how specialized methods and the combination of different data sources can be used to fill such gaps in data in a way, that accounts for the uncertainty underlying the imputations so that it can be taken into account when further utilizing the data. As part of this thesis, a variety of industrial automation data will be presented, providing valuable resources for future research projects and addressing a gap in the data available in this specific field of research.

Another potential deficiency in empirical research that is often overlooked is the presence of data bias. Failure to recognize and address these biases can result in misleading conclusions and erroneous interpretations of data (Rothman et al., 2008). This dissertation aims to investigate this issue in greater detail and intends to provide an illustrative example of how measurement error can distort the reality represented by the data collected. Unbiased and representative data form the foundation of data-driven decision-making, highlighting the crucial role of unbiased research findings for policy makers.

1.2. Summary of the chapters

This dissertation presents four independent studies, comprising a chapter each, that aim to address the objectives of the thesis. The following summarizes the chapters, providing

[1] See, e.g., the different assumptions and findings on automation induced net effects on job displacement versus creation in Prettner & Strulik (2017); Frey & Osborne (2017); Acemoglu & Restrepo (2020); G. Graetz & Michaels (2018); Aghion et al. (2020, 2022), amongst others.

some background information about the creation process.

1.2.1. Chapter 2: *Factor analysis for data with heterogeneous blocks*

The second chapter is co-authored with Prof. Dr. Tatyana Krivobokova. The economic background for this methodological paper is the literature on macroeconomic uncertainty (Bloom, 2009; Stock & Watson, 2012; Baker & Bloom, 2013; Bloom, 2014), that finds it to have mostly growth mitigating effects via various channels. One of these channels are so called wait-and-see effects, where economic agents delay their economic activity in times of uncertainty, because the presence of increased uncertainty implies more difficult risk assessment (Bloom, 2014). In the interest of fostering economic growth, and to counteract these wait-and-see effects, one approach is to provide now- and forecasts to economic agents, that are intended to reduce uncertainty about the current economic situation and future economic development Stock & Watson (2002a).

A common method linking economic uncertainty to the creation of now- and forecasts are so called factor-based now- and forecasting models, introduced by Stock & Watson (2002a,b). In this type of model, a set of reductive factors is extracted from the available macroeconomic time series data, which is informative about various facets of the state of the economy. These factors are then used as predictors for key economic indicators as GDP growth or industrial production in regression models. Typically, the factor estimation methods used to extract these underlying factors are designed to impose orthogonality on the factors. Stock & Watson (2002a,b) argue, that the orthogonality enables a link of the factors to distinct forms of macroeconomic uncertainty. These models have been shown to perform relatively well in now- and forecasting exercises (see, e.g., Boivin & Ng 2005). Subsequent research by Boivin & Ng (2006) showed that extracting the factors from ever larger data sets in the numbers of variables they contain, does not necessarily improve the performance of such models, and in some cases even deteriorates it. While abundant now- and forecasts are available for advanced economies and certain sparse model specifications have been repeatedly shown to perform well, the same can not be said in the case of developing countries and emerging market economies.

However, especially in the case of emerging market economies, macroeconomic time series data has become much more available over the last two decades (see, e.g., Cepni et al. 2020; Li & Chen 2014; Porshakov et al. 2016; Gupta & Kabundi 2011). The simultaneous abundance of such data, and the problem of performance deterioration observed with canonical methods led to a series of publications mostly in the field of sparse statistical modelling (see, e.g., Zou & Hastie 2005; Zou et al. 2006; Bai & Ng 2017; Ayesha et al. 2020).

In chapter one, we suggest that noise level differences between groups of data can cause the omission of relevant factors using the conventional factor extraction methods. We demonstrate this phenomenon with simulations in a controlled data environment and introduce blockPCA, an algorithm that clusters the data into blocks and extracts the factors from these blocks separately in a first step. In a second step, the factors are

concatenated and a second set of factors is extracted from the first set of concatenated factors. The resulting second set of factors is then used as predictors for key economic indicators. Using the original Stock & Watson (2002b) data revisited by Boivin & Ng (2006), and five macroeconomic data sets from emerging market economies, we show that this algorithm is much more robust to factor omission than the conventional factor extraction methods and thus performs better in a factor-based model setting.

1.2.2. Chapter 3: *Taking over the World? Automation and Market Power*

Chapter three is joint work with Dr. Henry Stemmler and Prof. Dr. Florian Unger. This chapter establishes a link between the recent literature on the measurement of global market power and concentration (see, De Loecker & Warzynski 2012; De Loecker & Eeckhout 2018) and the literature on industrial automation, which in many cases utilizes the macroeconomic panel data documenting the adoption of industrial robots by the International Federation of Robotics (IFR, Müller & Kutzbach 2019), see, e.g., G. Graetz & Michaels (2018); Acemoglu & Restrepo (2020); Krenz et al. (2021); Artuc et al. (2023). In general, automation technology comprises mostly either the use of automation enabling hardware such as industrial robots or the implementation of applications based on artificial intelligence often using image or sensor data as inputs (de Nigris et al., 2022). In this chapter we focus on the adoption of the industrial robots reported by the IFR.

This chapter is related to the literature on so-called "superstar" firms, which is a term coined by D. Autor et al. (2020) and describes high-tech firms excelling in their markets, claiming ever increasing market shares. Most closely related to our work is probably the work by Stiebale et al. (2020), who, using different data sources, aim to answer a similar research question as we aim to answer in this chapter, namely, whether firms pioneering in the adoption of automation technology manage to increase their market power, measured by the markup of price over marginal cost. While the work by Stiebale et al. (2020) builds on data from European firms, our analysis uses global firm data and thus extends to the global economy. We are also interested in the overall association between the increasing uptake of industrial automation via robots and changes in market concentration. Our research question implies that firms are heterogeneous in how quickly they adopt new technologies, which potentially improves their productivity. In order to derive a set of hypotheses about firms' heterogeneous responses to the increasing availability of robots and the implications for their market shares and market concentration as a whole, we modify the model of oligopolistic competition in Edmond et al. (2015). This modified model includes robots as a factor input and introduces firm heterogeneity in the extent to which robots are utilized in production. It predicts that firms with above average robot intensity will benefit from the increased availability of robots in contrast to firms with below average robot intensity. Moreover, it predicts that foreign high-robot-intensity firms selecting into exporting will exert downward pressure on the market shares of domestic firms in a symmetric two-countries version of the model.

Our empirical analysis shows that the increased adoption of industrial robots has had negative effects on average markups, but that firms in the highest markup quintile have experienced automation induced gains in market shares. Moreover, we also find that increasing automation of foreign competitors exerts downward pressure on local firms' markups across all quintiles, as predicted by the model. These findings corroborate the findings of the related superstar firm literature.

1.2.3. Chapter 4: *Expanding the industrial automation data universe: Prices, Production, Trade Flows*

The fourth chapter is joint work with Néstor Duch-Brown, PhD., who was one of my supervisors during my research stay at the Joint Research Centre (JRC) of the European Commission in Ispra, Italy. Having worked on the third chapter of this thesis it became apparent to me that data on the cost of automation, a variable frequently occurring in theoretical work modelling firms' decision to automate, lacked an empirical counterpart. In fact, the empirical analysis in chapter three relies on a proxy for the increased availability of industrial robots, which in the theoretical model is expressed as a reduction in the robot rental rate. At the time I joined the team at the JRC it was tasked to write a policy report discussing the evolution of European firms' market shares in the robot industry. It became clear, that the main data source available on country-level exports of industrial robots were the Comtrade data (UN, 1990). Unfortunately the Comtrade data posed many challenges that ended up being addressed following a very different strategy in the policy report than we utilize in this chapter (Duch-Brown et al., 2021; Duch Brown et al., 2023).

The lack of data on country-level production of industrial robots, data on the evolution of prices and trade flows were the starting point for this chapter. Using the Comtrade data alongside data from the IFR (Müller & Kutzbach, 2019), the OECD (OECD, 2015, 2021, 2023a,b), UNCDAT (UNCDAT, 2018), and some minor data sources, we derive new data and thereby contribute to the available data on industrial automation.

The main challenge inherent to the Comtrade data is the high degree of missing entries for traded robot units. We impute these missing values using a sophisticated imputation algorithm called `Amelia` (Honaker et al., 2011) that draws information from related data sources. From the imputed Comtrade data, we can derive an origin-destination-matrix of industrial robots covering 64 countries over the 1996-2018 period. The combination of country-level robot installations by the IFR, traded volumes in US Dollars, kilograms and units renders the derivation of unit prices and country-level robot production possible.

We explore the newly derived data using various descriptive statistics, such as measures for market concentration, market stability, the evolution of robot prices over time by exporting country, by weight quantile of the robot, etc. We find that a few robot exporting countries dominate the market claiming high market shares and that market concentration is not subject to large changes over time. Moreover, we find that robot prices adjusted for inflation have been declining over the period under consideration even without accounting for the

most likely having increased capabilities and capacities of the robots. Remarkably, we also observe a clear pattern of convergence amongst prices of robots across weight quantiles, potentially hinting that software could make up for an increasingly large share of the value added in industrial robots compared to hardware.

Finally, we demonstrate how the newly derived data on robot prices can be utilized in empirical research as a counterpart to the cost of automation often modelled theoretically. We do so by replicating part of the analysis from chapter three, comparing the estimation results with the ones obtained using the aforementioned proxy for the increased availability of industrial robots. We also compare the two instrumental variables constructed using the novel price data versus using the availability proxy based on standard metrics. We can confirm the findings from chapter three and find that instrument strength is comparable, however, we argue that the exclusion restriction is better complied using the price based instrumental variable.

1.2.4. Chapter 5: *Interviewer Biases in Medical Survey Data: The Example of Blood Pressure Measurements*

This chapter is joint work with Pascal Geldsetzer, Andrew Young Chang, Vivek Charu (all three Stanford University), Erik Meijer (University of Southern California), Nikkil Sudharsanan (Technical University of Munich) and Peter Kramlinger (University of California Davis). It addresses the often overlooked problem of biased data, in this case the biased measurement of medical survey data. Medical survey data is routinely collected to obtain representative data on populations and serves as the empirical foundation of research aiming to answer various medical, often epidemiological, research questions (see, e.g., Cockburn et al. 2023; Rahim et al. 2023a,b). Moreover, the prevalences of diseases estimated based on such medical survey data are often used by policy makers to asses the burden and development of such diseases and depict an important input to their decision-making process in taking measures to counteract them (for an influential contribution, see Ezzati et al. 2002).

Some well-established survey-based data sources designed to be nationally representative are the Demographic and Health Surveys (DHS), the Indonesia Family Life Survey (IFLS, Strauss et al. 2009; Sikoki et al. 2013; Strauss et al. 2016), the National Income Dynamics Study of South Africa (NIDS, Southern Africa Labour and Development Research Unit 2018a,b,c,d,e) and the Longitudinal Aging Study in India (LASI, International Institute for Population Sciences (IIPS), MoHFW, Harvard T. H. Chan School of Public Health (HSPH) and the University of Southern California (USC) 2020).

Data collection for these survey-based data sets is typically conducted by interviewer teams that are comprised of non-healthcare worker personnel. In some cases, training of such personnel may be insufficient to prevent variations in interviewer technique and demeanour impacting the measurements taken. This interviewer induced measurement bias is commonly referred to as "interviewer effects" (Svensson & Theorell, 1982; Ulijaszek

& Kerr, 1999; Ali & Rouse, 2002; Cernat & Sakshaug, 2020). Given the importance of the implications derived from these data sets and the fact, that these effects receive very little attention in the literature, this chapter aims to investigate the presence of interviewer effects in the IFLS, NIDS and LASI data. Unfortunately, the DHS data does not provide the necessary interviewer IDs in the data, so that it had to be excluded from the analysis.

We focus on the measurement of blood pressure, which is important to assess the prevalence of hypertension, the disease referring to elevated blood pressure. We employ a linear mixed model in which the interviewer effects are modelled as random effects, following the reasoning in Hodges (2013). This model allows us to adjust the observed measurements for the estimated interviewer effects. Using a bootstrap approach, we can then sample subsets of the adjusted and unadjusted data to quantify the uncertainty inherent to the prevalence of hypertension. While we find that hypertension prevalences were not substantially impacted at national level, we find numerically small, but significant interviewer effects. The smaller the geographic division however, the higher the risk that an extreme interviewer could cause substantial bias in measured prevalences. This is important, since estimates from smaller areas are increasingly used for mapping disease prevalences at subnational levels, sometimes in areas as small as a few square kilometres (Dwyer-Lindgren et al., 2019; Reiner Jr et al., 2018; Osgood-Zimmerman et al., 2018; N. Graetz et al., 2018).

Chapter 2

Factor analysis for data with heterogeneous blocks

Abstract

Factor-based now- and forecasting models are known to excel at capturing latent macroeconomic uncertainty and handling high-dimensional data. However, the increasing availability of macroeconomic time series data has revealed challenges. The predictive accuracy of such models often deteriorates when additional variables are added, in part due to relevant factors being dominated by other factors and thus not being adequately accounted for in the final regression model. We provide a theoretical foundation, highlighting noise level differences between groups of variables as a key driver of factor omission. In response, we introduce "blockPCA", a novel algorithm that preserves the strengths of PCR-based factor models while mitigating factor domination. BlockPCA identifies variable groups and separates them into distinct blocks, extracting factors from each block separately. These factors are then concatenated, and a second set of factors is derived from the resulting composite matrix, which serve as regressors in the final regression model. The application of BlockPCA to five wide datasets from emerging economies and a long dataset often revisited in the literature yields considerable improvements in industrial production nowcasting compared to conventional factor extraction methods.

This chapter is joint work with Tatyana Krivobokova (University of Vienna). We appreciate the feedback and suggestions received from the participants of the 41^{st} International Symposium on Forecasting 2021 and the 14^{th} International conference of the ERCIM, 2021, London.

2.1. Introduction

Over the past two decades, factor-based models have become a well-established tool in economic now- and forecasting (Stock & Watson, 2002a,b; Boivin & Ng, 2005). The more accurate the now- forecasts, the more they can be expected to counteract macroeconomic uncertainty, which is often associated with a slowdown in economic activity and growth (Bloom, 2009; Stock & Watson, 2012; Baker & Bloom, 2013).

In addition to factor-based models, several other econometric approaches have been introduced in the literature that are well suited to the large-scale macroeconomic panel data typically used in this context (see, e.g., Kelly & Pruitt 2015). Factor models, however, have often been associated with two distinct qualities. First, as Stock & Watson (2002a,b) state, the estimated factors can be seen as representations of different forms of economic uncertainty and thus the link to economic theory is considered compelling. Second, they were originally associated with the idea that no precise selection of variables needed to be made in preparation for the analysis, since relevant factors could be extracted even from large amounts of data.

However, with the steadily increasing availability of macroeconomic time series data, it became clear that more data does not necessarily improve the performance of such models (Boivin & Ng, 2006). In particular, Boivin & Ng (2006), suggest that under certain properties of the data, adding more variables deteriorates now- and forecasting results when using the method of principal component analysis (PCA) to extract the factors. In addition to naming cross-correlated errors as one of the reasons for this observation, they discuss the phenomenon of some factors being dominated by others. These findings led to a series of subsequent publications, mainly in the area of sparse modeling and thus often dealing with how to make the best possible preselection of variables (e.g., Zou & Hastie 2005; Zou et al. 2006; Bai & Ng 2017, for an overview see Ayesha et al. 2020).

In this paper, we propose an algorithm that retains the aforementioned advantages of principal-component-regression (PCR) based factor models while being robust to large data sets. This algorithm poses an alternative to pretesting variable combinations, while maintaining the link to economic theory.

We argue that large differences in noise levels between different groups of variables can lead to factors from some of these groups being dominated and thus omitted in the final regression model. Based on these theoretical remarks, we introduce "blockPCA", an algorithm that first clusters the input data into distinct blocks and then extracts factors from those blocks separately. In a second step, the extracted factors are used to generate new, more reductive factors that act as regressors in the nowcasting model.

Our simulations confirm that this approach estimates the true underlying factors more stably than the conventional principal component regression in a setting that imposes the aforementioned differences in noise levels between blocks.

Using long and wide macroeconomic data, we compare the accuracy of factor-based nowcasts generated using conventional factor extraction methods versus using blockPCA. As long

data, where the number of observations substantially exceeds the number of variables, we utilize the original data from Stock & Watson (2002b) that was revisited by Boivin & Ng (2006) to demonstrate that additional variables can be detrimental to now- and forecasting results. As wide data, i.e., data where the number of variables substantially exceeds the number of observations, we utilize large-scale macroeconomic panel data sets from five emerging market economies.

The paper is structured as follows. First, we introduce the basic factor Stock & Watson (2002a,b) model in section 2.2, from which we depart to then provide some theoretical background on how relevant factors become dominated so that they are not sufficiently taken into account in typical factor models in section 2.3. Section 2.4 demonstrates the problem of factor omission in a controlled data environment, before we move to observational data in section 2.5 to investigate the differences in nowcasting performance by method. Finally, section 2.6 concludes the paper.

2.2. Basic factor model and motivation

2.2.1. Basic factor model

The basic factor model on which subsequent research has been based can be traced back to Stock & Watson (2002a,b). It is based on the idea that there are various forms of macroeconomic uncertainty that cannot be measured directly. However, they can be thought of as latent variables, that can be represented by the orthogonal factors of a principal component analysis. Research has shown that these factors serve as useful predictors in a panel data regression setting with an outcome variable that is affected by latent uncertainty, such as economic growth. This basic model is conventionally estimated with ordinary least squares and often includes a lag term in addition to the extracted factors. Assuming that the N dimensional time series X_t can be represented by a factor structure, that is, it can be written as

$$X_t = \Lambda F_t + e_t, \quad t = 1, \ldots, T, \tag{2.1}$$

where Λ is the loading matrix of dimension $N \times p$, $p \leq N$, F_t is the p-dimensional vector of extracted factors used as regressors in the forecasting regression model, and e_t is the residual variation in X_t unexplained by the factor structure. The specification of the forecasting regression model takes the form

$$y_{t+h} = F_t^T \beta_F + \omega_t^T \beta_\omega + \epsilon_{t+h} \tag{2.2}$$

where F_t contains the factors extracted from the original explanatory data as specified in equation 2.1, ω_t is a vector containing lags of y_t, ϵ_{t+h} is an i.i.d. error term and β_F, β_ω are unknown regression coefficients. The index t indicates a specific point in time, the index h indicates how many steps into the future the target variable y is being forecasted. In

the case of nowcasting, $h = 0$, so that a prediction for y_t is made in a period for which observations of the explanatory variables in t are already available. Different methods are available for extracting the factor matrix F_t from X_t. However, the originally proposed and most widely used approach is the method of principal components. The use of the factors extracted by principal component analysis is commonly referred to as principal component regression (Jolliffe, 1982). In contrast to the method of maximum-likelihood based factor analysis, the factors extracted with PCA are orthogonal to each other, and in line with economic theory, therefore, represent distinct components of uncertainty. The advantage of principal components is that large data sets, which would lead to underspecification in a normal regression model, can be reduced to a few regressors while preserving most of the variation in the original data.

However, Boivin & Ng (2006) find that adding more data to a principal component based forecasting model does not necessarily improve forecast accuracy, and under certain circumstances even worsens it. The authors suggest two possible explanations of this finding. First, they note that worse forecasting results induced by adding more data to the model may occur when the idiosyncratic errors are cross-correlated, and second, relevant factors may be dominated by other factors in large datasets. In this paper, we focus on the second point and propose an algorithm that builds on the typical principal component approach and makes it robust to relevant factors being dominated.

2.3. Factor analysis of heterogenous data

We now provide the theoretical background on how relevant factors can be dominated by other factors in the context of principal component analysis.

2.3.1. Factor models and principal component analysis

Let $x = (x_1, \ldots, x_N)^T$ be a random vector. Without loss of generality let $\mathrm{E}(x) = 0_N$. Let $S = \mathrm{diag}(s_1^{-1}, \ldots, s_N^{-1})$ with $s_i^2 = \mathrm{var}(x_i)$, $i = 1, \ldots, N$. Assume for the standardised Sx a population factor model, that is, $Sx = \Lambda f + e$, where $\Lambda \in \mathbb{R}^{N \times p}$, is the fixed matrix of unknown loadings of full rank $p < N$, $f \in \mathbb{R}^p$ is the random vector of factors and $e \in \mathbb{R}^N$ is the random error term.
The standard assumptions on the factors and error terms are $\mathrm{E}(f) = 0_p$, $\mathrm{E}(e) = 0_N$, $\mathrm{cov}(e, f) = 0_{N \times p}$, $\mathrm{cov}(f) = I_p$. To simplify subsequent notations and calculations we assume that $\mathrm{cov}(e) = \sigma^2 I_N$. With these assumptions follows that $\Sigma = \mathrm{cov}(Sx) = \mathrm{cor}(x) = \Lambda\Lambda^T + \sigma^2 I_N$.
From the eigendecomposition of $\Lambda\Lambda^T = UDU^T$ with $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p, 0_{N-p})$, it follows that up to a rotation $\Lambda = U_p D_p^{1/2}$, where $U_p \in \mathbb{R}^{N \times p}$ is the matrix of first p columns of the matrix of eigenvectors U and $D_p = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Hence, one could estimate the factors via $\hat{f} = D_p^{-1/2} U_p^T x$. Since the loadings are not observable, one

can derive U_p and D_p using the identity

$$\Sigma = \Lambda\Lambda^T + \sigma^2 I_N = UDU^T + \sigma^2 I_N = U(D + \sigma^2 I_N)U^T = UD_\sigma U^T,$$

where $D_\sigma = \text{diag}(\eta_1, \ldots, \eta_N)$, with $\eta_i = \lambda_i + \sigma^2$, $i = 1, \ldots, p$ and $\eta_i = \sigma^2$ for $i = p+1, \ldots, N$. Note that since all eigenvalues of Σ are shifted by the same number σ^2, the order of the eigenvalues and herewith eigenvectors of Σ and $\Lambda\Lambda^T$ is exactly the same. Hence, U_p can be derived as the first p eigenvectors of Σ and λ_i as the i-th eigenvalue of Σ, reduced by σ^2.
Observing T realisations of the random vector x leads to the sample factor model $SX_t = \Lambda F_t + E_t,\ t = 1, \ldots, T$, where X_t, F_t and E_t are the t-th (independent) realisation of x, f and e, respectively. Given the sample correlation matrix $\widehat{\Sigma}$ of X_t, one can derive estimators of loadings (up to a rotation) $\widehat{\Lambda} = \widehat{U}_p \widehat{D}_p^{1/2}$ and of factors $\widehat{F}_t = \widehat{D}_p^{-1/2} \widehat{U}_p^T X_t$. Matrix $\widehat{U}_p$ contains the first p eigenvectors of the sample correlation matrix $\widehat{\Sigma}$ and $\widehat{D}_p = \text{diag}(\hat{\eta}_1 - \hat{\sigma}^2, \ldots, \hat{\eta}_p - \hat{\sigma}^2)$, where $\hat{\eta}_i$, $i = 1, \ldots, p$ are the first p eigenvalues of $\widehat{\Sigma}$ and $\hat{\sigma}^2 = (N-p)^{-1} \sum_{i=p+1}^{N} \hat{\eta}_i$ estimates σ^2.
It has been shown under various asymptotic scenarios for N and T that the first p (scaled empirical) principal components of $\widehat{\Sigma}$ are consistent estimators of the factors f even under much more general assumptions on $\text{cov}(E_t)$; also, the assumption on independence of realisations X_t can be relaxed, see e.g., Boivin & Ng (2006); Bai et al. (2008); Stock & Watson (2011).

2.3.2. Factor analysis on heterogeneous blocks

In practice, the random vector $x \in \mathbb{R}^N$ often contains groups of intrinsically different variables, that are typically measured on different scales and may have different magnitudes of the eigenvalues of corresponding correlation matrices together with the different noise levels.
To simplify the notation, we will consider only two heterogeneous groups, the extension to more groups is straightforward. Let $x = (x_1^T, x_2^T)^T \in \mathbb{R}^N$ with $x_j \in \mathbb{R}^{N_j}$, $j = 1, 2$ and $N_1 + N_2 = N$. Assume for each x_j that $S_j x_j = \Lambda_j f_j + e_j$, with unknown loading matrices $\Lambda_j \in \mathbb{R}^{N_j \times p_j}$ of full rank $p_j < N_j$, $p_1 + p_2 = p$, random factors $f_j \in \mathbb{R}^{p_j}$, random errors $e_j \in \mathbb{R}^{N_j}$ and diagonal matrices of inverse standard deviations $S_j = \text{diag}(s_{j1}^{-1}, \ldots, s_{jN_j}^{-1})$. As in section 2.3.1 we assume $\text{E}(f_j) = 0_{p_j}$, $\text{E}(e_j) = 0_{N_j}$, $\text{cov}(e_j, f_j) = 0_{N_j \times p_j}$, $\text{cov}(f_j) = I_{p_j}$, as well as $\text{cov}(e_j) = \sigma_j^2 I_{N_j}$ with $\sigma_1^2 \neq \sigma_2^2$. We also assume that $\text{cov}(f_1, f_2) = 0_{p_1 \times p_2}$ and $\text{cov}(e_1, e_2) = 0_{N1 \times N_2}$, which implies that $\text{cov}(x_1, x_2) = 0_{N_1 \times N_2}$. The assumptions on independence between x_1 and x_2 can be relaxed. Then the correlation matrix of x results in

$$\begin{aligned}
\Sigma &= \text{cov}(Sx) = \text{cor}(x) = \text{blockdiag}(\Lambda_1\Lambda_1^T + \sigma_1^2 I_{N_1},\ \Lambda_2\Lambda_2^T + \sigma_2^2 I_{N_2}) \\
&= \text{blockdiag}\left(U_1 D_1 U_1^T + \sigma_1^2 I_{N_1},\ U_2 D_2 U_2^T + \sigma_2^2 I_{N_2}\right) \\
&= \begin{pmatrix} U_1 & 0_{N_1 \times N_2} \\ 0_{N_2 \times N_1} & U_2 \end{pmatrix} \begin{pmatrix} D_1 + \sigma_1^2 I_{N_1} & 0_{N_2 \times N_1} \\ 0_{N_1 \times N_2} & D_2 + \sigma_2^2 I_{N_2} \end{pmatrix} \begin{pmatrix} U_1^T & 0_{N_1 \times N_2} \\ 0_{N_2 \times N_1} & U_2^T \end{pmatrix} \\
&= UD_\sigma U^T,
\end{aligned}$$

where $D_j = \text{diag}(\lambda_{j1}, \ldots, \lambda_{jp_j}, 0_{N_j - p_j})$ with $\lambda_{j1} > \lambda_{j2} > \ldots > \lambda_{jp_j}$, $j = 1, 2$.
Obviously, the ordering of the eigenvalues in the matrix D_σ depends on the magnitude of σ_j^2 and λ_{ji}, $i = 1, \ldots, p_j$, $j = 1, 2$. If $\min\{\lambda_{1p_1} + \sigma_1^2, \lambda_{2p_2} + \sigma_2^2\} > \max\{\sigma_1^2, \sigma_2^2\}$, then all relevant eigenvectors corresponding to the non-zero eigenvalues $\lambda_{j1}, \ldots, \lambda_{jp_j}$, $j = 1, 2$ would enter the first p eigenvectors of matrix Σ and factors $f = (f_1^T, f_2^T)^T$ can be estimated by the first p (scaled) principal components of Σ. In particular, this happens, when λ_{1i} and λ_{2i} have a very similar magnitude and a decay rate, while σ_1^2 and σ_2^2 are either relatively small or equal to λ_{1p_1} and λ_{2p_2}, respectively. In the very extreme case, where $\min\{\lambda_{11} + \sigma_1^2, \lambda_{21} + \sigma_2^2\} < \max\{\sigma_1^2, \sigma_2^2\}$, all eigenvectors of one of the blocks would not enter the first p eigenvectors of Σ, because they are indistinguishable from the noise and no corresponding factors would be estimated. Hence, if the data contains heterogeneous blocks, it is more advantageous to estimate the factors in each block separately, since the factors can be estimated consistently in any constellation.
The detection of heterogeneous blocks can be done based on σ_{ji}^2, $j = 1, 2$, $i = 1, \ldots, N_j$. Indeed from $\text{cov}(S_j x_j) = U_j D_j U_j^t + \sigma_j^2 I_{N_j}$ follows

$$\sum_{k=1}^{p_j} \{U_j\}_{ik}^2 \lambda_{jk} = 1 - \sigma_j^2, \quad j = 1, 2, \quad i = 1, \ldots, N_j. \tag{2.3}$$

Since U_j is an orthogonal matrix, its elements $\{U_j\}_{ik} = O(N_j^{-1/2})$, at the same time $\sigma_j = O(1)$, $j = 1, 2$. Now, the eigenvalues of a correlation matrix λ_{jk} can be represented as λ_{jk}^*/sjk^2, where λ_{jk}^* is the k-th eigenvalue of $S_j^{-1}\Lambda_j\Lambda_j^t S_j^{-1}$ (that is, based on x, the data before standardisation) and $s_{jk}^2 = \text{var}(x_{jk})$, $j = 1, 2$, $k = 1, \ldots, p_j$. It is easy to see that, if $s_{1i}^2 \ll s_{2i}^2$ or $s_{2i}^2 \ll s_{1i}^2$ is found for all $i = 1, \ldots, \min\{N_1, N_2\}$, this could indicate that s_{ji}^2 compensates for a different magnitude and/or different decay of λ_j^*, $j = 1, 2$. For example, if $\lambda_{j1}^* = c_1 N_j/p_j$ for some constant c_1, while $\lambda_{ji}^* = o(N_j/p_j^2)$ for $i = 2, \ldots, N_j$, then s_{ji}^2 should be of order $O(p_j^{-1})$ in order equation 2.3 to hold. If all $\lambda_{ji}^* = c_i N_j/p_j$ for some constants c_i, $i = 1, \ldots, N_j$, then $s_{ji}^2 = O(1)$ is needed for equation 2.3 to hold. Hence, a different behaviour of the eigenvalues in corresponding blocks may be identified by a different magnitude of s_{1i}^2 and s_{2i}^2.
In the sample model based on T observations of X_t, the same considerations apply to the sample correlation matrix $\widehat{\Sigma}$.

2.3.3. Algorithm

The practical implementation of the factor analysis on heterogeneous blocks is straightforward. Based on the data matrix $X \in \mathbb{R}^{T \times N}$, sample standard deviations $\hat{s}_1, \ldots, \hat{s}_N$ of each column of X are calculated. Since for large T the sample standard deviations are asymptotically normally distributed, one can apply a stochastic clustering algorithm based on a normality assumption of the data, for example, the one implemented in the R function `Mclust`, see Scrucca et al. (2016). This function also estimates the optimal number of clusters based on a BIC. In principle, the same clustering algorithm can be

applied to sample variances $\hat{s}_1^2, \ldots, \hat{s}_N^2$ as well, but since their distribution in small samples is more skewed than that of the standard deviations, we found that in practice working with standard deviations is more stable. Based on the identified clusters, the matrix X is divided into corresponding blocks and a principal component analysis is run on each block. Of course, it may happen that not all clusters based on standard deviations correspond to different behaviours of the eigenvalues in these clusters, or that the algorithm finds additional clusters. However, taking more clusters than necessary typically does not deteriorate the performance. The choice of the number of principal components to extract from each block can be based, for example, on the criterion suggested in Sobczyk et al. (2017). An implementation of this approach is available from the authors upon request.

2.4. Simulations

For the following simulations, we generate a data matrix consisting of two blocks of variables $X = (X_1^T, X_2^T)^T \in \mathbb{R}^N$ with $x_j \in \mathbb{R}^{N_j}$, $j = 1, 2$ and $N_1 + N_2 = N$, where N is the number of variables contained in X, for each of which we generate T observations, so that X has dimensions $T \times N$. The variables in each block are subject to a factor structure, so that there is a separate set of underlying factors in each block. The factor representation for each block is given by

$$X_i = \Lambda_i f_i + e_i,$$

where index $i = 1, 2$ denotes the respective block of variables, Λ_i the matrix of loadings, f_i the matrix of factors, and e_i the error matrix, that explains the variation in X_i that is not explained by the factor structure. The factors in each block are independent of the factors in the other block, and there is no cross-block dependence of variables on factors, i.e., their joint probability distribution is equal to the product of the respective marginal probability distributions $P(f_1, f_2) = P(f_1)P(f_2)$ and thus also $\text{cov}(f_1, f_2) = 0$, $\text{cov}(X_1, f_2) = 0$ and $\text{cov}(X_2, f_1) = 0$.

In addition, we introduce differently sized errors e_i in the factor specifications between blocks, so that the factors in one block are much more noisy than those in the other block. The errors are drawn from a Gaussian normal distribution with different second moments, such that $e_i \sim N(0, \sigma_i^2)$ and $\sigma_1 < \sigma_2$.

The composite data matrix X consisting of both blocks takes the form

$$X = \begin{bmatrix} f_1 \Lambda_1^T & | & f_2 \Lambda_2^T \end{bmatrix} + \varepsilon,$$

where ε is the error matrix capturing the idiosyncratic errors e_i of both blocks and is given by

$$\varepsilon = \begin{bmatrix} e_1 & 0 \\ 0 & e_2 \end{bmatrix},$$

implying that there is no cross-correlation of errors between blocks. The variance of the

matrix $\varepsilon \sim N(0, \Sigma)$ is given by $\Sigma = \text{blockdiag}\left(\sigma_1^2 I_{N_1}, \sigma_2^2 I_{N_2}\right)$. In addition, we create a target variable denoted $y = \mu_y + \epsilon$, that is driven by the factorial component $\mu_y = 1 + F\beta$, where F is the composite factor matrix $F = \begin{bmatrix} f_1 & | & f_2 \end{bmatrix}$. The coefficient vector, $\beta \sim \text{Poisson}(\lambda_\beta)$, determines the dependence of the factorial component on the factors in F and is drawn from a Poisson distribution. Finally, $\epsilon \sim N(0, 1)$ represents a Gaussian i.i.d. error term that is added to the factorial component in y.

Based on this framework, we develop two different cases. For Case 1 we assume perfect information about the block structure of the data and apply conventional PCA to the blocks separately estimating $\hat{F}$ on a training split to nowcast $\hat{y}$ for test data. In Case 2, we apply our blockPCA algorithm as described in section 2.3.3. We make two changes in order to make Case 2 more similar to the real-world nowcasting scenario. First, we no longer assume perfect information about the block structure of the data. The blockPCA algorithm is designed to detect the blocks of variables as described above. Second, we relax the fixed parameter choice about the number of factors used in the linear regression models and determine the optimal number of factors to use in each iteration via leave-one-out cross-validation.

Each case is run for $M = 500$ iterations generating $T = 300$ observations for each of the $N = 100$ variables. These chosen dimensions depict the case of long data, with the number of observations substantially exceeding the number of variables. We ran the simulations with the opposite dimensions to create wide data as well. The results are so similar that we do not show them in addition to the long data results.

Tables 2.2 and 2.1 summarize the expressions, dimensions and parameter choices used.

2.4.1. Case 1 - separate versus single factor extraction

As described above, for Case 1 we assume perfect information about the block structure of the data. This includes knowledge of which variables belong to which block and the true number of underlying factors to be estimated. In this setting, we compare two different approaches of how to estimate the factor structure before using the estimated factors to nowcast the target variable y. The difference between these two approaches is that we compare the nowcasting results using factors estimated on the composite data matrix X versus using factors estimated on the block matrices X_1 and X_2 separately. Both approaches use conventional PCA to estimate the factors. The estimated factor matrix $\hat{F}$ thus constitutes the matrix of regressors in both approaches. In the block-wise estimation, however, it is a composite matrix consisting of the separately estimated block factor matrices, i.e., $\hat{F} = \begin{bmatrix} \hat{f}_1 & | & \hat{f}_2 \end{bmatrix}$.

The data is divided into a training and a test split, so that the factor structure and coefficients of the OLS nowcasting model are estimated on the first half of the data. In a first step, the factor structure is estimated. Then, the resulting estimated factors are used

Table 2.1: Parameter choices for simulation exercises

Definition	Parameter	Value
Number of repetitions	M	500
Number of observations	T	300
Number of variables	N	100
Number of factors per block	k	20
First moment of distribution loadings are drawn from	μ_Λ	4
Second moment of distribution errors in first block are drawn from	σ_1	1
Second moment of distribution errors in second block are drawn from	σ_2	50
Parameter of Poisson distribution	λ_β	1

as regressors in an OLS nowcasting model of the form

$$y_{\text{train}} = \hat{F}\gamma + \zeta,$$

where γ are the OLS coefficients and ζ is an i.i.d. error term. We conduct the evaluation of the approaches on the training data, imposing the factor structure on the test partition of the matrix X, such that

$$\hat{y}_{\text{test}} = X_{test}\hat{\Lambda}\hat{\gamma},$$

where $\hat{\Lambda}$ represents the matrix of loadings estimated on the training data. Finally, we compare the errors in nowcasting y_{test}, using the mean squared error as the evaluation criterion

$$\text{MSE}_y = \left(y_{\text{test}} - \hat{y}_{\text{test}}\right)^2.$$

Figure 2.A1a illustrates the differences in MSE for the two approaches. We find that estimating the factor structure separately results in better predictions of the outcome variable y, which is driven by those factors.

2.4.2. Case 2 - blockPCA versus PCA

In Case 2, we apply our blockPCA algorithm as described in section 2.3.3. We make two changes in order to make Case 2 more similar to the real-world nowcasting scenario. First, we no longer assume perfect information about the block structure of the data. The blockPCA algorithm is designed to detect the blocks of variables as described above. Second, we relax the fixed parameter choice about the number of factors k used in the linear regression models and determine the optimal number of factors to use in each iteration via leave-one-out cross-validation. Apart from these modifications, the scenario resembles Case 1. Again, we evaluate the mean squared error between the outcome variable in the test period and the predictions from the linear regression models based on the estimated factor structure in the data.

In the case of blockPCA, extracting the factors from the individual blocks adds an additional layer. Thus, an additional loading matrix $\hat{\Lambda}_1$ enters the equation, imposing the factor

Table 2.2: Definitions, expressions and dimensions used in simulation exercises

Definition	Expression	Dimensions
Loadings	$\Lambda_{1,2} \sim N(\mu_\Lambda, 1)$	$N/2 \times k$
Factors	$f_{1,2} \sim N(0,1)$	$T \times k$
Idiosyncratic errors in blocks	$e_{1,2} \sim N(0, \sigma_{1,2})$	$T \times N/2$
Variance of errors	$\Sigma = \text{blockdiag}\left(\sigma_1^2 I_{N_1}, \sigma_2^2 I_{N_2}\right)$	$N \times N$
Error matrix	$\varepsilon \sim N(0, \Sigma)$	$T \times N$
Composite data matrix	$X = \left[f_1 \Lambda_1^T \quad \vert \quad f_2 \Lambda_2^T\right] + \varepsilon$	$T \times N$
Composite factor matrix	$F = \left[f_1 \quad \vert \quad f_2\right]$	$T \times k * 2$
Coefficients of factors	$\beta \sim \text{Poisson}(\lambda_\beta)$	$2 * k \times 1$
Factor driven component of target	$\mu_y = 1 + F\beta$	$T \times 1$
Target variable	$y = \mu_y + \epsilon$	$T \times 1$
Noise in target variable	$\epsilon \sim N(0,1)$	$T \times 1$
Coefficient vector from OLS regression	$\gamma = (\hat{F}^T \hat{F})^{-1} \hat{F}^T y$	$2 * k \times 1$
Error term from OLS regression	$\zeta = y - \hat{F}\gamma$	$T \times 1$

structure from the block layer. In this case, the estimator for β takes the form

$$\hat{\beta}_{\text{bPCA}} = \hat{\Lambda}_1 \hat{\Lambda}_2 \hat{\gamma}_{\hat{F}_2}, \tag{2.4}$$

where $\hat{\Lambda}_2$ are the loadings from the second principal component regression, in which the composite factor matrix consisting of the concatenated factors extracted from each block represents the matrix of regressors. The estimated loading matrix $\hat{\Lambda}_1$ transforms the observed data into this composite factor matrix. The estimated OLS coefficients $\hat{\gamma}_{\hat{F}_2}$ in this case come from regressing y on the the second layer factor matrix $\hat{F}_2$ from the principal component regression.

The nowcasts simply follow as

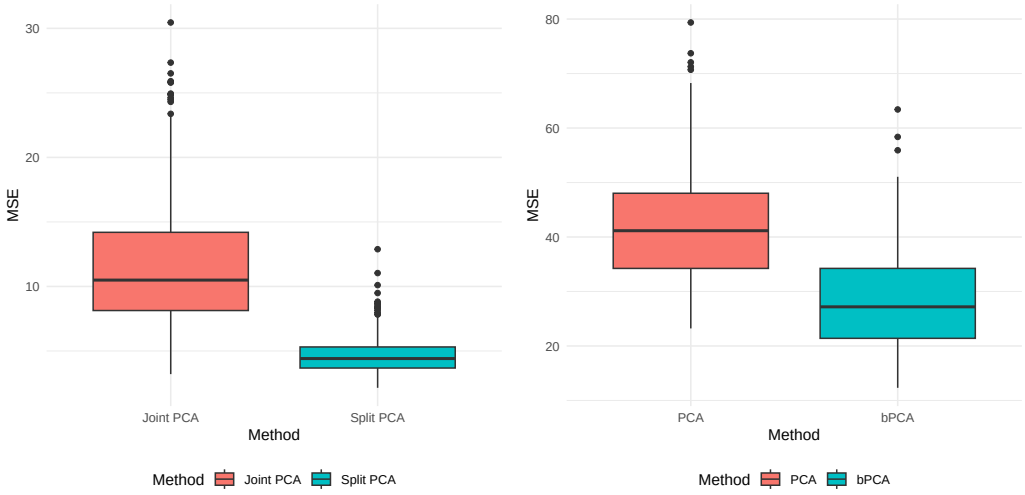
$$\hat{y}_{bPCA} = X_{\text{test}} \hat{\beta}_{\text{bPCA}}.$$

As described in section 2.3.3, several methods are available to optimize the number of factors to extract from each block. While they perform very similarly, we stick to the Bayesian approach described in Sobczyk et al. (2017). Using a maximum-likelihood-approach leads to very similar results.

In addition, the algorithm provides the option to set an upper limit for the number of clusters to be identified. For this simulation, we set the maximum number of clusters to detect to two. Few clusters usually already lead to huge differences in the consistency of factor estimates. In comparison, marginal effect of further clusters is tends to be relatively small.

Figure 2.A1b in the Appendix illustrates additonally the differences in mean squared error between the blockPCA algorithm and conventional principal component regression. Clearly, the blockPCA algorithm handles the heterogeneous blocks of data better in predicting the

Figure 2.1: Simulation results by MSE for Case 1 and 2.



(a) Case 1 (b) Case 2

outcome variable driven by the underlying factors than does the conventional principal component regression.

2.5. Data analysis

To illustrate how the blockPCA algorithm compares to the conventional methods of principal component analysis and partial least squares in real-world macroeconomic nowcasting problems, we apply it to two types of data. Firstly, we make nowcasts using the original monthly macroeconomic time series data from Stock & Watson (2002b) used by Boivin & Ng (2006) to argue that more data does not always improve the results of factor-based now- and forecasting models. Secondly, we compile large-scale monthly macroeconomic data from five emerging economies. To keep the specifications consistent, we select industrial production as the nowcasting target in both cases. The two types of data differ mostly in their dimensions. While the emerging market data is what is typically considered wide data, i.e., the number of variables exceeds the number of observations by far, the Stock & Watson (2002b) data is considered long, with the number of observations substantially exceeding the number of variables. The idea is to put the blockPCA algorithm to the test in both data environments.

For the wide data, we choose emerging economies for the data analysis section of this paper because, compared to developing countries, there is an abundance of data to work with and, compared to advanced economies, there are no established sets of variables known to produce the best nowcasting results. We obtain the data for the analysis from Thomson Reuters Datastream. For Brazil, Chile, India, Malaysia, Mexico and Turkey, we extract all monthly economic time series marked as active at the time of extraction over the period 01.01.2000 to 31.12.2019, as shown in Table 2.3.

Table 2.3: Time frames, number of observations and variables of country data sets.

Country	From	To	Observations	Variables
Data from Stock & Watson (2002b) (long data)				
USA	July 1959	February 1999	238	147
Data from emerging market economies (wide data)				
Brazil	July 2000	September 2019	231	1845
Chile	July 2000	September 2019	231	552
India	July 2000	July 2019	229	769
Malaysia	July 2000	March 2019	225	1144
Turkey	July 2000	October 2018	220	1888

2.5.1. Preprocessing

At the edges, the emerging market economy datasets are patchy, i.e., the first and last observations for many variables are subject to a high degree of missingness. Our initial data filtering process excludes variables with missing observations, so we would lose a significant amount of variables if we were to prioritize a longer time span over trimming the edges to preserve variables. Thus, we shift the start and end points of the datasets individually to preserve variables in exchange for a few observations. Moreover, we exclude variables with a standard deviation close to zero, because they do not represent time dynamics. Finally, some of the variables contain many null values, which can lead to computational problems and are therefore also filtered out.

Since we use an ordinary least squares estimator for nowcasting, the assumption of identically and independently distributed error terms is crucial. This requires that all time series entering the model are stationary. Using the KPSS (Kwiatkowski et al., 1992) and Augmented-Dickey-Fuller (Dickey & Fuller, 1979) tests, we individually determine the order of integration for all variables entering the model and apply an appropriate number of first differences to ensure stationarity.

Lastly, in the case of the emerging market datasets, we select a monthly measure of industrial production as the nowcasting target and filter out all explanatory variables that either contain "industrial production" in the variable description and or are highly correlated (> 0.95) with the target variable. This step is necessary, because some datasets contain several similar measures for industrial production, which would dominate the estimation and thus weaken the comparability of the factor estimation as measured by the mean squared error in predicting the target variable. In the case of the Stock & Watson (2002b) data, we exclude all other variables depicting some form of industrial production, following the same reasoning.

2.5.2. Nowcasting

We employ an ordinary least squares model in which the estimated factors are the only explanatory variables following the structure outlined in section 2.2. Estimations are run

independently for all countries, and the available data for each country is split in half to create a training period and a test period. We estimate the factor structure and OLS coefficients on the training data and apply them to the test data to create predictions for the target variable.

More specifically, the factor structure estimated on the training data is imposed on the test data by multiplying with the estimated loadings and OLS coefficients. The number of estimated factors used as explanatory variables is determined by leave-one-out cross-validation on the training data, minimizing the root mean squared error. The estimation procedure is the same as the the one used for the simulations outlined in section 2.4.

We fix the number of blocks to two for the scope of the analysis. While we observe large improvements in nowcasting accuracy applying the blockPCA algorithm compared to conventional methods already with two blocks, the blockPCA induced improvements are not very sensitive to using more blocks. Major improvements are made by imposing the block structure, the difference between two and three or four blocks is often not significant.

In addition to comparing the performance of our proposed blockPCA algorithm with conventional principal component regression, we add the method of partial least squares as a benchmark. Partial least squares is a related method that extracts the factors maximizing the covariance between the target variable y and the set of explanatory variables in X. In other words, the extracted factors are estimated in such a way as to explain as much variation as possible in the target variable y (see, e.g., Garthwaite 1994). Even though the factor estimation with partial least squares should be similarly affected by the factor domination problem outlined above, we are interested in investigating potential differences between conventional principal component regression and partial least squares in coping with large noisy datasets.

2.5.3. Results

We evaluate the nowcast accuracy using two metrics. In addition to calculating the mean squared error for all three competing methods for each point-nowcast, we calculate the correlation between the series of nowcasts and the true series over the entire test period. Figures 2.A2 to 2.A7 show the results in terms of mean squared error and correlation, the eigenvalues by block, and the target variable over time alongside the three competing nowcasts over time for each country sample individually.

Figure 2.A2 comprises the results for the long Stock & Watson (2002b) data. We observe a slight improvement across both evaluation metrics, with slightly higher correlation between the series of nowcasts and the series of testing data and a slightly lower mean squared nowcasting error in comparison to conventional PCA and PLS.

The emerging differences between blockPCA, PCA and PLS are more significant for the wide data. Across all five samples, bPCA produces more accurate nowcasts in terms of mean squared error than PCR and PLS. Improvements in correlation are less significant,

however the bPCA generated nowcasts are at least slightly more correlated with the true target series than PCA and PLS in all cases. The time series plots in Figures 2.A3 to 2.A7 indicate that PCA and PLS seem to generally create nowcasts with the correct sign, but the amplitudes tend to be off compared to bPCA. We hypothesize that this is due to the dominance of certain factors, causing the omittance of relevant factors better accounted for by bPCA.

The differences in results between the types of data we observe can be seen as evidence for the assumption that the wider the data, the more likely relevant factors end up being omitted.

2.6. Discussion

This paper addresses the observation made in Boivin & Ng (2006) that relevant factors can be dominated by other factors in a large data principal component regression setting. Over the past two decades, numerous attempts have been made to deal with the challenges of very wide data sets, where the number of variables substantially exceeds the number of observations. While many approaches involve reducing the number of variables in advance, we propose an approach that is much more robust to the use of large datasets and prevents factors from being dominated. We derive this approach from the theoretical background we provide and present the blockPCA algorithm, which identifies different groups of variables and extracts a first set of factors from these groups separately in a first step. In a second step, the resulting factors are concatenated and a second set of factors is extracted from the resulting matrix. This second set of factors is then be used as regressors in the canonical factor-based now- and forecasting model described in Stock & Watson (2002a,b). We argue that this approach strongly mitigates the omission of relevant factors caused by noise level differences between groups of variables.

To demonstrate how the blockwise factor extraction improves the estimation of the true underlying factors under heteroskedastic idiosyncratic errors, we run a set of simulations imposing the corresponding error structure on the simulated data. Applying our algorithm to long and wide real-world data, we show that it handles large-scale macroeconomic panel data better than the conventional PCA and PLS algorithms. We observe major improvements using very wide data, i.e., data for which the number of variables substantially exceeds the number of observations, and minor improvements for long data, where the number of observations substantially exceeds the number of variables, with the number of variables being relatively small also in absolute terms. Our proposed algorithm has the advantage of being easy to use and not requiring preselection of variables, while being much more robust to the problem of relevant factors being dominated than conventional methods.

In future research the blockPCA algorithm can be applied and further tested in settings, where conventional principal component analysis could encounter the problems laid out here. This refers to canonical principal component regression models, but also to other

contexts in which it is of interest to account for the problem of factor omission.

2.A. Appendix

2.A.1. Simulations

Case 1.A - comparison of factor estimation approaches using matrix norm

In Case 1.A, we point out that conventional principal component analysis applied to the data matrix recovers the factors less consistently than principal component analysis applied to each block of variables separately. We compare the factors estimated by PCA applied to the joint data and the factors estimated by PCA applied to each block separately to the true underlying factors using a matrix norm, that evaluates the distance between the true factors and the factor estimates.

We generate a data matrix X consisting of two blocks of variables, such that $X = (X_1^T, X_2^T)^T \in \mathbb{R}^N$ with $x_j \in \mathbb{R}^{N_j}$, $j = 1, 2$ and $N_1 + N_2 = N$, where N is the number of variables contained in X, for each of which we generate T observations, so that X has dimensions $T \times N$. Each block is subject to the typical factor structure, and otherwise we stay in within the framework outlined in section 2.3. With each iteration, new factors and errors are generated, and the factors are estimated using the two competing approaches. The matrix norm we use is called the Frobenius norm, which we apply to matrix A, where A is of the form

$$A = F - \hat{F}\left(\hat{F}^T\hat{F}\right)^{-1}\hat{F}F.$$

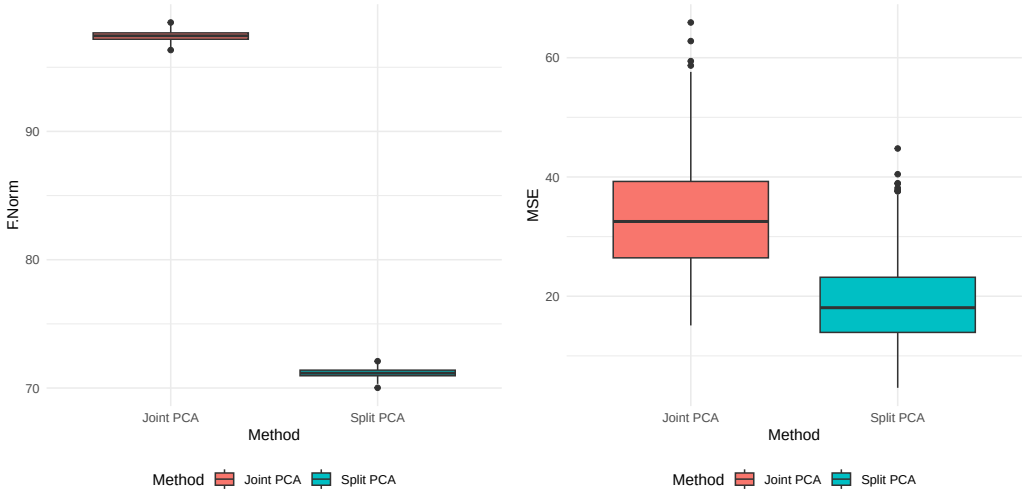
The matrix A captures the difference between the true and estimated factors. The smaller the Frobenius norm applied to matrix A, the better the underlying factors have been recovered. The parameter choices are given in Table 2.1. The two blocks of variables differ only in the amount of noise that is added to the underlying factor component by setting $\sigma_1 < \sigma_2$. Figure 2.A1a illustrates the Frobenius norm results over $M = 500$ iterations by approach.

Clearly, applying PCA to both blocks separately and then concatenating the factors from each block into a joint factor matrix provides more consistent estimates of the true underlying factors than applying PCA to the data matrix at once.

Case 2.A - comparison of factor estimation approaches estimating factorial component

Case 2.A extends Case 1 by introducing a regression setting, that includes the outcome variable y. Instead of estimating the underlying factors, we estimate the the factorial component of the outcome variable denoted μ_y, as defined in Table 2.2. This factorial component μ_y depends on the factor matrix F as determined by the coefficient vector β, whose coefficients are drawn from a Poisson distribution at each iteration. We evaluate both approaches by calculating the mean squared error in estimating μ_y via y, which, as

Figure 2.A1: Simulation results by MSE for Case 1.A and Case 2.A



(a) Case 1 (b) Case 2

shown in Table 2.2 is the sum of the factorial component and an i.i.d. error term. The evaluation criterion is thus

$$\text{MSE}_{\mu_y} = (\mu_y - \hat{y})^2 .$$

Figure 2.A1b illustrates the results in terms of mean squared error. Estimating the factors separately before using them as explanatory variables in a linear regression model with outcome y leads to a more consistent estimation of the factorial component μ_y.

2.A.2. Nowcasting

Figure 2.A2: USA: Results for nowcasting industrial production with data from Boivin & Ng (2006), originally from Stock & Watson (2002b), by MSE, correlation, and eigenvalue decay by block.

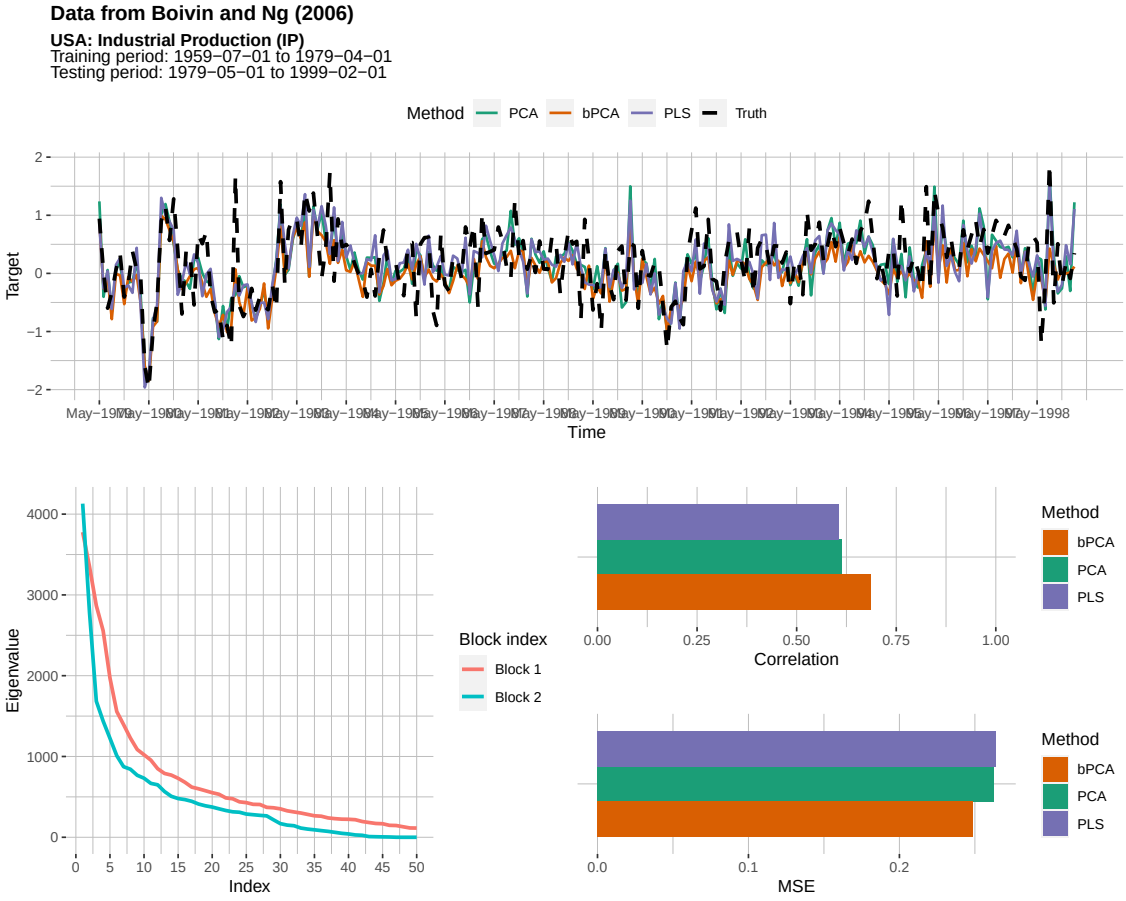


Figure 2.A3: Brazil: Results for nowcasting industrial production with data from Thomson Reuters Datastream, by MSE, correlation, and eigenvalue decay by block.

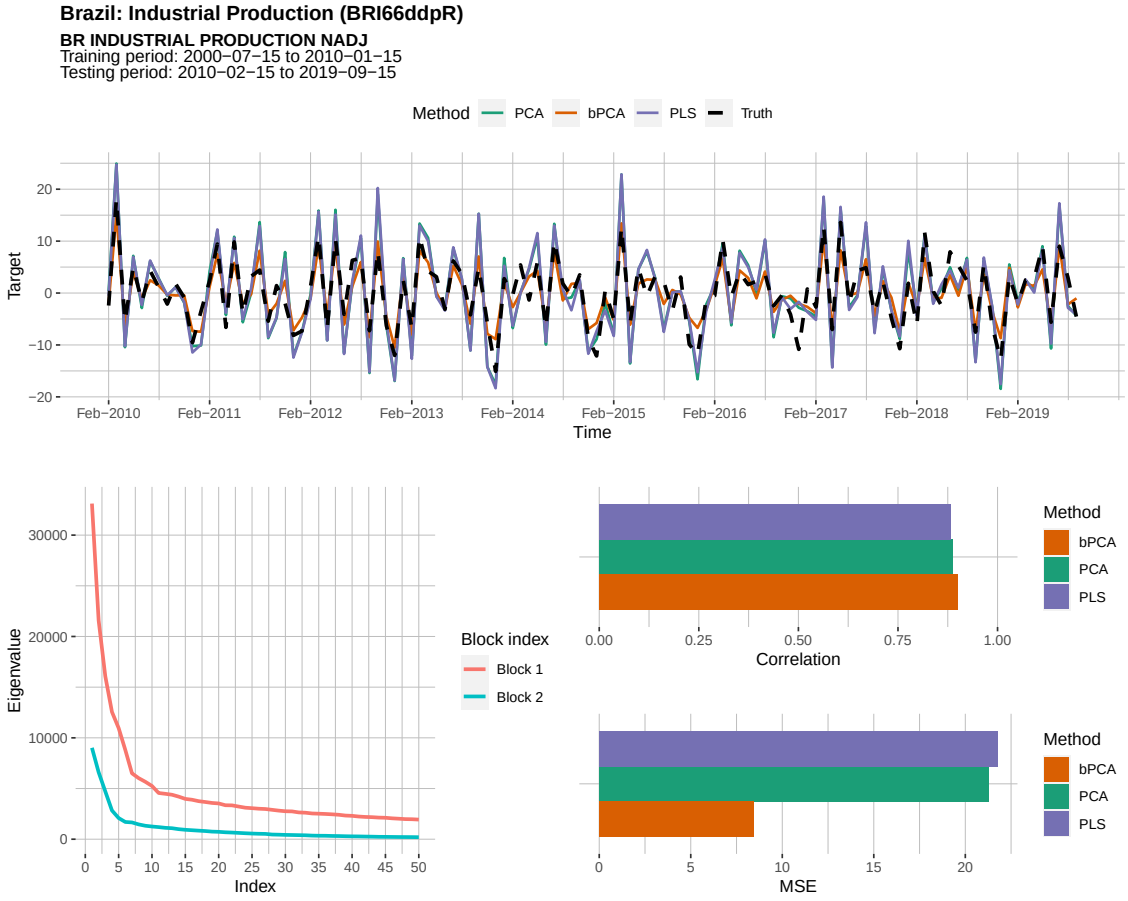


Figure 2.A4: Chile: Results for nowcasting industrial production with data from Thomson Reuters Datastream, by MSE, correlation, and eigenvalue decay by block.

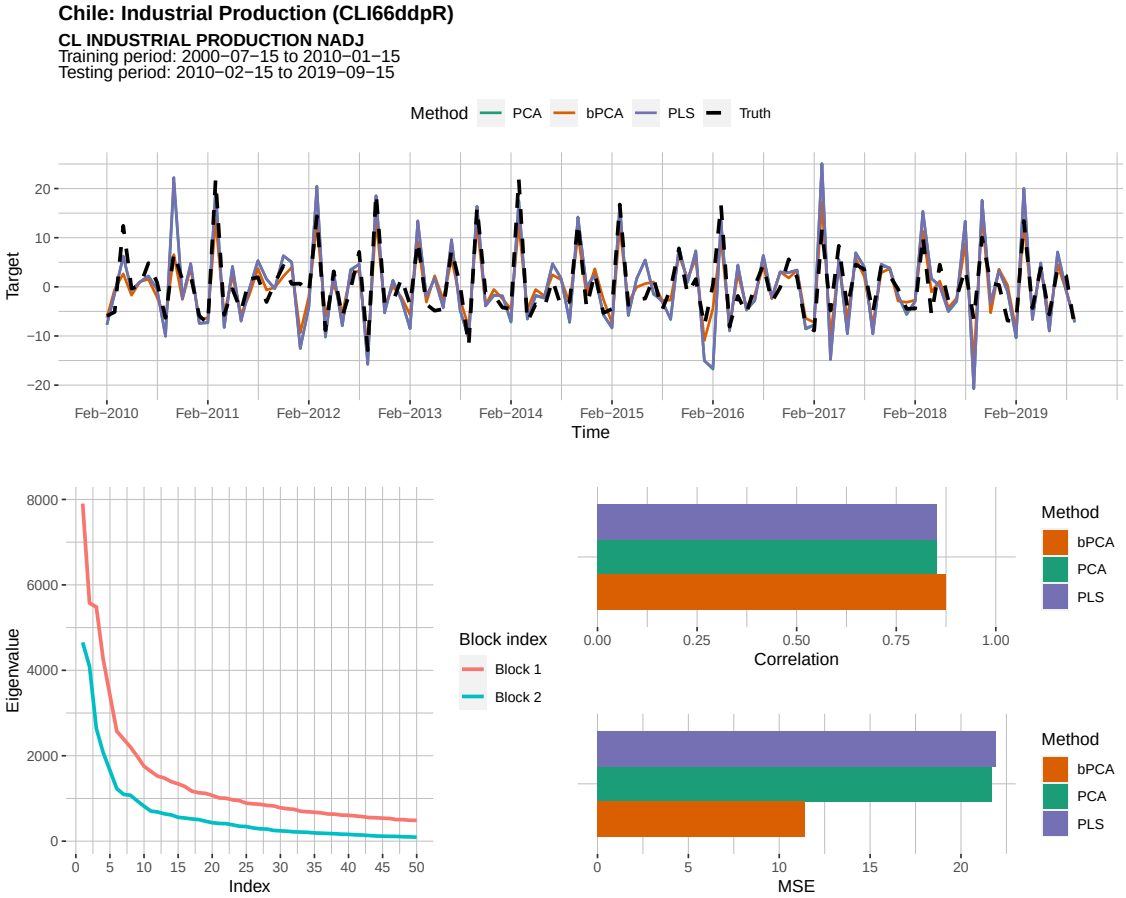


Figure 2.A5: India: Results for nowcasting industrial production with data from Thomson Reuters Datastream, by MSE, correlation, and eigenvalue decay by block.

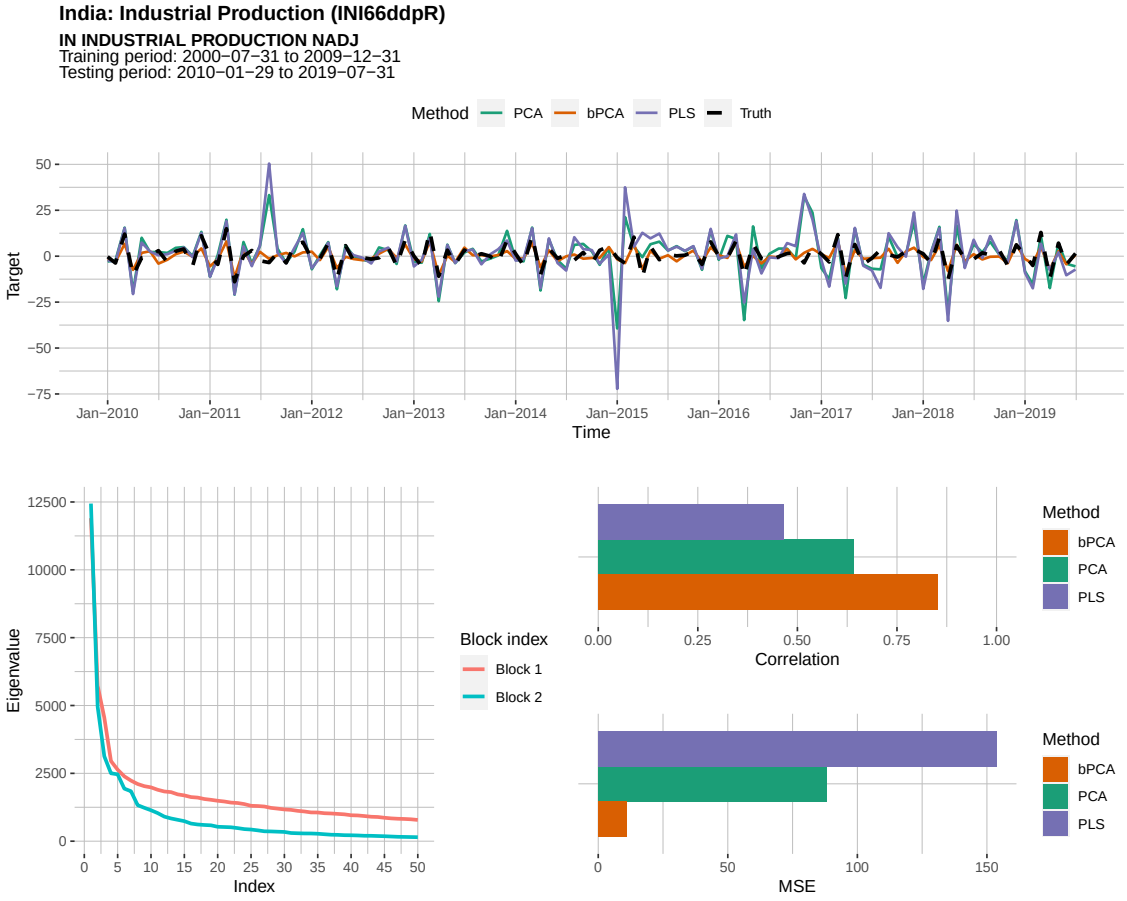


Figure 2.A6: Malaysia: Results for nowcasting industrial production with data from Thomson Reuters Datastream, by MSE, correlation, and eigenvalue decay by block.

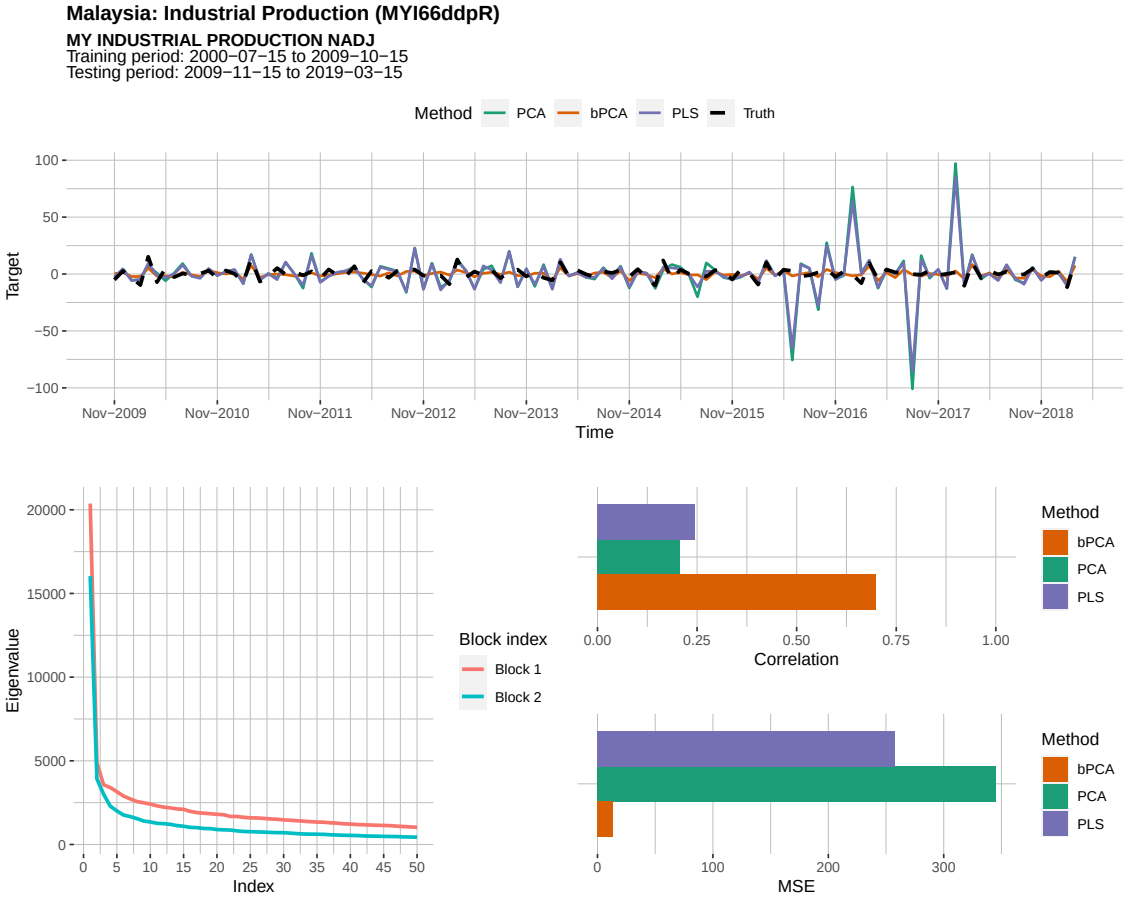
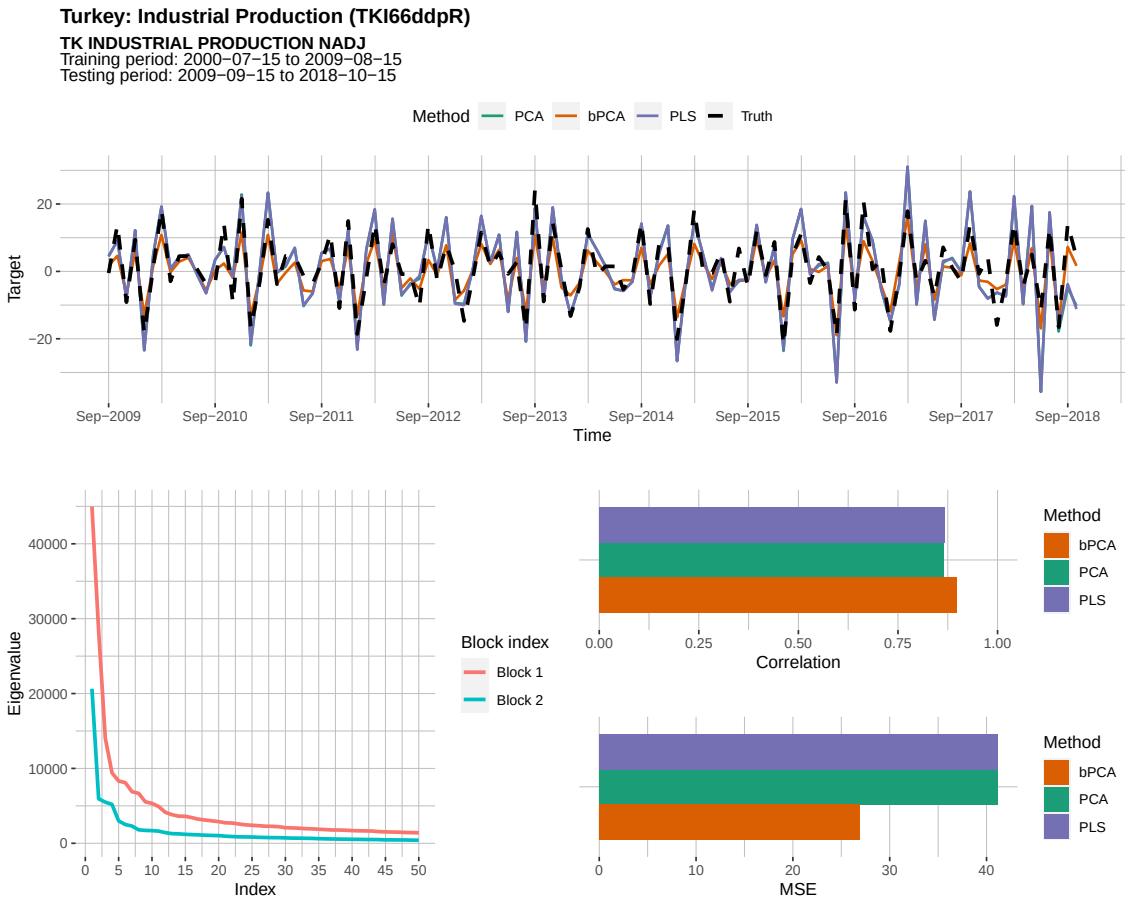


Figure 2.A7: Turkey: Results for nowcasting industrial production with data from Thomson Reuters Datastream, by MSE, correlation, and eigenvalue decay by block.



Chapter 3

Taking over the World? Automation and Market Power

Abstract

This paper studies how automation technology affects market power in the global economy. We develop a theoretical model in which firms' markups are endogenous to factor input choices based on technology levels, but are also affected by technology adoption of other domestic and foreign firms. In an empirical analysis, we find that market power, measured as the markup of price over marginal cost, declines on average with higher levels of automation. However, there is substantial heterogeneity, with firms in the highest revenue and markup quintile gaining market power. Moreover, we find that exposure to foreign automation increases competition in the local market.

This chapter is joint work with Henry Stemmler (University of Göttingen) and Florian Unger (University of Göttingen). We are thankful for valuable comments and feedback from Holger Strulik, Gianmarco Ottaviano, Joel Stiebale, Katharina Erhardt, Krisztina Kis-Katos and participants of the GlaD Seminar at the University of Göttingen.

3.1. Introduction

In recent years, there has been a surge of interest in three related phenomena in international economics: The declining share of labor, the accelerating concentration of market power, and the increasing use of automation technology. Although a link between these phenomena has been established (Karabarbounis & Neiman, 2014; Berg et al., 2021), the exact interrelationships remain a matter of ongoing debate and are yet to be better understood (Grossman & Oberfield, 2021).

In this paper, we investigate whether automation technology contributes to the rise of market power in the form of markups. Specifically, we develop a theoretical model of oligopolistic competition in which firms' markups are endogenous to factor input choices, total factor productivity and the competitive environment created by other domestic and foreign firms. We test the model empirically, distinguishing between domestic robot adoption and exposure to robots in foreign economies. Our empirical analysis reveals considerable heterogeneity across firms. We find that firms in the highest markup quintile further increase their markups and market shares through sectoral robot adoption, while firms in lower quintiles suffer losses in terms of markups and market shares.

The recent literature on the evolution of global market power and hence market concentration has been largely influenced by De Loecker & Warzynski (2012), both methodologically and descriptively. In this paper the authors introduced an innovative method for estimating firms' markups based on a control function approach, which led to a large number of subsequent publications building on this methodology. For example De Loecker & Eeckhout (2018) and Diez et al. (2019) document a global rise in markups, which they mostly attribute to a reallocation of market shares from low to high markup firms. In this strand of literature, firms' markups are assumed to be proportional to firms' market shares, so that the documented rise in markups implies increasing market concentration.

Corroborating the notion that fewer firms are increasingly dominating markets, D. Autor et al. (2020) coined the term "superstar firms", to describe how high-tech firms excel in a "winner takes all" economy. In related work, D. Autor & Salomons (2018) and Dorn et al. (2017) link this to the labor share debate, arguing that the emergence of technology, and hence capital-intensive superstar firms has played a crucial role in the decline of the labor share. While most "superstar" firms have been documented in the digital, IT and service sectors, benefiting from platform economies (Lashkari et al., 2018; D. Autor et al., 2020), similar but somewhat weaker trends have also been observed for technological leaders in manufacturing (Andrews et al., 2016; Stiebale et al., 2020).

Advances in industrial robot technology and subsequent commercialization have led to a steady increase in uptake over the past three decades (International Federation of Robotics, 2018). Several dimensions of robot adoption and its consequences have been studied in recent years. Dinlersoz & Wolf (2018) and Koch et al. (2019) show that the most affluent and technologically advanced manufacturers pioneer the adoption of industrial robots in

manufacturing. A number of papers find that robot adoption at the firm level increases sales and employment, although it typically reduces the labor share (Humlum, 2019; Acemoglu et al., 2020; Aghion et al., 2020; Bonfiglioli et al., 2020). At the more aggregate labor market level, Acemoglu & Restrepo (2020) argue that job displacement rather than job creation effects are the predominant consequence of robot adoption in the US economy. Dauth et al. (2021) document that more robot-exposed labor markets in Germany experience declines in manufacturing employment, but these are offset by increasing employment in services.

However, little research has been done on the impact of robot adoption on market power. In recent work closely related to this paper, Stiebale et al. (2020) investigate the existence of European superstar firms in manufacturing. In line with our results, they report within-sector heterogeneity across firms in the effect of robot adoption on markups. Our work adds to the literature by confirming the findings of Stiebale et al. (2020) using a different international firm dataset, providing a theoretical model, and extending the scope of the analysis to robot adoption by foreign firms.

Our theoretical framework builds on a model of oligopolistic competition from Edmond et al. (2015), which we adapt to the objective of our analysis. Specifically, we introduce a Cobb-Douglas production technology in which industrial robots serve as an input alongside labor to intermediate good producing firms. Moreover, we allow output elasticities to vary at the firm level, so that firms operate with different labor and robot intensities, similar to Harrigan & Reshef (2015). The model predicts that firms operating with above-average robot intensities benefit from a reduction in the robot rental rate in terms of market shares and markups at the cost of firms with below-average robot-intensity in the one-country economy. Furthermore, the model predicts that robot adoption by foreign competitors exerts downward pressure on the market shares and markups of all domestic firms in a two-country economy.

We combine publicly available firm-level balance sheet data, used to estimate average sectoral markups, with data on industry-level robot uptake from the International Federation of Robotics (2018) on 29 countries and 20 sectors between 1995 and 2015 for the empirical analysis. We employ an instrumental variable (IV) approach to account for endogenous uptake of robots within sectors. Our empirical results suggest that increased automation is associated with higher markups and larger market shares for the most productive quintile of firms in our sample. Analogously, we find that firms in the lower quintiles suffer losses in market shares and markups as a consequence of increased automation. Taken together, these findings reconcile the notions that average markups in manufacturing have not increased much over the past years and that automation technologies increase profits for some firms. We take this as evidence for the hypothesis that the increasing use of industrial robots amplifies market concentration and makes only a few firms better off. Moreover, we find that the adoption of robots by foreign competitors exerts downward pressure on all local firms' markups and market shares.

The paper is structured as follows. In section 3.2, we develop the one-country and two-

countries economy versions of the model, and derive the model's hypotheses about the effect of a change in the robot rental rate on markups and market shares. In section 3.3 we present our empirical strategy and in section 3.4 we present all the relevant results. We conclude the analysis in section 3.5.

3.2. Theory

In the following, we derive a theoretical model to motivate our analysis of the effects of increasing industrial robot adoption on the distribution of firm-level markups. In order to obtain a framework that allows the derivation of hypotheses about the interplay between robot adoption and markups, we combine a number of assumptions.

First, we assume that firms use industrial robots alongside labor as an input to production. We also assume that firms differ in the intensity with which they use factor inputs, i.e., we allow for firm-level heterogeneity in output elasticities. This assumption builds on the findings of previous work by, for example, Koch et al. (2019), who report firm-level heterogeneity in the adoption of industrial robots across but also within sectors.

Second, we assume that firms differ in terms of total factor productivity (TFP), as is common in the literature (see, for example, Melitz 2003). In sum, firms are thus subject to two sources of heterogeneity, which they obtain by drawing from probability distribution functions. A joint distribution function of the two technology parameters allows for correlation between the two, so that, for example, a high level of robot intensity is more likely to be drawn alongside a high level of TFP than a low level of TFP, as in Harrigan & Reshef (2015). While Koch et al. (2019) find a positive association between firm productivity and robot intensity, our data are insufficient to calibrate such a joint distribution function. Therefore, we refrain from calibrating the model and instead derive purely theoretical results allowing for different technology parameterizations.

Third, we assume that markups vary at the firm-level and are endogenous to a firm's competitiveness, which is determined by its technology relative to that of its competitors. Thus, a firm's robot intensity, which depends on its technology draw, is one of the determinants of its markup. To provide a theoretical framework that allows for the combination of these assumptions, we adapt the model in Edmond, Midrigan, & Xu (2015) (hereafter EMX model), which is a model of oligopolistic competition based on the Atkeson & Burstein (2008) model. Although it was originally designed as a trade model, we first simplify the model to a one-country economy version in order to derive the effect of decreasing robot prices on markups without interference from foreign competitors or trade effects. In this setting, we show how a reduction in the robot rental rate makes firms with above-average robot intensity better off in terms of market shares and markups.

We then extend the model to the two-country case and show how additional competition via trade aggravates this polarizing effect. Due to fixed costs of trade, only firms with high productivity and robot intensity choose to export. Thus, a reduction in the price

of robots increases the average productivity and robot intensity in the export market. Firms that would have been on the margin of benefiting from the robot price reduction in the one-country economy are crowded out by foreign high-robot-intensity firms in the two-country case.

3.2.1. Small open economy: domestic competition

We model a two-stage economy, in which heterogeneous intermediate good producers provide inputs to homogeneous final good producers. While intermediate good producers operate under oligopolistic competition, final good producers operate under perfect competition. Consumers purchase the homogeneous final good and supply labor to the economy.

Final Good Producers

In the final good stage firms produce a homogeneous final good denoted Y under perfect competition

$$Y = \left(\int_0^1 y\left(s\right)^{\frac{\sigma-1}{\sigma}} ds \right)^{\frac{\sigma}{\sigma-1}}, \tag{3.1}$$

where $\sigma > 1$ is the elasticity of substitution across a continuum of sectors $s \in [0, 1]$ from which inputs $y(s)$ are sourced. Consumers buy the final good at price P, which is the price index for the final good and given by

$$P = \left(\int_0^1 p\left(s\right)^{1-\sigma} ds \right)^{\frac{1}{1-\sigma}}, \tag{3.2}$$

where $p(s)$ is a sector specific price index defined below in Equation 3.5.

Intermediate Good Producers

The number of intermediate good producers is finite and assumed to be exogenous, as in the benchmark EMX model. Intermediate good producers use Cobb-Douglas production technology, where labor L and robots R are the only inputs. In addition, intermediate producers are subject to two sources of heterogeneity, which are imposed by draws from a joint distribution function. These two draws determine the total overall factor productivity of intermediate producers φ_i, as well as their output elasticity for labor in production θ_i, where the subscript i denotes the intermediate good producing firm. The joint distribution function is denoted as $g(\varphi_i, \theta_i)$ as in Harrigan & Reshef (2015). Assuming constant returns to scale, the draw of θ_i entails the output elasticity for robots, which follows as $1 - \theta_i$. This firm-level variation in output elasticities implies that producers of intermediate goods operate with different factor intensities, i.e., different factor input ratios. Their production

technology for output in a given sector s takes the form

$$y_i(s) = \varphi_i(s) L_i(s)^{\theta_i} R_i(s)^{1-\theta_i}, \tag{3.3}$$

where firm-specific input of labor and robots in sector s are denoted $L_i(s)$ and $R_i(s)$ respectively.

In the interest of parsimony, we do not include conventional, non-automation-related capital, typically denoted K, in the production function. We assume that automation capital, here represented as robots R, differs from conventional capital conceptually in that it comprises capital directly linked to automation technology and no other forms of capital. Moreover, we hypothesize that it is also different from conventional capital in its degree of usage across firms. Strictly speaking, we assume that there is a difference in the underlying empirical distributions of the respective output elasticities, with the use of automation related capital being more heterogeneous across firms than the use of conventional capital. While we argue that a firm can be operational even with virtually no use of automation capital, we consider the use of conventional capital to be less variable. However, firm-level data would be required to estimate the corresponding output elasticities to verify these assumptions by interpreting the means and variances of the estimated underlying distributions. As data availability steadily increases, we expect such data to become available in the future so that we will then be able to calibrate the model we present here, including conventional capital. For the scope of this work, we argue that its inclusion in the production function would not alter the core predictions of our model regarding market concentration. We thus decide to keep the production function as simple as possible for deriving our hypotheses of interest. Nevertheless, extending the model to include non-automation-related capital in the production function would be a natural extension and of interest for future calibration. Following a similar reasoning, we make the simplifying assumption that the sum of the output elasticities equals one and that we are thus in the classical Cobb-Douglas scenario with constant returns to scale. Future empirical research must show whether this assumption should be relaxed in order for the derived hypotheses to match empirical observations as closely as possible. A deviation from the assumption of constant returns to scales at this point would add another layer of complexity not clearly being warranted by theoretical arguments nor the current body of evidence.

There is an ongoing debate in the literature as to whether automation has a positive or negative effect on labor demand. While the potential channels for both, job displacement and job creation effects, have been described in detail, evidence to which ultimately dominates is mixed. Moreover, apart from the labor demand effects observed at the level of the automating firm, the resulting industry-level changes may differ in a general equilibrium setting. In the context of modeling firm-level production, however, it has probably been more common to assume that technology-related capital and labor function as substitutes. We depart from this view building on the evidence from Aghion et al. (2020, 2022), and hence assume that robots and workers are complementary in the production of intermediate goods.

Demand for Intermediate Goods. Since the demand for intermediate goods in our one-country economy version is equivalent to the demand for intermediate goods on the home market in the EMX model, we keep the derivation thereof brief. It is derived from the final good producer's profit-maximization problem.[1]

Demand for the intermediate good produced by firm i in sector s is given by

$$y_i(s) = \left(\frac{p_i(s)}{p(s)}\right)^{-\gamma} \left(\frac{p(s)}{P}\right)^{-\sigma} Y, \tag{3.4}$$

where $p(s)$ is the intermediate good price index for any given sector s and γ depicts the within-sector-elasticity of substitution, which is assumed to be larger than the cross-sector-elasticity of substitution, so that $\gamma > \sigma$. Equation 3.4 implies that the more competitive a firm is within its sector, the larger its share of aggregate demand Y will be. A firm's competitiveness is determined by its marginal cost advantage over its competitors, which results from its technology draws. The lower a firm's marginal cost, the more pricing power it has and the greater its potential to gain market share. Analogously, the more competitive the sector in which the firm operates is relative to other sectors, the larger that firm's share of aggregate demand Y will be.

The sectoral price index is based on the prices of active firms in a given sector and the within-sector-elasticity γ and is defined as

$$p(s) = \left(\sum_{i=1}^{n(s)} p_i(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}}. \tag{3.5}$$

Market Structure. We impose Bertrand competition on the intermediate goods market. The choice between Cournot and Bertrand competition mainly affects the derivation of the demand elasticity that firms face. Since Edmond et al. (2015) show that Cournot and Bertrand lead to similar results in the EMX framework, we do not derive the results for Cournot competition.

Profit Maximization of Intermediate Good Producers. In the interest of parsimony, we do not introduce fixed operating costs. Intermediate good producers therefore maximize profits via

$$\pi_i(s) = \max_{p_i(s), L_i(s), R_i(s)} \left[p_i(s) y_i(s) - w L_i(s) - r R_i(s)\right], \tag{3.6}$$

where $p_i(s)$ is the price intermediate producer i charges, w denotes the wage rate, i.e. the cost of labor, and r denotes the robot rental rate. Indirect demand for goods produced by

[1] See eq. A1 in appendix 3.A.1.

firm i follows from equation 3.4 and takes the form

$$p_i(s) = y_i(s)^{-\frac{1}{\gamma}} p(s) \left(\frac{p(s)}{P}\right)^{-\frac{\sigma}{\gamma}} Y^{\frac{1}{\gamma}}. \tag{3.7}$$

By plugging indirect demand into the intermediate producers' profit maximization problem (equation 3.6) we can derive the respective profit-maximizing factor demands using first order conditions. Profit-maximizing demand for labor, $L_i^*(s)$ and for robots $R_i^*(s)$ take the form

$$L_i^*(s) = \frac{y_i(s)}{\varphi_i(s)} \left(\frac{1-\theta_i}{\theta_i}\frac{w}{r}\right)^{-(1-\theta_i)}, \tag{3.8}$$

$$R_i^*(s) = \frac{y_i(s)}{\varphi_i(s)} \left(\frac{1-\theta_i}{\theta_i}\frac{w}{r}\right)^{\theta_i}. \tag{3.9}$$

An intermediate good producers' profit-maximizing price is obtained by plugging the profit-maximizing factor demands into the profit-maximization problem given by equation 3.6 and deriving with respect to the price p_i, which gives

$$p_i(s) = \frac{\epsilon_i(s)}{\epsilon_i(s) - 1} \frac{V_i}{\varphi_i(s)}, \tag{3.10}$$

where an intermediate producing firm's marginal costs are defined as

$$V_i = w^{\theta_i} r^{1-\theta_i} \theta_i^{-\theta_i} (1-\theta_i)^{\theta_i - 1}. \tag{3.11}$$

We denote the demand elasticity intermediate producer i faces with ϵ_i. In line with the EMX Bertrand model, the demand elasticity depends on the underlying within-sector-elasticity of substitution γ and across-sector-elasticity of substitution σ in the form

$$\epsilon_i(s) = \gamma(1 - \omega_i(s)) + \sigma\omega_i(s), \tag{3.12}$$

where $\omega_i(s)$ denotes an intermediate producing firm's sectoral market share and is defined as

$$\omega_i(s) = \left(\frac{p_i(s)}{p(s)}\right)^{1-\gamma}. \tag{3.13}$$

An intermediate producer's market share is thus determined by its profit-maximizing price relative to the price index of its sector. Consequently, a reduction in the profit-maximizing price $p_i(s)$ is generally associated with an increase in market share $\omega_i(s)$.

Markups. An intermediate good producing firm's markup is a function of its demand

elasticity and given by

$$\mu_i(s) = \frac{\epsilon_i(s)}{\epsilon_i(s) - 1}. \tag{3.14}$$

Hence, the lower the demand elasticity faced by an intermediate producer, the higher its markup. Accordingly, the higher a firm's market share, the lower the demand elasticity it faces and thus the higher its markup.

Market Clearing

Markets clear according to the factor shares in the economy. Aggregate demands for labor and robots take the form

$$L = \int_0^1 \left(\sum_{i=1}^{n(s)} L_i^*(s) \right) ds = \bar{\theta} Y, \tag{3.15}$$

$$R = \int_0^1 \left(\sum_{i=1}^{n(s)} R_i^*(s) \right) ds = \left(1 - \bar{\theta}\right) Y, \tag{3.16}$$

where $\bar{\theta}$ is the average draw of the output elasticity for labor. We assume that labor supply is perfectly elastic, so that changes in the demand for labor are reflected in changes in the wage w. In the case of robots, we assume that they are not produced domestically, but are imported from a foreign economy in exchange for the final good produced in the domestic economy. The production of robots is thus exogenous to the domestic economy and not modelled explicitly. We consider them to be inputs to production that fully depreciate each period, so that the robot rental rate equals the price of robots in exchange for final goods. Similarly to the classical setting of a small open economy, we assume that demand from the domestic economy does not affect the price for robots, but that it is determined on the world market. The assumption of inelastic robot supply implies that aggregate demand for robots as given by equation 3.16 is therefore met by foreign supply without affecting the world market price for robots.
In related empirical work, Duch-Brown & Haarburger (2023) investigate the development of market concentration for the world market of industrial robots. They find that a few robot exporting countries provide the majority of world robot supply. The economy modelled here can be seen as a small economy sourcing robot supply from one of these large-scale exporters.

3.2.2. Reduction in the robot rental rate: only domestic competition

A reduction in the robot rental rate r, directly affects firms' marginal costs and profit-maximizing demands for labor and robots. In response, *both* firms' profit-maximizing prices and sectoral price indexes change, which affects market shares and markups. We are

interested in identifying which firms gain market share and markups and which firms do not. To derive this result, we construct a set of robot price elasticities, that allow us to trace the effect of a change in the robot rental rate.

Effect on marginal costs. Due to the output elasticity of labor being constrained by $0 < \theta < 1$, all firms use both factor inputs in production. The direct effect of a reduction in the price of robots r is therefore a reduction in the firm's marginal cost as defined in eq. (3.11). Using the differential of the marginal cost equation, we can solve for the elasticity of a firm's marginal costs with respect to the robot rental rate

$$\frac{d\ln V_i}{d\ln r} = \theta_i \frac{d\ln w}{d\ln r} + (1-\theta_i)\,. \tag{3.17}$$

We interpret the two terms on the right-hand side of equation 3.17 as the direct and indirect marginal cost effects induced by robot price changes. The higher a firm's robot intensity in production, i.e. the smaller θ_i, the larger is the direct effect $(1-\theta_i)$ on a firm's marginal cost in response to changes in the robot rental rate. The indirect effect $(\theta_i \frac{d\ln w}{d\ln r})$ represents an adjustment of the wage in response to shifts in aggregate demand for both input factors in general equilibrium. Since robots and labor enter the production technology of intermediate firms as complements, a decline of the robot rental rate leading to increased robot uptake would entail a positive wage response, given that we model labor supply as perfectly inelastic. As indicated by θ_i, this affects firms proportionally to their labor-intensity of production.

Thus, a decrease in the robot rental rate implies a decrease in a firm's marginal costs V_i as long as the direct effect is larger than the indirect effect. We discuss the wage response effect in more detail in section 3.2.3 on the general equilibrium effects.

Effect on profit-maximizing price. To illustrate the effect of changes in the robot rental rate on an intermediate firm's profit-maximizing prices, we again construct the differential of our equation of interest, which in this case is the price equation (eq. 3.10). Based on the differential, we construct the elasticity of the profit-maximizing price with respect to the robot rental rate, which takes the form

$$\frac{d\ln p_i(s)}{d\ln r} = -\frac{1}{\epsilon_i(s)-1}\frac{d\ln \epsilon_i(s)}{d\ln r} + \frac{d\ln V_i}{d\ln r}. \tag{3.18}$$

In addition to the effect on the marginal costs as depicted in equation 3.17, a firm's price is affected by a change in its demand elasticity, which, as shown above, is a function of its market share. We construct the differential of the demand elasticity to again rearrange for its elasticity with respect to the robot rental rate and obtain

$$\frac{d\ln \epsilon_i(s)}{d\ln r} = -(\gamma-\sigma)\frac{\varphi_i(s)}{\epsilon_i(s)}\frac{d\ln \omega_i(s)}{d\ln r}, \tag{3.19}$$

which is a function of the elasticity of the market share with respect to the robot rental rate.

Effect on market shares. A firm's market share is defined as a relative measure of its profit-maximizing price to the price index of the sector it is active in. Thus, how a firm's market share reacts to decreasing prices of robots depends on its factor intensity draw, i.e., its output elasticity of robots $(1-\theta_i)$. The higher a firm's output elasticity for robots, the larger the magnitude of the price reduction effect. The firm with the highest output elasticity for robots in a given sector will experience the largest increase in market share in a given sector. We find the elasticity of the market share with respect to the robot rental rate based on equation 3.13, it takes the form

$$\frac{d\ln\omega_i(s)}{d\ln r} = (1-\gamma)\left(\frac{d\ln p_i(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right). \tag{3.20}$$

Since $\gamma > 1$, a firm's market share will increase in response to a reduction in the robot rental rate, if its own price decreases by more than the price index.

Effect on sectoral price indexes. The elasticity of the sectoral price index with respect to the robot rental rate can be written as a market share weighted sum of the changes in individual firm prices.[2] We can write it as

$$\frac{d\ln p(s)}{d\ln r} = \sum_{i=1}^{n}(s)\omega_i(s)\frac{d\ln p_i(s)}{d\ln r}. \tag{3.21}$$

Markups. The final step to fully gauge the effect of a change in the robot rental rate on a firm's markup is to combine the above derived elasticities. We again refer to the appendix for details and present the fully expanded solution for equation 3.19

$$\frac{d\ln\epsilon_i(s)}{d\ln r} = \frac{(\gamma-\sigma)(\gamma-1)}{1+\Omega_i}\frac{\varphi_i(s)}{\epsilon_i(s)}\left(1-\theta_i-\frac{\sum_{i=1}^{n(s)}\varphi_i(s)\frac{1-\theta_i}{1+\Omega_i}}{\sum_{i=1}^{n(s)}\frac{\Omega_i}{1+\Omega_i}}\right), \tag{3.22}$$

where $\Omega_i = \frac{(\gamma-\sigma)(\gamma-1)}{\epsilon_i(s)-1}\frac{\varphi_i(s)}{\epsilon_i(s)}$. Whether the demand elasticity increases (decreases) and therefore the markup decreases (increases) in response to a reduction in the rental rate depends on a firm's robot intensity relative to the average robot intensity in the same sector. We can distinguish between two cases

i) If $(1-\theta_i) > \frac{\sum_{i=1}^{n(s)}\varphi_i(s)\frac{1-\theta_i}{1+\Omega_i}}{\sum_{i=1}^{n(s)}\frac{\Omega_i}{1+\Omega_i}}$ then $\frac{d\ln\epsilon_i(s)}{d\ln r} > 0$ and $\mu_i(s)$ increases in response to reduction in r,

[2] See appendix 3.A.1

ii) If $(1-\theta_i) < \frac{\sum_{i=1}^{n(s)} \varphi_i(s) \frac{1-\theta_i}{1+\Omega_i}}{\sum_{i=1}^{n(s)} \frac{\Omega_i}{1+\Omega_i}}$ then $\frac{d \ln \epsilon_i(s)}{d \ln r} < 0$ and $\mu_i(s)$ decreases in response to reduction in r.

3.2.3. General equilibrium

In the general equilibrium, firms will adjust their factor demands according to the changes in the robot rental rate. With robots and labor being complementary in the production technology we introduce, a decrease in the rental rate of robots will lead to increased labor demand, which implies upward pressure on wages with supply being perfectly inelastic. The feedback on wages following a decline in the robot rental rate will thus further exacerbate the effect of market concentration, since higher wages affect firms inversely to their robot intensity. High robot intensity firms are thus relatively better off compared to low robot intensity firms not only because they benefit more from the decreased robot rental rate, but also, because they are less affected by the increase in wages.

3.2.4. Small open economy: foreign competition

We extend the model to a simple two-country case, in which intermediate good producing firms can sell to the final stage in the country foreign to them, in addition to selling to the final stage in their home economy. We use this simplistic two-country economy model to illustrate, what we call, the international competition effect. As we have seen in the one-country model, the domestic effect of a reduction in the robot rental rate will make the high-robot-intensity firms better off, because they will be able to reduce their marginal costs the most, allowing them to achieve higher market shares and markups while setting lower prices. We introduce fixed costs, that a firm must pay in order to gain access to the respective foreign market. Firms therefore choose to export based on their technology draws. Increasing robot use by foreign exporters will thus exert downward pressure on domestic firms' markups across all technology levels. Firms that were on the verge of benefiting from decreasing robot prices in the one-country economy are displaced by more productive, more robot-intense foreign competitors in the two-country economy. Overall, firms operating with above-average robot intensity will benefit from a reduction in the robot rental rate in both countries, while labor-intensive firms, i.e. firms with below-average robot-intensity, will be crowded out in both markets. In the following, we derive the effect of increased foreign competition for firm's domestic outcomes.

Intermediate good producers

Due to constant returns, the markup a firm generates in its home and foreign markets are the result of separate firm problems. A firm therefore faces two separate demand functions, one representing demand from its home market and one from its foreign market. Demand

for intermediate goods from domestic producers in the home market takes the form

$$y_i^H(s) = \left(\frac{p_i^H(s)}{p(s)}\right)^{-\gamma} \left(\frac{p(s)}{P}\right)^{-\sigma} Y, \tag{3.23}$$

while demand for intermediate goods from foreign producers in the home market is

$$y_i^F(s) = \left(\frac{p_i^F(s)}{p(s)}\right)^{-\gamma} \left(\frac{p(s)}{P}\right)^{-\sigma} Y. \tag{3.24}$$

Conceptually, the aggregate price index P remains unchanged from the one-country economy model. The sectoral prices $p(s)$ now include the prices of not only domestic but also foreign firms. Thus, the aggregate price index P now reflects the prices of domestic and foreign firms operating in the home country. This is illustrated by the two-country sectoral price index equation

$$p(s) = \left(\sum_{i=1}^{n(s)} p_i^H(s)^{1-\gamma} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} p_i^F(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}}, \tag{3.25}$$

where $\tau \geq 1$ depicts iceberg trade costs. A firm's market share in its home market is therefore determined not only by its competitiveness vis-à-vis domestic competitors, but also vis-à-vis foreign competitors operating in its home market, whose revenue enters in the denominator

$$\omega_i^H(s) = \frac{p_i^H(s) y_i^H(s)}{\sum_{i=1}^{n(s)} p_i^H(s) y_i^H(s) + \tau \sum_{i=1}^{n(s)} p_i^F(s) y_i^F} = \left(\frac{p_i^H(s)}{p(s)}\right)^{1-\gamma}. \tag{3.26}$$

We also introduce fixed costs of exporting denoted f_x. Due to profit-maximizing behavior some firms select into exporting. The exporting decision for foreign firms can be written as

$$\left(p_i^F(s) - \frac{V_i}{\varphi_i(s)}\right) y_i^F(s) \geq f_x. \tag{3.27}$$

3.2.5. Reduction in the robot rental rate with foreign competition

Using the equations adapted for the two-countries case laid out in the previous section, we pursue a similar strategy as in the one-country economy to examine the effect of foreign robot adoption on home market firm outcomes. We construct a set of elasticities, that, in combination illustrate the effect of foreign robot adoption on home firms' market shares and markups.

Effect on the domestic market share. In contrast to the one-country economy model, in the two-countries economy a firm's domestic market share is additionally determined by the prices of foreign competitors, as formulated in equation 3.26. In order to capture the

full effect on firms' domestic market shares in the two-countries economy, we construct the market share elasticity with respect to the robot rental rate. It takes the form

$$\frac{d\ln\omega_i^H(s)}{d\ln r} = (1-\gamma)\left(\frac{d\ln p_i^H(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right). \tag{3.28}$$

The presence of foreign firms implies downward pressure on domestic firms' markups, if it increases the elasticity of the sector price with respect to the robot rental rate. More specifically, the sign of equation 3.28 remains negative as long as the elasticity of the firm price is larger than the elasticity of the sector price.

A negative sign implies that a decrease in the robot rental rate leads to an increase in the domestic market share of firm i. If the sector price elasticity were larger than the firm price elasticity, the sign of equation 3.28 were positive, which would imply that a decrease in the robot rental rate led to a decrease in firm i's domestic market share. Therefore, the next step is to derive the elasticity of the sectoral price with respect to the robot rental rate.

Elasticity of the sector price index. The elasticity of the sectoral price with respect to the robot rental rate in the two-countries case takes the form

$$\frac{d\ln p(s)}{d\ln r} = \sum_{i=1}^{n(s)} \omega_i^H(s)\frac{dp_i^H(s)}{d\ln r} + \tau^{1-\gamma}\sum_{i=1}^{n(s)} \phi_i^F(s)\omega_i^F(s)\frac{dp_i^F(s)}{d\ln r}, \tag{3.29}$$

where ϕ_i^F is a binary variable indicating firm activity, based on a firm's exporting decision formulated in equation 3.27. The summand on the right-hand side represents the effect of foreign firms on the sectoral price index in the home country. Depending on their technology draws, some foreign firms will be able to lower their profit-maximizing prices in response to a reduction in the robot rental rate, while others will not. If the presence of foreign firms increases the sector price elasticity, or in other words, if the right-hand summand is positive, this puts downward pressure on the market shares of domestic firms. Due to exporting fixed costs the firms selecting into exporting are more competitive than firms not selecting into exporting. Assuming symmetric countries and thereby equal technology distributions, the average active foreign firm in the home market will be more competitive than the average domestic firm.

Effect on demand elasticity. Recall, that a firm's demand elasticity determines its markup, as shown in equation 3.14. Deriving the results for changes in markups therefore requires deriving changes in firms' demand elasticities in response to changes in the robot rental rate. We construct the corresponding elasticity

$$\frac{d\ln\epsilon_i^H(s)}{d\ln r} = -(\gamma-\sigma)\frac{\omega_i^H(s)}{\epsilon_i^H(s)}\frac{d\ln\omega_i^H(s)}{d\ln r}, \tag{3.30}$$

which again depends on the change in a firm's market share. Thus, if the presence of foreign

firms causes a firm's market share elasticity to change from a negative sign to a positive sign as discussed above, the sign of the demand elasticity equation formulated in equation 3.30 changes from negative to positive in response. In this case, a firm that would have benefited from the decrease in the robot rental rate in the one-country economy would lose in terms of market share and markups due to the adoption robot by foreign competitors.

In general, all domestic firms, regardless of their technology level, will experience downward pressure on market shares and markups as long as the foreign firms contribute to a decline in the sectoral price. For symmetric countries, this is the expected outcome, given the selection of above-average competitive firms into exporting.

3.3. Empirical Strategy

3.3.1. Markup Estimation

We estimate industry-level markups by slightly adapting the procedure developed by De Loecker & Warzynski (2012) to include robots in production. The firm-level data needed for the estimation comes from from Worldscope. Worldscope contains financial statements for more than 80,000 companies worldwide. The sample consists mainly of publicly traded firms, with few privately held firms.[3] Markups are the ratio of price (P) to marginal cost (MC) and are a direct measure of market power (De Loecker et al., 2020). The advantage of using markups instead of standard concentration indices such as the Herfindahl-Hirschman index is that the latter do not measure market power when there is product differentiation (De Loecker et al., 2020) and that one would require data on all firms in the market, which we do not have. The method builds on the observation that markups can be estimated using expenditure shares and output elasticities, which follows from standard cost minimization via a Lagrange function. Markups can thus be expressed as

$$\mu_{ist} = \frac{P_{ist}}{MC_{ist}} = \frac{\theta^V_{it}}{\alpha^V_{ist}},$$

where θ^V_{ist} is the output elasticity of variable input V and α^V_{ist} is the expenditure share on input V of firm i in sector s at year t. The expenditure shares are directly be observable in the data.

To obtain output elasticities, we estimate a Cobb-Douglas production function separately for each industry, following De Loecker & Eeckhout (2018). Unfortunately, since we do not have information about robots in the firm-level Worldscope data, but only at the sector level, it is not possible for us to directly estimate firm-level robot output elasticities. To adhere as much as possible to the established procedure for estimating markups on the on hand, and to incorporate robots in the markups estimation on the other hand, we alter the

[3] De Loecker & Eeckhout (2018) use the same data and perform some robustness tests to ensure that the selection of firms in the data does not lead to biased results.

standard production function used for the markup estimation in De Loecker & Eeckhout (2018) by adding sector level robots. The result is a production function that extends the one introduced in the theoretical part of this paper including labor l and the stock of robots R, by variable inputs v, and capital k. We argue that omitting variable inputs v and capital k in the estimation equation could raise omitted variable bias concerns and thus include them in the estimation. The resulting Cobb-Douglas production function takes the form

$$q_{ist} = \beta_v v_{ist} + \beta_k k_{ist} + \beta_l l_{ist} + \beta_r R_{st} + \omega_{ist} + \epsilon_{ist}$$

with q denoting output and all variables being in logs and deflated.[4] Unobserved productivity is given by ω. Estimating the production function yields output elasticities β. The estimation follows Ackerberg et al. (2015), who use a control function approach to overcome simultaneity bias between input demand and unobserved productivity. In a first step, expected output (ϕ_{ist}) is estimated

$$q_{ist} = \phi_t(v_{ist}, k_{ist}, l_{ist}, R_{st}, z_{ist}) + \epsilon_{ist},$$

where z are other variables that affect the demand for variable inputs (we use a set of fixed effects to control for other variables) and ϵ_{ist} is the residual of estimating expected output. Following the authors, we correct for variation in expenditure not correlated to variables impacting input demand using ϵ_{ist}: $\hat{\alpha}^V_{ist} = \frac{P^V_{ist} V_{ist}}{P_{ist} \hat{Q}_{ist} / exp(\hat{\epsilon}_{ist})}$, where we use a set of fixed effects to control for other variables that affect the demand for variable inputs.

Next, the inverse demand of variable input $h_t(\cdot)$ is used to rewrite expected output as

$$\phi_{ist} = \beta_v v_{ist} + \beta_k k_{ist} + \beta_l l_{ist} + \beta_r R_{st} + h_t(v_{ist}, k_{ist}, l_{ist}, R_{st}, z_{ist}).$$

With the expected output, productivity can be computed as $\omega_{ist}(\beta) = \hat{\phi_{ist}} - \beta_v v_{ist} - \beta_k k_{ist} - \beta_l l_{ist} - \beta_r R_{st}$ (De Loecker & Warzynski, 2012). The productivity innovation ξ_{ist} is recovered by non-parametrically regressing $\omega_{ist}(\beta)$ on its lag. With this, all coefficients of the production function can be obtained through GMM with the moment conditions

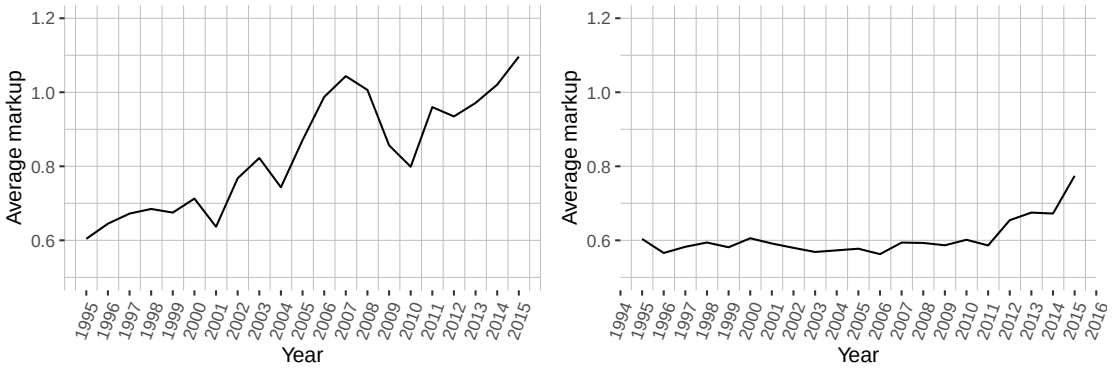
$$E\left[\xi_{ist}(\beta) \begin{pmatrix} v_{ist-1} \\ k_{ist} \\ l_{ist-1} \\ R_{st-1} \end{pmatrix} \right] = 0. \tag{3.31}$$

The output elasticity of variable input v is then given by $\theta_{st} = \hat{\beta}_v$.

After estimating markups at the firm level, we aggregate them to the sector level. In the main specification, we weight each markup by the firm's share of industry output. As a robustness test, we use the average markups as a measure. Figure 3.1 shows that markups have steadily increased over the past decades and, that our markup estimates are similar to

[4] We obtain capital, price and GDP deflators from Worldbank's WDI and OECD's STAN database.

Figure 3.1: Estimated average markups over time for all sectors versus manufacturing sectors using Worldscope data.



(a) All sectors

(b) Manufacturing sectors

those of De Loecker & Eeckhout (2018). In panel 3.1b, we plot the evolution of markups in the manufacturing sector only. While markups have increased after 2011, there is not as strong an overall upward trend as in panel . This suggests that the service sector was largely responsible for the strong markup increases between 1995 and 2015 (Lashkari et al., 2018; D. Autor et al., 2020).

3.3.2. Estimation Equation

To estimate the impact of automation on markups and other outcomes related to market power and concentration, we first employ a simple regression model

$$y_{cst} = \alpha_{cst} + \beta_R R_{cst} + \beta_\chi \chi_{cst} + \gamma_{cs} + \delta_{ct} + \eta_{st} + \epsilon_{cst}, \tag{E.1}$$

where c denotes the country, s the sector, t the year and the outcome of interest is y_{cst}. R_{cst} is the stock of domestic robots per 1000 workers. In addition, χ_{cst} represents a vector of control variables, γ_{cs} country sector fixed effects, δ_{ct} country year fixed effects and η_{st} sector year fixed effects.[5] Thus, we observe changes only within sectors of countries over time, while controlling for all other larger-scale developments and characteristics.

Data on the stock of robots by country, industry, and year are obtained from the International Federation of Robotics (IFR). The IFR provides the annual number of "multi-purpose industrial robots"[6] installations at the country, industry and application levels (International Federation of Robotics, 2018). Industries are defined at the three-digit or two-digit level according to ISIC classifications.

Estimation equation E.1 already gives a first indication of the relationship between au-

[5] Our main controls are the number of patents and the capital stock, both of which we take from the Worldscope database, and net exports, which we take from the OECD ICIO.

[6] A robot is defined by ISO 8373:2012 as an automatically controlled, re-programmable, multi-purpose manipulator, programmable in three or more axes, which can be either fixed in place or mobile for use in industrial automation applications (International Federation of Robotics, 2018)

tomation and markups. However, the choice to use robots in production is likely to be endogenous to markups. For instance industries with higher markups could have more resources to employ robots. Therefore, we use an IV approach to obtain exogenous variation in robot uptake. Following the current literature, we argue that the global stock of robots is likely to be exogenous to single industries, and represents the overall decline in robot prices (G. Graetz & Michaels, 2018; Artuc et al., 2019). We construct a similar but novel IV,

$$R_{cst}^{IV} = R_t^G \frac{O_{cs}}{L_{cs}} I_c,$$

which interacts the global stock of robots R^G with country- and sector-level predictors of the degree of automation. The fraction O_{cs}/L_{cs}, output per worker of sector s in country c in 1995, reflects the *potential* of a sector to employ robots. The source of these data are the OECD's ICIO tables and the OECD's Annual Labor Force Statistics (OECD, 2021, 2023b), respectively. I is a measure of technological *capacity* in 1990, developed by Archibugi & Coco (2004). Thus, our IV exploits exogenous variation over time and cross-sectional capabilities to install automation technologies.

In our main specification we estimate equation E.1 in a two-stage procedure, where R_{cst} is instrumented with R_{cst}^{IV} in the first stage. Given the data requirements, we are able to estimate the equation for 29 countries and 20 sectors, between 1995 and 2015. However, we do not have a complete panel for all combinations of countries and sectors.

In a second step, we test the theoretical predictions made in section 3.2.5 and include a measure of foreign robot competition in our model. Following De Benedictis & Tajoli (2007a,b), we construct a similarity index for the correlation between sectoral exports between two countries of the following form:

$$m_{cdst} = 1 - \frac{\sum_{ps} |x_{cst} - x_{dst}|}{\sum_{ps} x_{cst} + x_{dst}}.$$

Within each sector s, the index compares the exports of two countries c and d over a range of products p in each year t.[7] The resulting index is bounded between 0 and 1, where the closer it is to 1, the more similar the exports of two countries are in that sector.

We expect that the more similar the domestic and foreign economies are, the greater the competition from foreign robots. Therefore, to construct a measure of foreign automation, we weight the stock of foreign robots per worker R_{dst} by the similarity index, which yields a competition-weighted measure of foreign robots F_{cst}:

$$F_{cst} = \sum_d m_{cds} R_{dst}.$$

We can therefore test for the differential effects of domestic and foreign automation by

[7] We use exports at the 4-digit level from Comtrade, over our 18 sectors.

Table 3.1: Automation and markups - OLS

	All sectors		Only manufacturing	
	(1)	(2)	(3)	(4)
Stock of robots p.w.	-0.049***	-0.047***	-0.052***	-0.050***
	(0.01)	(0.01)	(0.02)	(0.02)
Observations	4580	4185	3354	3354
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls		✓		✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The stock of robots per worker is mean standardized to a standard deviation of one. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Markups are aggregated on the industry level by each firm's share of sales. Regressions run from 1995 to 2016.

Table 3.2: Automation IV - First stage

	All sectors		Only manufacturing	
	(1)	(2)	(3)	(4)
Robot IV	0.716	0.688	1.170***	1.123***
	(0.43)	(0.42)	(0.21)	(0.20)
Observations	4530	4155	3354	3354
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls		✓		✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The stock of robots per worker is mean standardized to a standard deviation of one. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Regressions run from 1995 to 2016.

including F_{cst} in the estimation equation E.1.

3.4. Estimation Results

3.4.1. Automation and Markups

In this section, we test the theoretical predictions of the model, by estimating how an increase in automation has affected markups. As laid out in section 3.2, automation is likely to affect firms differently depending on their level of productivity. We start with estimating how increasing usage of robots affects markups and other measures of market power domestically, before moving to the effects of foreign automation. We thereby establish a complete picture of the effects of automation on market power.

Table 3.1 presents the results of estimating equation E.1 with an OLS model. In all specifications, the standard errors are two-way clustered at the country and sector level. The outcome is the logarithm of markups, where industry-level markups are obtained by

Table 3.3: Automation and markups - IV

	All sectors			Only manufacturing		
	(1)	(2)	(3)	(4)	(5)	(6)
Stock of robots p.w.	-0.246**	-0.245**	-0.237**	-0.176***	-0.183***	-0.172***
	(0.11)	(0.10)	(0.10)	(0.02)	(0.02)	(0.02)
Observations	4530	4155	3628	3354	3354	2843
Country × Sector Dummies	✓	✓	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓	✓	✓
Controls		✓	✓		✓	✓
Additional Controls			✓			✓
KP F-Statistic	2.77	2.65	2.9	31.8	30.8	57.5

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The stock of robots per worker is mean standardized to a standard deviation of one. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Markups are aggregated on the industry level by each firm's share of sales. Regressions run from 1995 to 2016.

weighting each firm's markup by its share of sales in the industry total. In the first two columns, the regressions are run over all sectors, while in the latter two only manufacturing sectors are examined. Columns 2 and 3 add industry-level production and net exports in logarithms as controls. The coefficient on the stock of robots per worker is statistically significant in all specifications and indicates a negative relationship between the stock of robots and average markups on average. The effect is stronger for manufacturing sectors.

As laid out above, a firm's market power reflects idiosyncratic characteristics of firms that are associated with the likelihood of robot adoption. The results are thus likely to be biased by reverse causality and we therefore use an IV approach to obtain unbiased estimates. As outlined in section 3.3.2, we address the endogeneity in the decision to automate by using an instrumental variable. The results of the first-stage regression are presented in Table 3.2. While the instrumental variable is not significant for all sectors, it is highly significant and has a positive coefficient for the manufacturing sectors. This is not surprising, as industrial robots are almost exclusively used in manufacturing production. The inclusion of the control variable does not change the coefficient or the precision of the instrument. The instrument is therefore a valid predictor of robot adoption.

Table 3.3 shows the results of the second-stage. The first thing to note is that the Kleibergen-Paap F-statistic is low with all sectors, but is above the usual thresholds in the manufacturing sectors, indicating that the instrument is not valid in the service and agricultural sectors, as found in Table 3.2.

As in the OLS setting, the stock of domestic robots has a negative and statistically significant coefficient. Since the coefficient is free of endogeneity concerns, we can now interpret the coefficient as a causal effect. The coefficient on the stock of robots is statistically significant and negative throughout. In columns 3 and 6, we add the logarithm of the number of patents at the industry-level as well as the logarithmized industry-level capital stock. The number of patents controls for the industry's innovation capacity and the capital stock for

Table 3.4: Automation and markups - Quintile regressions

		Sales-quintiles		Markup-quintiles	
	(1)	(2)	(3)	(4)	(5)
Stock of robots p.w.	-0.004**				
	(0.00)				
1. Quintile × Stock of robots p.w.		-0.016***	-0.013**	-0.035**	-0.037**
		(0.01)	(0.00)	(0.01)	(0.01)
2. Quintile × Stock of robots p.w.		-0.011***	-0.009***	-0.018**	-0.018**
		(0.00)	(0.00)	(0.01)	(0.01)
3. Quintile × Stock of robots p.w.		-0.006***	-0.005**	-0.004**	-0.003
		(0.00)	(0.00)	(0.00)	(0.00)
4. Quintile × Stock of robots p.w.		0.003	0.005	0.012*	0.015**
		(0.00)	(0.00)	(0.01)	(0.01)
5. Quintile × Stock of robots p.w.		0.014*	0.015**	0.033*	0.039**
		(0.01)	(0.01)	(0.02)	(0.01)
Observations	11654	11641	10389	11618	10365
Country × Sector Dummies	✓	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Additional Controls			✓		✓
KP F-Statistic	29	.835	1	.858	.998

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. Quintiles are based on firms' sales in the previous year in columns 2 and 3 and on firms' markups in the previous year in columns 4 and 5. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

the overall capital intensity, both of which are correlated with the adoption of robots. The inclusion of the additional control variables reduces the sample size because the variables are not available for all observations, but the coefficient in column 6 remains statistically significant at the 1% level. Thus, a larger stock of robots appears to reduce industry markups, on average. This finding points to a distribution of technology across firms which according to our model (section 3.2) suggests that: New technology benefits only a few firms at the expense of others, leading to an average negative effect on markups. The effect of automation on markups is substantial: A one standard deviation increase in the stock of robots per worker reduces average markups by 17%.

To examine whether it is only high-productivity and high-sales firms which benefit from automation, we split the firms in our sample into quintiles within each sector, based on their sales and markups in the previous year, to obtain a fuller picture of the distributional effects of automation.[8] D. H. Autor et al. (2016) show that the rise of markups is driven by "superstar" firms. Furthermore, in another recent study using a similar setting, Stiebale et al. (2020) find no effect of automation on markups for manufacturing firms on average, but an increase for the highest quintile of firms.

In Table 3.4, the level of observation is now sector-quintiles. The first column reproduces

[8]For observations without information on the previous year, we use sales and markups of the same year.

Table 3.5: Automation, production and exports

	log Production		log Exports	
	(1)	(2)	(3)	(4)
Stock of robots p.w.	0.199***	0.156***	1.845	1.239
	(0.02)	(0.02)	(1.89)	(1.45)
Observations	3353	2842	3353	2842
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Additional Controls		✓		✓
KP F-Statistic	31.2	59.7	31.3	58.3

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The stock of robots per worker is mean standardized to a standard deviation of one. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the log industry production in columns 1 and 2 and the log net exports in columns 3 and 4. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

the previous results at the alternative level of observation. In columns 2 and 3, we interact the sales quintile with the domestic stock of robots and the corresponding instrument. The same procedure is repeated in columns 4 and 5, using firms' markups to construct quintiles. In both settings, and in line with the current literature (Stiebale et al., 2020), we also find that the decline in average markups is driven by firms in the lowest 3 quintiles. Conversely, firms in the top quintile experience an increase in markups.[9]

This suggests interesting within-industry heterogeneity. The largest and most productive firms are able to reap disproportional benefits from automation. At the same time, less productive firms face greater competition due to the lower production costs of automating firms. As a consequence, markups of these firms decrease.

To see whether this pattern is driven by individual industries, we disaggregate the manufacturing sector in Table 3.A3. We run the quintile-level analysis for each individual industry. The Table shows that most sectors have a similar pattern. Although the estimation power is limited due to the smaller number of observations, the coefficient of the interaction between the stock of robots and the highest quintile is positive in almost all industries. Similarly, the coefficients of the first and second quintile are almost entirely negative. Notably, there are are negative and statistically significant coefficients in the computer electronics industry.

3.4.2. Alternative Outcomes

Having provided evidence above that domestic automation reduces average markups, we now turn to alternative outcomes related to output and market concentration.

[9] We find the same pattern when estimating markups using a translogirathmized instead of a Cobb-Douglas function. Table 3.A2 shows that robots have, on average, a negative effect on markups estimated in this way, and that the negative effect is driven by lower quantiles.

Table 3.6: Automation and alternative outcomes

	log Number of Firms		log Output Prices		log Operating Margin	
	(1)	(2)	(3)	(4)	(5)	(6)
Stock of robots p.w.	-0.032	0.423	-0.137*	-0.140*	-0.622***	-0.838***
	(0.35)	(0.28)	(0.07)	(0.07)	(0.21)	(0.17)
Observations	1565	1282	2637	2216	2365	1932
Country × Sector Dummies	✓	✓	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
KP F-Statistic	9.63	10.3	21.5	31.6	13.7	26.5

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The stock of robots per worker is mean standardized to a standard deviation of one. The sample consists of manufacturing sectors only. All outcome variables are measured in logs. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Markups are aggregated on the industry level by each firm's share of sales. Regressions run from 1995 to 2016.

First, in Table 3.5, we examine how automation affects production and exports. Production increases with a larger stock of robots (columns 1 and 2), as might be expected and as has been found, for example, by G. Graetz & Michaels (2018) and Koch et al. (2019). Net exports, on the other hand, are not affected by automation.[10] Table 3.A1 in the appendix displays firms' sales as an outcome, based on quintiles by their sales and markups in the previous year. The results show that the average decline is again driven by the firms in the lowest quintile, which are less likely to install robots in production and thus face more competition. As with markups, more productive firms benefit from automation.

Next, we turn to alternative measures of market concentration. First, in columns 1 and 2 of Table 3.6, we find no changes in the total number of firms in a sector associated with an increased robot adoption. It should be noted, however, that the sample size here is relatively small, as data on the number of firms are not available for all countries.

In columns 3 and 4, we examine how prices are affected by automation. We find a negative association between the stock of robots and prices in the baseline setting and with additional controls, which supports the previous results. Robotization reduces sectoral prices and only the most productive firms benefit due to lower marginal costs. Lastly, we use firms' operating margin as an alternative measure of markups. Firms report their operating margin directly in the Worldscope data, which eliminates the possibility of estimation error.[11] The coefficient on the stock of robots per worker is again statistically significant and even larger in size, both in the baseline setting and with additional controls. Our results are thus robust to alternative measures of markups.

[10] The same holds true for the log of exports, rather than the log of net exports.
[11] The operating margin is defined as the operating income divided by net sales.

Table 3.7: Foreign automation and markups

	Only manufacturing sectors			
	(1)	(2)	(3)	(4)
Foreign weighted robots	-0.033***	-0.035***	-0.015**	-0.022**
	(0.01)	(0.01)	(0.01)	(0.01)
Stock of robots p.w.			-0.046**	-0.047**
			(0.02)	(0.02)
Observations	3354	3354	3354	2843
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls		✓	✓	✓
Additional Contols				✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. The weighted foreign robot stock and the stock of domestic robots per worker are mean standardized to a standard deviation of one. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Markups are aggregated on the industry level by each firm's share of sales. Regressions run from 1995 to 2016.

3.4.3. Foreign Automation

Having established that domestic automation reduces average industry level markups, driven by low-sales, and low-markup firms, we now turn to the question of how foreign automation affects domestic markups. Our theoretical model predicts that foreign automation will depress domestic markups, due to increased competition through lower production costs abroad.

Table 3.7 presents the results of including the foreign weighted robot measure F_{cst} (see section 3.3.2) into our estimation equation E.1. We focus on the manufacturing sectors, as these were found to drive our previously found results. The coefficient of foreign-weighted robots is statistically significant and negative throughout. Adding controls in column 2 doesn't change the coefficient. While including the domestic stock of robots in column 3 reduces the size of the coefficient, it remains statistically significant at the 5% level. Moreover, The finding is robust to the inclusion of additional controls.

Compared to domestic automation, we expect that increasing competition from foreign automating firms will not only affect lower productivity and smaller firms. Indeed, in the fully specified model with sales quintiles we find a decrease in markups along the entire distribution of firms, as shown in column 2 of Table 3.8. With quintiles based on markups, we find a statistically significant effect only for the fourth quintile. In contrast to the previous results, it is rather the firms in the middle quintiles that experience a larger reduction in markups. These firms seem to face the strongest competition from foreign firms.

Foreign automation thus seems to put additional strain on domestic firms, but not only on the smallest ones. Competition from foreign producers, which can reduce their production costs, reduces the market power and market share of domestic firms. Table 3.8 provides

Table 3.8: Foreign automation and markups - Quintile regressions

	Sales-quintiles		Markup-quintiles	
	(1)	(2)	(3)	(4)
1. Quintile × Stock of robots p.w.	-0.004***	-0.003**	-0.007***	-0.008***
	(0.00)	(0.00)	(0.00)	(0.00)
2. Quintile × Stock of robots p.w.	-0.002**	-0.002	-0.003***	-0.004***
	(0.00)	(0.00)	(0.00)	(0.00)
3. Quintile × Stock of robots p.w.	-0.001	-0.001	-0.001	-0.002*
	(.)	(0.00)	(0.00)	(0.00)
4. Quintile × Stock of robots p.w.	0.003***	0.003**	0.003***	0.004**
	(0.00)	(0.00)	(0.00)	(0.00)
5. Quintile × Stock of robots p.w.	0.004**	0.006**	0.007***	0.010**
	(0.00)	(0.00)	(0.00)	(0.00)
1. Quintile × Foreign weighted robots	-0.002	-0.011**	0.003	-0.004
	(.)	(0.01)	(0.02)	(0.02)
2. Quintile × Foreign weighted robots	-0.013	-0.025**	0.003	-0.007
	(0.01)	(0.01)	(0.02)	(0.01)
3. Quintile × Foreign weighted robots	-0.013	-0.022**	-0.002	-0.011
	(0.01)	(0.01)	(0.01)	(0.01)
4. Quintile × Foreign weighted robots	-0.019**	-0.038***	-0.010	-0.031*
	(0.01)	(0.01)	(0.01)	(0.02)
5. Quintile × Foreign weighted robots	0.006	-0.023**	-0.007	-0.042
	(0.01)	(0.01)	(0.02)	(0.03)
Observations	11641	10389	11618	10365
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Additional Controls		✓		✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. Quintiles are based on firms' sales in the previous year in columns 2 and 3 and on firms' markups in the previous year in columns 4 and 5. The sample consists of only manufacturing sectors. The coefficients of foreign robot exposure are displayed in 1000s, to ensure visibility. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

further evidence of this pattern. While domestic automation leads to larger industry-level production, competition to foreign automation is associated with lower production levels. No effect is found for net exports.

While exporting firms face greater competition from foreign firms that can produce at lower costs, increased production by the latter could increase demand for inputs. Therefore, foreign automation may have countervailing effects. Increased demand for inputs may spur prices and output of input-providing firms. We therefore add an additional measure of exposure to foreign robots in Table 3.A4, which captures input-output linkages. We weight each foreign sector's stock of robots per worker by the share of input exports (imports) from a domestic sector to the respective foreign sector.[12] In columns 1 and 2, we weight foreign robots with imports and in columns 3 and 4 with exports. Contrary to domestic robots and similarity-weighted foreign robots, the coefficient of input-trade-weighted foreign robots is positive. However, the coefficients are only statistically significant when not including

[12]Data on input exports and imports are taken from the OECD's ICIO database.

Table 3.9: Foreign automation, production and exports

	log Production		log Exports	
	(1)	(2)	(3)	(4)
Stock of robots p.w.	0.106***	0.087***	1.127	1.059
	(0.03)	(0.03)	(0.78)	(0.84)
Foreign weighted robots	-0.094***	-0.081***	-0.132	0.553
	(0.03)	(0.02)	(1.04)	(0.77)
Observations	3353	2842	3353	2842
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Additional Controls		✓		✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. Coefficients of foreign robot exposure are displayed in 1000s, to ensure visibility. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the log industry production in columns 1 and 2 and the log net exports in columns 3 and 4. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

additional control variables and thereby losing observations. Moreover, the coefficient of export-weighted foreign robots is larger in both magnitude and statistical significance than the the coefficient of import-weighted robots. Therefore, input-providing firms appear to profit from automation abroad.

On the one hand, finding a positive coefficient for trade-weighted foreign robots reinforces confidence that we are indeed capturing increasing competition with our similarity-weighted robot measure. It also shows that automation affects different types of firms differently. Those which compete with automating firms are crowded out, while firms that provide inputs to these firms are may benefit from their increased production.

3.5. Conclusion

In this paper, we examine how automation shapes economies through the channel of market power. We develop a theoretical model that links automation, technological capability and markups. Building on the model of Edmond et al. (2015), we show that a reduction in the robot rental rate benefits high-productivity and high-technology firms. These firms are able to reap the benefits of automation and can reduce their production costs, allowing them to further increase their markups. This comes at the expense of low-productivity firms, which are unable to take advantage of the lower prices of robots in production. As high-productivity firms lower their output prices, low-productivity firms lose in terms of market share and market power. Furthermore, in a two-country-case, we show that additional competition from foreign automating firms increases the burden on lower-productivity firms. Firms which are able to export and can lower their production costs take away additional market share and thus market power from lower-productivity firms.

We test these theoretical predictions by estimating how markups are affected by automation. Since domestic automation is endogenous to market power and productivity levels, we employ a Two-Stage-Least-Squares (2SLS) strategy. Our findings indicate that automation has a negative impact on markups, in manufacturing industries. While markups decline on average, there is substantial heterogeneity within the economy. High-markup firms are able to increase their markups at the cost of low-markup and low-productivity firms. Firms in the three lowest markup and productivity quintiles experience large declines in markups. Complementing these results, we find that automation leads to lower average prices and lower average operating profit margins.

To empirically investigate how foreign automation affects domestic markups, we develop a novel measure of competition to foreign automation, which builds on the similarity of two countries' export structure. The more similar two countries are, the more we expect competition to increase when one country adopts more robots. Adding this measure to our estimations, we find that foreign automation does indeed lead to a reduction in markups. Again differentiating firms by markup and productivity quintile, we show that firms in the lower quintiles are the ones which lose market power due to foreign competition.

Our results add to the growing literature on the distributional effects of automation technology. While markups have risen sharply on average in recent decades (De Loecker & Eeckhout, 2018), relatively few firms are able to dominate whole markets (D. Autor et al., 2020). New technologies such as robots in production could exacerbate this trend. Making technology more readily available is therefore key to counteracting monopolistic markets.

3.A. Appendix

3.A.1. Theory

One-country economy

Intermediate goods producers

Profit-maximization-problem of the final good producer.

$$PY - \int_0^1 \left(\sum_{i=1}^{n(s)} p_i(s)\, y_i(s) \right) ds. \tag{A1}$$

Demand for intermediate goods. The relative demand for two varieties i and j within the same sector s is given by:

$$\frac{y_i(s)}{y_j(s)} = \left(\frac{p_i(s)}{p_j(s)} \right)^{-\gamma}. \tag{A2}$$

We multiply with the price of one variety and aggregate over $n(s)$ varieties within a sector:

$$p_i(s)\, y_i(s) = p_j(s)^{\gamma}\, y_j(s)\, p_i(s)^{1-\gamma},$$

$$\sum_{i=1}^{n(s)} p_i(s)\, y_i(s) = p(s)\, y(s) = p_j(s)^{\gamma}\, y_j(s) \sum_{i=1}^{n(s)} p_i(s)^{1-\gamma}.$$

By taking into account the definition of the price index (3.5), we obtain the demand for one variety:

$$p(s)\, y(s) = p_j(s)^{\gamma}\, y_j(s)\, p(s)^{1-\gamma},$$

$$y_j(s) = \left(\frac{p_j(s)}{p(s)} \right)^{-\gamma} y(s). \tag{A3}$$

Combining equation A3 with the demand function for one sector leads to:

$$y_j(s) = \left(\frac{p_j(s)}{p(s)} \right)^{-\gamma} \left(\frac{p(s)}{P} \right)^{-\sigma} Y. \tag{A4}$$

One country economy - reduction in the robot rental rate

Effect on marginal costs

Marginal costs of a firm i :

$$V_i = w^{\theta_i} r^{1-\theta_i} {\theta_i}^{-\theta_i} (1-\theta_i)^{\theta_i - 1} \tag{A5}$$

The total differential of marginal costs V_i with respect to changes of endogenous variables can be derived as follows:

$$dV_i = \theta_i w^{\theta_i - 1} dw r^{1-\theta_i} {\theta_i}^{-\theta_i} (1-\theta_i)^{\theta_i - 1} + w^{\theta_i}(1-\theta_i) r^{-\theta_i} dr {\theta_i}^{-\theta_i} (1-\theta_i)^{\theta_i - 1}$$

The change of θ_i can be neglected, as this is an exogenous firm-specific draw. We can simplify the total differential by using the definition of marginal costs V_i:

$$dV_i = \theta_i w^{-1} dw V_i + V_i^{-1}(1-\theta_i) dr$$

$$\frac{d \ln V_i}{d \ln r} = \frac{dV_i}{dr}\frac{r}{V_i} = \theta_i \frac{d \ln w}{d \ln r} + (1-\theta_i)$$

Effect on price

The optimal price of a firm is given by

$$\begin{aligned}
p_i(s) =& \frac{\epsilon_i(s)}{\epsilon_i(s)-1}\frac{V_i}{\varphi_i} \\
\frac{dp_i(s)}{dr} =& \frac{\frac{d\epsilon_i(s)}{dr}(\epsilon_i(s)-1) - \epsilon_i \frac{d\epsilon_i(s)}{dr}}{(\epsilon_i(s)-1)^2}\frac{V_i}{\varphi_i} + \frac{\epsilon_i(s)}{\epsilon_i(s)-1}\frac{dV_i}{dr}\frac{1}{\varphi_i} \\
\frac{dp_i(s)}{dr} =& -\frac{1}{\epsilon_i(s)-1}\frac{p_i(s)}{\epsilon_i(s)}\frac{d\epsilon_i(s)}{dr} + p_i(s)\frac{dV_i}{dr}\frac{1}{V_i} \\
\frac{d \ln p_i(s)}{d \ln r} =& -\frac{1}{\epsilon_i(s)-1}\frac{d \ln \epsilon_i(s)}{d \ln r} + \frac{d \ln V_i}{d \ln r}
\end{aligned}$$

Effect on demand elasticity

In the Bertrand version of the model, the demand elasticity is defined as

$$\begin{aligned}
\epsilon_i(s) =& \gamma(1-\omega_i(s)) + \sigma\omega_i(s) \\
\frac{d\epsilon_i(s)}{dr} =& -(\gamma-\sigma)\frac{\omega_i(s)}{dr} \\
\frac{d \ln \epsilon_i(s)}{d \ln r} =& -(\gamma-\theta)\frac{\omega_i(s)}{\epsilon_i(s)}\frac{d \ln \omega_i(s)}{d \ln r}
\end{aligned}$$

We now use the expression of the market share

$$\begin{aligned}
\omega_i(s) =& \left(\frac{p_i(s)}{p(s)}\right)^{1-\gamma} \\
\frac{d\omega_i}{dr} =& (1-\gamma)\left(\frac{p_i(s)}{p(s)}\right)^{-\gamma} \frac{\frac{dpi(s)}{dr}p(s) - p_i(s)\frac{dp(s)}{dr}}{p(s)^2} \\
\frac{d\omega_i(s)}{dr} =& (1-\gamma)\left(\frac{p_i(s)}{p(s)}\right)^{1-\gamma}\left(\frac{d\ln p_i(s)}{dr} - \frac{d\ln p(s)}{dr}\right) \\
\frac{d\ln\omega_i(s)}{d\ln r} =& (1-\gamma)\left(\frac{d\ln p_i(s)}{dr} - \frac{d\ln p(s)}{dr}\right)
\end{aligned}$$

Summary of effects

$$\frac{d\ln p_i(s)}{d\ln r} = -\frac{1}{\epsilon_i(s)-1}\frac{d\ln\epsilon_i(s)}{d\ln r} + \frac{d\ln V_i}{d\ln r} \tag{A6}$$

$$\frac{d\ln V_i}{d\ln r} = (1-\theta_i) > 0 \tag{A7}$$

$$\frac{d\ln\epsilon_i(s)}{d\ln r} = -\gamma(\gamma-\sigma)\frac{\omega_i(s)}{\epsilon_i(s)}\frac{d\ln\omega_i(s)}{d\ln r} \tag{A8}$$

$$\frac{d\ln\omega_i(s)}{d\ln r} = (1-\gamma)\left(\frac{d\ln p_i(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right) \tag{A9}$$

Combining eqs. A6 and A7 leads to:

$$\frac{d\ln p_i(s)}{d\ln r} = -\frac{1}{\epsilon_i(s)-1}\frac{d\ln\epsilon_i(s)}{d\ln r} + (1-\theta_i) \tag{A10}$$

Combining eqs. A8 and A9 leads to:

$$\frac{d\ln\epsilon_i(s)}{d\ln r} = (\gamma-\sigma)(\gamma-1)\frac{\omega_i(s)}{\epsilon_i(s)}\left(\frac{d\ln p_i(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right) \tag{A11}$$

Inserting eq. A11 into eq. A10 leads to:

$$\frac{d\ln p_i(s)}{d\ln r} = -\frac{1}{\epsilon_i(s)-1}(\gamma-\sigma)(\gamma-1)\frac{\omega_i(s)}{\epsilon_i(s)}\left(\frac{d\ln p_i(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right) + (1-\theta_i)$$

$$\frac{d\ln p_i(s)}{d\ln r}\left(1 + \frac{(\gamma-\theta)(\gamma-1)}{\epsilon_i(s)-1}\frac{\omega_i(s)}{\epsilon_i(s)}\right) = \frac{(\gamma-\theta)(\gamma-1)}{\epsilon_i(s)-1}\frac{\omega_i(s)}{\epsilon_i(s)}\frac{d\ln p(s)}{d\ln r} + (1-\alpha_i)$$

Define $\Omega = \frac{(\gamma-\theta)(\gamma-1)}{\epsilon_i(s)-1}\frac{\omega_i(s)}{\epsilon_i(s)}$, so that

$$\begin{aligned}
\frac{d\ln p_i(s)}{d\ln r} =& \frac{1-\theta_i}{1+\Omega_i} + \frac{\Omega_i}{1+\Omega_i}\frac{d\ln p(s)}{d\ln r} \\
\frac{d\ln \epsilon_i}{d\ln r} =& (\gamma-\sigma)(\gamma-1)\frac{\omega_i(s)}{\epsilon_i(s)}\left(\frac{1-\theta_i}{1+\Omega_i} + \frac{\Omega_i}{1+\Omega_i}\frac{d\ln p(s)}{d\ln r} - \frac{d\ln p(s)}{d\ln r}\right) \\
\frac{d\ln \epsilon_i(s)}{d\ln r} =& \frac{(\gamma-\sigma)(\gamma-1)}{1+\Omega_i}\frac{\omega_i(s)}{\epsilon_i(s)}\left(1-\theta_i - \frac{d\ln p(s)}{d\ln r}\right)
\end{aligned}$$

Sector price

The sector price is defined as

$$\begin{aligned}
p(s) =& \frac{1}{1-\gamma}\left(\sum_{i=1}^{n(s)} p_i(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \\
\frac{dp(s)}{dr} =& \left(\sum_{i=1}^{n(s)} p_i(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}-1} (-1)\sum_{i=1}^{n(s)} p_i(s)^{-\gamma}\frac{dp_i(s)}{dr}
\end{aligned}$$

The derivative takes into account that all prices adjust to a change in the rental rate.

$$\begin{aligned}
\frac{dp(s)}{dr} =& p(s)\left(\sum_{i=1}^{n(s)} p_i(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \sum_{i=1}^{n(s)} p_i(s)^{-\gamma}\frac{dp_i(s)}{dr} \\
\frac{d\ln p(s)}{d\ln r} =& \sum_{i=1}^{n(s)}\left(\frac{p_i(s)}{p(s)}\right)^{1-\gamma}\frac{d\ln p_i(s)}{d\ln r}
\end{aligned}$$

Note, that $\left(\frac{p_i(s)}{p(s)}\right)^{1-\gamma} = \omega_i(s)$, so

$$\frac{d\ln p(s)}{d\ln r} = \sum_{i=1}^{n(s)} \omega_i(s)\frac{d\ln p_i(s)}{d\ln r}$$

Thus, the change in the sector price is a weighted sum of changes in firm prices, where the weights are the respective market shares of goods. We now insert the reaction of a firm's price into the response of the sector price:

$$\begin{aligned}
\frac{d\ln p_i(s)}{d\ln r} &= \frac{1-\theta_i}{1+\Omega_i} + \frac{\Omega_i}{1+\Omega_i}\frac{d\ln p(s)}{d\ln r} \\
\frac{d\ln p(s)}{d\ln r} &= \sum_{i=1}^{n(s)} \omega_i(s)\left(\frac{1-\theta_i}{1+\Omega_i} + \frac{\Omega_i}{1+\Omega_i}\frac{d\ln p(s)}{d\ln r}\right) \\
\frac{d\ln p(s)}{d\ln r} - \sum_{i=1}^{n(s)} \omega_i(s)\frac{\Omega_i}{1+\Omega_i}\frac{d\ln p(s)}{d\ln r} &= \sum_{i=1}^{n(s)} \omega_i(s)\frac{1-\theta_i}{1+\Omega_i} \\
\frac{d\ln p(s)}{d\ln r} &= \frac{\sum_{i=1}^{n(s)} \omega_i(s)\frac{(1-\theta_i)}{1+\Omega_i}}{1-\sum_{i=1}^{n(s)} \omega_i(s)\frac{\Omega_i}{1+\Omega_i}}
\end{aligned}$$

Next, we insert the reaction of the sectoral price into the response of a firm's price:

$$\begin{aligned}
\frac{d\ln p_i(s)}{d\ln r} &= \frac{1-\theta_i}{1+\Omega_i}\frac{\Omega_i}{1+\Omega_i}\frac{\sum_{i=1}^{n(s)} \omega_i(s)\frac{(1-\theta_i)}{1+\Omega_i}}{1-\sum_{i=1}^{n(s)} \omega_i(s)\frac{\Omega_i}{1+\Omega_i}} \\
\frac{d\ln p_i(s)}{d\ln r} &= \frac{1}{1+\Omega_i}\left(1-\theta_i+\Omega_i\frac{\sum_{i=1}^{n(s)} \omega_i(s)\frac{(1-\theta_i)}{1+\Omega_i}}{1-\sum_{i=1}^{n(s)} \omega_i(s)\frac{\Omega_i}{1+\Omega_i}}\right)
\end{aligned}$$

We then insert the reaction of the sectoral price into the response of a firm's markup:

$$\begin{aligned}
\frac{d\ln \epsilon_i(s)}{d\ln r} &= \frac{(\gamma-\sigma)(\gamma-1)}{1+\Omega_i}\frac{\omega_i(s)}{\epsilon_i(s)}\left(1-\theta_i-\frac{d\ln p(s)}{d\ln r}\right) \\
\frac{d\ln \epsilon_i(s)}{d\ln r} &= \frac{(\gamma-\sigma)(\gamma-1)}{1+\Omega_i}\frac{\omega_i(s)}{\epsilon_i(s)}\left(1-\theta_i-\frac{\sum_{i=1}^{n(s)} \omega_i(s)\frac{(1-\theta_i)}{1+\Omega_i}}{1-\sum_{i=1}^{n(s)} \omega_i(s)\frac{\Omega_i}{1+\Omega_i}}\right)
\end{aligned}$$

The markup effect depends on the size of $1-\theta_i$ relative to the weighted average of $1-\theta_i$ across all firms in sector s.

Two-countries economy

Derivation of the sector price elasticity with respect to the robot rental rate.

$$\begin{aligned}
\frac{dp(s)}{dr} = &\left(\sum_{i=1}^{n(s)} \phi_i^H(s)p_i^H(s)^{1-\gamma} + \tau^{1-\gamma}\sum_{i=1}^{n(s)} \phi_i^F(s)p_i^F(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}-1} \\
&\left(\sum_{i=1}^{n(s)} \phi_i^H(s)p_i^H(s)^{-\gamma}\frac{dp_i^H(s)}{dr} + \tau^{1-\gamma}\sum_{i=1}^{n(s)} \phi_i^F(s)p_i^F(s)^{-\gamma}\frac{dp_i^F(s)}{dr}\right)
\end{aligned}$$

$$\frac{dp(s)}{dr} = p(s) \left(\sum_{i=1}^{n(s)} \phi_i^H(s) p_i^H(s)^{1-\gamma} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) p_i^F(s)^{1-\gamma} \right)^{-1} \left(\sum_{i=1}^{n(s)} \phi_i^H(s) p_i^H(s)^{-\gamma} \frac{dp_i^H(s)}{dr} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) p_i^F(s)^{-\gamma} \frac{dp_i^F(s)}{dr} \right)$$

$$\frac{d\ln p(s)}{d\ln r} = \left(\sum_{i=1}^{n(s)} \phi_i^H(s) p_i^H(s)^{1-\gamma} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) p_i^F(s)^{1-\gamma} \right)^{-1} \left(\sum_{i=1}^{n(s)} \phi_i^H(s) p_i^H(s)^{-\gamma} \frac{dp_i^H(s)}{d\ln r} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) p_i^F(s)^{-\gamma} \frac{dp_i^F(s)}{d\ln r} \right)$$

$$\frac{d\ln p(s)}{d\ln r} = \frac{1}{p(s)^{1-\gamma}} \left(\sum_{i=1}^{n(s)} \phi_i^H(s) p_i^H(s)^{-\gamma} \frac{dp_i^H(s)}{d\ln r} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) p_i^F(s)^{-\gamma} \frac{dp_i^F(s)}{d\ln r} \right)$$

$$\frac{d\ln p(s)}{d\ln r} = \sum_{i=1}^{n(s)} \phi_i^H(s) \frac{p_i^H(s)^{1-\gamma}}{p(s)^{1-\gamma}} \frac{dp_i^H(s)}{d\ln r} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) \frac{p_i^F(s)^{1-\gamma}}{p(s)^{1-\gamma}} \frac{dp_i^F(s)}{d\ln r}$$

$$\frac{d\ln p(s)}{d\ln r} = \sum_{i=1}^{n(s)} \phi_i^H(s) \omega_i^H(s) \frac{dp_i^H(s)}{d\ln r} + \tau^{1-\gamma} \sum_{i=1}^{n(s)} \phi_i^F(s) \omega_i^F(s) \frac{dp_i^F(s)}{d\ln r}$$

3.A.2. Empirical Analysis

Table 3.A1: Automation and sales - Quintile regressions

	Sales-quintiles		Markup-quintiles	
	(1)	(2)	(3)	(4)
1. Quintile × Stock of robots p.w.	-0.337**	-0.336**	-0.092**	-0.056
	(0.15)	(0.14)	(0.04)	(0.03)
2. Quintile × Stock of robots p.w.	-0.125**	-0.117*	-0.053*	-0.026
	(0.06)	(0.06)	(0.03)	(0.02)
3. Quintile × Stock of robots p.w.	0.038**	0.052***	0.027	0.053***
	(0.01)	(0.02)	(0.02)	(0.01)
4. Quintile × Stock of robots p.w.	0.186**	0.225***	0.086**	0.103***
	(0.07)	(0.07)	(0.03)	(0.03)
5. Quintile × Stock of robots p.w.	0.369**	0.408***	0.115**	0.123*
	(0.15)	(0.13)	(0.05)	(0.06)
Observations	11641	10389	11618	10365
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Additional Controls		✓		✓
KP F-Statistic	.832	1	.843	.998

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. Quintiles are based on firms' sales in the previous year in columns 2 and 3 and on firms' markups in the previous year in columns 4 and 5. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

Table 3.A2: Automation and Markups - Translog function

	(1)	(2)
Stock of robots p.w.	-0.149***	
	(0.02)	
1. Quintile × Stock of robots p.w.		-0.009**
		(0.00)
2. Quintile × Stock of robots p.w.		-0.005**
		(0.00)
3. Quintile × Stock of robots p.w.		-0.002
		(0.00)
4. Quintile × Stock of robots p.w.		0.003
		(0.00)
5. Quintile × Stock of robots p.w.		0.007*
		(0.00)
Observations	2842	10389
Country × Sector Dummies	✓	✓
Country × Year Dummies	✓	✓
Sector × Year Dummies	✓	✓
Controls	✓	✓
Additional Controls	✓	✓
KP F-Statistic	57.5	1

Notes: In this table, markups are estimated with a translog function. Standard errors, in parentheses, are two-way clustered on the country and sector level. Quintiles are based on firms' sales in the previous year. The sample consists of manufacturing sectors only. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

Table 3.A3: Automation and markups - industry-level quintile regressions

	Food, beverages and tobacco	Textiles and leather products	Other manufacturing	Paper and printing	Chemicals and pharmaceuticals	Rubber and plastic	Mineral products
1. Quintile ×	0.063	-0.616	-0.012	0.191	-0.016	-0.015	-0.084
Stock of robots p.w.	(0.17)	(1.78)	(0.06)	(0.48)	(0.01)	(0.01)	(0.13)
2. Quintile ×	0.058	-0.707	-0.007	0.201	-0.009	-0.013	-0.049
Stock of robots p.w.	(0.17)	(2.02)	(0.05)	(0.50)	(0.01)	(0.01)	(0.14)
3. Quintile ×	0.076	-0.388	0.024	0.315	-0.009	-0.010	-0.035
Stock of robots p.w.	(0.16)	(2.02)	(0.05)	(0.51)	(0.01)	(0.01)	(0.13)
4. Quintile ×	0.122	0.065	0.038	0.422	-0.003	-0.001	0.027
Stock of robots p.w.	(0.17)	(1.89)	(0.05)	(0.51)	(0.01)	(0.01)	(0.15)
5. Quintile ×	0.152	0.439	0.090	0.518	0.002	0.008	0.084
Stock of robots p.w.	(0.17)	(1.75)	(0.06)	(0.50)	(0.01)	(0.02)	(0.15)
Observations	1152	732	792	756	1196	416	527
	Basic metals	Fabricated metals	Computer electronics	Electrical equipment	Machinery and equipment	Automotive	Other vehicles
1. Quintile ×	-0.073	0.089	-0.055**	-0.034	-0.062	-0.000	-0.998
Stock of robots p.w.	(0.04)	(0.09)	(0.02)	(0.09)	(0.06)	(0.01)	(1.19)
2. Quintile ×	-0.063	0.099	-0.041**	-0.030	-0.052	-0.001	-0.926
Stock of robots p.w.	(0.04)	(0.10)	(0.01)	(0.09)	(0.06)	(0.01)	(1.14)
3. Quintile ×	-0.054	0.096	-0.046**	-0.023	-0.043	0.000	-0.991
Stock of robots p.w.	(0.04)	(0.09)	(0.02)	(0.09)	(0.06)	(0.01)	(1.22)
4. Quintile ×	-0.027	0.103	-0.023	-0.014	-0.027	0.002	-0.894
Stock of robots p.w.	(0.05)	(0.09)	(0.02)	(0.09)	(0.06)	(0.01)	(1.13)
5. Quintile ×	-0.020	0.113	0.004	0.005	0.007	0.007	-0.401
Stock of robots p.w.	(0.05)	(0.09)	(0.03)	(0.09)	(0.06)	(0.01)	(1.02)
Observations	675	545	875	982	1048	413	295
Country Dummies	✓	✓	✓	✓	✓	✓	✓
Year Dummies	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓
Additional Controls	✓	✓	✓	✓	✓	✓	✓

Notes: Standard errors, in parentheses, are clustered on the country level. Quintiles are based on firms' sales in the previous year. All specifications include country dummies and year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

Table 3.A4: Trade-weighted foreign automation and markups

	Import weighted		Export weighted	
	(1)	(2)	(3)	(4)
Stock of robots p.w.	-0.052***	-0.053***	-0.054***	-0.052**
	(0.01)	(0.02)	(0.02)	(0.02)
Foreign weighted robots	-0.035***	-0.039***	-0.021***	-0.025***
	(0.00)	(0.01)	(0.00)	(0.01)
Export weighted foreign robots	0.032**	0.025		
	(0.01)	(0.02)		
Import weighted foreign robots			0.018*	0.011
			(0.01)	(0.01)
Observations	3354	2843	3354	2843
Country × Sector Dummies	✓	✓	✓	✓
Country × Year Dummies	✓	✓	✓	✓
Sector × Year Dummies	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Additional Controls		✓		✓

Notes: Standard errors, in parentheses, are two-way clustered on the country and sector level. All variables are mean standardized to a standard deviation of one. All specifications include country sector dummies, country year dummies and sector year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

Table 3.A5: Foreign automation and markups - industry-level quintile regressions

	Food, beverages and tobacco	Textiles and leather products	Other manufacturing	Paper and printing	Chemicals and pharmaceuticals	Rubber and plastic	Mineral products
1. Quintile ×	-0.629	-15.527***	-0.121	-3.494	0.064	-0.066	-2.618**
Foreign weighted robots	(0.48)	(3.70)	(0.35)	(4.00)	(0.59)	(0.09)	(0.79)
2. Quintile ×	-0.514	-24.368***	0.367	-6.541*	-0.513	0.006	-2.752**
Foreign weighted robots	(0.47)	(6.10)	(0.27)	(3.42)	(0.58)	(0.09)	(0.92)
3. Quintile ×	-0.580	-9.853***	0.226	-4.912	-0.336	0.103	-3.403***
Foreign weighted robots	(0.44)	(2.97)	(0.34)	(2.85)	(0.57)	(0.09)	(0.71)
4. Quintile ×	-0.658	2.157	0.273	-0.463	-0.120	0.103	-2.290*
Foreign weighted robots	(0.41)	(13.34)	(0.23)	(3.00)	(0.55)	(0.10)	(1.06)
5. Quintile ×	-0.245	26.492**	0.005	0.777	0.153	0.288**	2.153
Foreign weighted robots	(0.38)	(10.51)	(0.48)	(2.55)	(0.51)	(0.10)	(3.95)
Observations	1152	732	792	756	1196	416	527
	Basic metals	Fabricated metals	Computer electronics	Electrical equipment	Machinery and equipment	Automotive	Other vehicles
1. Quintile ×	0.249	0.167	-0.118	0.101	0.268	-0.001	-1.884*
Foreign weighted robots	(0.20)	(0.12)	(0.09)	(0.07)	(0.58)	(0.02)	(0.96)
2. Quintile ×	0.082	0.253*	-0.141*	0.091	0.016	0.001	-2.217
Foreign weighted robots	(0.15)	(0.14)	(0.07)	(0.06)	(0.64)	(0.01)	(1.27)
3. Quintile ×	0.134	0.243	-0.034	0.104*	0.098	-0.007	-0.917
Foreign weighted robots	(0.16)	(0.14)	(0.05)	(0.05)	(0.54)	(0.01)	(1.22)
4. Quintile ×	0.190	0.265	0.003	0.048	0.408	-0.007	-0.244
Foreign weighted robots	(0.23)	(0.16)	(0.04)	(0.07)	(0.63)	(0.01)	(0.93)
5. Quintile ×	0.068	0.367**	0.044	0.014	1.704***	0.010	1.699**
Foreign weighted robots	(0.31)	(0.16)	(0.06)	(0.10)	(0.49)	(0.01)	(0.60)
Observations	675	545	875	982	1048	413	295
Country Dummies	✓	✓	✓	✓	✓	✓	✓
Year Dummies	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓
Additional Controls	✓	✓	✓	✓	✓	✓	✓

Notes: Standard errors, in parentheses, are clustered on the country level. Quintiles are based on firms' sales in the previous year. All specifications include country dummies and year dummies. Controls are the logs of net exports and industry production. Additional controls are the log number of patents and the log capital stock. Regressions run from 1995 to 2016.

Chapter 4

Expanding the industrial automation data universe: Prices, Production, Trade Flows

Abstract

Empirical research on industrial automation is often limited by data availability. This paper addresses these data limitations by utilizing industrial robot trade data from the Comtrade database to extract time series data on robot prices and panel data on countries' robot production. Our analysis also explores the global industrial robot market, examining concentration and stability trends over time, comparing countries' comparative advantages and market shares, and investigating potential country-level specialization. To address the issue of missing data in the Comtrade database, we employ an imputation algorithm precisely calibrated to our problem at hand. Our findings reveal that a few exporting countries dominate the industrial robot market, and that concentration is relatively stable over time. Moreover, our novel price data indicate a decline in inflation-adjusted robot prices over time, even without adjusting for the growth in robot capacity. To show potential further applications of our data, we develop an instrumental variable based on the prices of robots to reproduce prior research that had relied on proxies due to the unavailability of price data. By employing standard metrics, we demonstrate that this new instrument, which is arguably more closely connected to underlying theory, is sound for use in regression analysis and confirms the outcomes of the replicated study.

This chapter is joint work with Néstor Duch-Brown (Joint Research Centre of the European Commission, Seville). We are grateful for the comments received at the CORA 2022 - Conference on Robots and Automation, Frankfurt, Germany.

4.1. Introduction

In recent years, the multifaceted effects of industrial automation have been discussed extensively in the literature of international economics (see, e.g., the different assumptions and findings on automation induced net effects on job displacement versus creation in Prettner & Strulik 2017; Frey & Osborne 2017; Acemoglu & Restrepo 2020; G. Graetz & Michaels 2018; Aghion et al. 2020, 2022). Empirical research in the field is however limited by available data on the adoption of industrial robots. The majority of studies investigating the effects of global automation use data from the International Federation of Robotics (IFR, Müller & Kutzbach, 2019). While the IFR data covers countries' yearly stocks and installations of industrial robots across industries, other dimensions of global robot adoption remain unrepresented. In this paper we aim to expand the data universe of international robot adoption by processing and analyzing mostly Comtrade data on international trade of industrial robotics.

The result of our data expansion strategy is an origin-destination-matrix of industrial robotics covering 64 countries over the 1996-2018 period subject to no missing entries, from which we discern various descriptive statistics such as countries' shares in the international industrial robot market, overall market concentration and new data such as yearly average unit prices. Our novel data is perfectly integrable with the commonly used IFR data and therefore expands the available data on industrial robotics. Besides price data, we derive novel country-year-level data on industrial robot production by combining data on robot trade and data on robot installations.

As will be shown, the raw Comtrade data pose some challenges, mainly due to missing observations. The first part of this paper therefore deals with the problem of missingness by employing a multiple-imputation algorithm called `Amelia` (Honaker et al., 2011), which imputes missing observations in the Comtrade data. It does so by drawing supportive information from related data sources, such as the IFR data on installations, and imputes missing observations also amongst those.

Comtrade does not provide information about the destination industry of reported trade flows. Therefore, it is not possible to track shipped robots to specific industries in the importing countries such as the automotive industry using only Comtrade data. Since a sectoral dimension in the destination of robot trade would allow for finer-grained analyses, we suggest a derivation of industry weights for trade flows adding a sectoral dimension to the reported trade flows. Specifically, we extract weights from the OECD inter-country input-output tables (OECD, 2021), with which we distribute yearly bilateral country flows of industrial robots to the sectors available in the IFR data in the destination country.

We further demonstrate how the newly compiled data can contribute to the literature in the field of global industrial automation by opening up new avenues for empirical analysis. The availability of prices and quantities enables the modelling of derived demand, allows for trade analyses, and may be used in other contexts. In particular, a consecutive time series of world robot prices has been known to be a missing component in empirical

analysis since theoretical models often include the cost of automation, which so far were not straightforward to capture empirically. We replicate a part of the empirical analysis of related work by Haarburger et al. (2023), which features a theoretical model involving the price of robots. While the authors previously could only use a proxy variable depicting firms' easier access to industrial robots driven by declining robot prices, we can now use robot unit world prices to re-examine their results and more accurately fit the empirical analysis to the theoretical predictions of the model. We argue that a relatively greater plausibility of the exclusion restriction and the linking of theory and empirics improves on the previous analysis and illustrates how the novel data can be used in the future.

4.2. Data sources

In the following we introduce the data sources used in the scope of this paper. The literature, and also the available data in the field of automation by means of robotization, can be broadly divided into two main categories. Firstly, there is panel data available on industrial robots at the country-, sector- and application level. This data tends to be more relevant for macroeconomic research. Secondly, there is comparatively less data available on service robots, commonly utilised in domestic settings, which tends to make them more relevant in a microeconomic context. This paper focuses only on industrial robots and is related to the macroeconomic literature.

4.2.1. Comtrade data

We extract trade flow data from the Comtrade database (UN, 1990). At the 6-digit HS level, Comtrade provides inter-country trade volumes on industrial robotics.[1] While the base unit in which all trade flows in Comtrade are reported is current U.S. dollars, information on the number of traded units of robots, and the weight of trade volumes in kg are also provided. A trade flow in a given year *only* appears in the Comtrade data base, if and only if a trade volume in current USD is reported by a reporting country. Reporting countries can be either importing or exporting countries. For the majority of exporter-importer-year combinations, two observations are available, one reported by the exporter, one reported by the importer. Thus, missing values of trade volumes in USD within the data only exist in cases where one of the reporting countries reports a trade flow, while the other does not. In cases where both countries do not report a trade flow, there is no entry for the respective exporter-importer-year combination and thus no missing value in the classical sense. In contrast, the data on the number of units shipped and total weight of the shipment is less consistent within the reported data, as illustrated by Figure 4.1. The absence of entries also implies that for some country-year combinations reported trade flows in Comtrade are not necessarily symmetric, i.e., for a given year exports in a certain HS-category from country

[1]The 6-digit HS category including industrial robots is "847950, machinery and mechanical appliances; industrial robots n.e.c. or included".

A to country B might be reported, while imports of country A from country B in the same category might not be. The primary reason for the absence of entries is that zero trade flows are only reported in very rare cases and otherwise left unreported. Whether zero trade flows are reported or not most likely depends on the reporting countries' trade flow documentation methodologies employed by the respective statistical offices. We assume that the absence of an entry for an exporter-importer-year combination generally implies that trade flows were zero. In a later section we argue that this assumption is supported empirically by showing that for countries that have close to or zero robot production, the sum of their imports corresponds relatively well to the reported installations in any given year in the IFR data. In other words, assuming unreported trade flows to correspond to zero seems not to cause an underestimation of shipped volumes using IFR data as the benchmark.

Since we are interested in putting trade flow data into context with existing data on industrial robots, primarily installations reported in robot units by the IFR, we impute the missing unit observations. Moreover, the combination of trade volumes in current USD and in the number of units shipped allows for the calculation of unit prices per shipment, which can be aggregated or averaged in manifold ways. Information on unit prices are otherwise largely unavailable, mainly due to manufacturers' data accessibility policies. The provision of unit prices over time, differentiable by exporter- and importer-level, robot weight groups, real- and nominal values, country of origin, etc., therefore is one of our major contributions to the data universe on industrial robots.

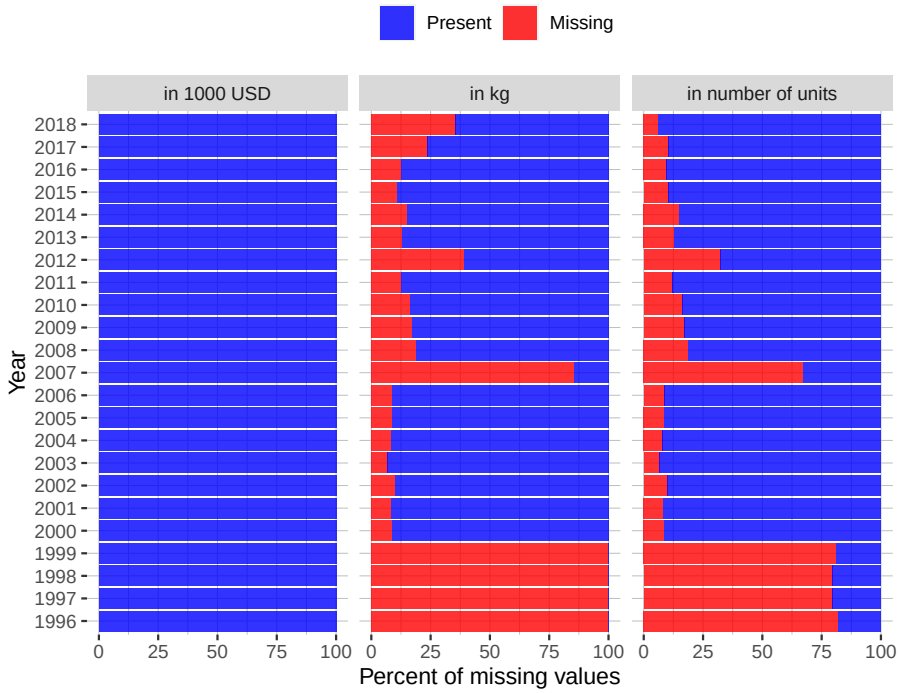
4.2.2. Other data sources

The imputation algorithm we use incorporates covariates that help to estimate missing observations. Further details on the imputation strategy including feature selection are given in section 4.4. The candidate covariates are selected based on economic theory and data availability. More precisely, we add candidate covariates from various publicly accessible databases, which are commonly associated with the use of robots, as, e.g., a country's share of manufacturing in total value added in a given year. In the following we discuss what covariates we extract from what sources to include in the analysis.

IFR. The IFR provides data on the number of yearly installed robot units and the total operating robot stock by country, industry, year and application (Müller & Kutzbach, 2019). The available industries are derived from the International Standard Industrial Classification (ISIC) system and can thus be related to other data sources using the ISIC system. Since Comtrade only reports data on country linkages, the industry dimension provided by the IFR cannot be matched with Comtrade data alone. In section 4.3 we describe our approach to impose a sectoral dimension on the Comtrade data deriving weights from OECD input-output tables.

OECD. We use the OECD key indicators to extract "Industrial production, seasonally

Figure 4.1: Missing value structure in Comtrade industrial robot trade flows amongst selected 64 countries in 1000 USD, in kg and in number of units, 1996-2018.



adjusted" and "Total manufacturing, seasonally adjusted" on the country-year level (OECD, 2015). Additionally, we extract "GDP, expenditure approach" from the OECD national accounts data (OECD, 2023a). For deriving sectoral weights which we use to add a sectoral dimension on the importer side of trade flows, we use the OECD inter-country input-output tables (OECD, 2021).

Larch RTA-DB. Mario Larch's Regional Trade Agreements Database from Egger & Larch (2008) provides various bilateral indicators as whether countries share a currency unit, have a free trade agreement, etc., and are thus expected to contribute to explaining bilateral trade flows.

TRAINS. We construct a binary bilateral variable that indicates whether an importing country imposed a tariff on products in the HS-6 category of industrial robots in a given year towards an exporting country. The base for this binary variable are MFN-tariffs from the UNCDAT TRAINS database (UNCDAT, 2018). The binarisation corresponds better to the required distributional characteristics of variables of the imputation algorithm.

ILO. We extract yearly data on the total labor force in a given country, which we use to compute output-labor-ratio (ILOSTAT, 2022).

4.3. Matrix construction strategy

As described above, the Comtrade data provide exporter- and importer-reported values for the same observations, so that the trade flows over time can be seen either through the lens of exporters or of importers, respectively. Whether to use the exporter-reported values, the importer-reported values or averages of the two is a re-occurring issue in the economic trade literature. For the sake of deriving traded robot units, we argue that using exporter-reported units will result in less biased data, because robot-producing and hence exporting countries are a relatively few highly advanced economies subject to mostly tight-knit reporting regulations imposed by local statistical offices. Constructing traded unit averages involving importer-reported units carries the risk of underestimating traded units, since many importers tend to report few or no traded units, even when reporting substantial trade flows in current USD. This observation suggests that there is a systematic difference between countries in the reporting of traded robot units, while there seems to be no such systematic difference in the reporting of trade volumes in current USD. An overestimation bias in exporter-reported trade units on the other hand seems unlikely, emphasizing the argument of using exporter-reported data on units. The same reasoning applies to reported trade flows in weight and thus we apply the same approach to prioritize exporter-reported weights.

However, if we were to limit the analysis to the exporter data only, we would potentially neglect entire country-year trade linkages, even though the importer data might contain entries different from zero for such linkages. This is because, as described above, Comtrade does not provide entries for exporter-importer-year combinations for which the reporting country did not report a trade flow in current USD, resulting in an unbalanced panel overall. Further, we address these completely missing entries as follows.

We code the respective trade flows as zeros, so that no imputation is performed in these cases. A balanced panel structure of the input-output-matrix is thus established by filling in zero flows for combinations for which no data entries are available. The number of such cases however can be reduced by about 10% by combining exporter-reported and importer-reported data on trade flows vis-a-vis using exporter-reported data only. Thus, we draw additional information on trade volumes in current USD from the importer-reported data by averaging exporter- and importer-reported values in cases in which both are reported and use the one available in cases where the other is missing.

Table 4.A1 indicates that the imputed origin-destination-matrix explains the majority of robot installations reported by the IFR, which we interpret as supportive evidence that systematic under-reporting in the national trade statistics data is not more pronounced than in the IFR data. Since the IFR installation data are based on sales data provided by robot manufacturers, as are the consolidated data reported by countries' statistical offices in Comtrade, a possible general systematic under- or over-reporting by robot manufacturers cannot be analyzed.

Since Comtrade reports inter- and no intra-country flows of robots, we can only derive the number of robots being produced and installed in a given country ex-post by subtracting total reported imports in a year from installations in that year. Analogously, we derive country-year level robot production values adding exports to the number of robots produced and installed in a given country and year. Our data do not allow to control for inventory effects, which potentially bias installation data, when they are derived from shipment data. Shipped robots could be installed in a later time period than delivered or in rare cases may even never be installed. However, as the IFR data are derived from manufacturer reported shipments, they do not account for inventory-effects either and thus no systematic discrepancy between Comtrade and IFR data in terms of inventory-effects should be expected. Unfortunately, the noise caused by potential inventory-effects in installations and all data derived from installations can not be addressed using the data sources available to us.

Comtrade does not include a sectoral dimension of trade flows. In order to investigate the world robot market along the sectoral dimension in subsection 4.5.4, we introduce sectoral variation using export weights of machinery and equipment from the OECD inter-country input-output tables (OECD, 2021). Industrial robots as defined in the 6-digit harmonized system (HS) represent only a fraction of all products belonging to sector 28 titled "machinery and equipment" as defined in ISIC rev. 4. However, we assume that the coarser weights serve as acceptable approximations for the unobservable, finer-grained industrial robot specific weights. We create exporter-importer-sector-year level weights according to

$$w_{eits} = \frac{V_{eits=28}}{\sum_{s=1,\dots,S} V_{eits}}, \tag{4.1}$$

where w_{eits} represents the weight w of flows (volume) of machinery and equipment V from exporter e to importer i's sector s in year t. By construction, the weights sum to 1 on the exporter-importer-year level.

We multiply these exporter-importer-sector-year level weights with the exporter-importer-year flows from our imputed origin-destination-matrix

$$\varrho_{eits} = w_{eits}\rho_{eit}, \tag{4.2}$$

where ϱ_{eits} represent the obtained yearly sectoral robot trade flows and ρ_{eit} aggregated yearly robot trade flows from the imputed origin-destination-matrix.

4.4. Imputation

As shown by Table 4.A1, no information on the number of traded robot units is available for 21.4% of observations. We impute these missing values using the R-package `Amelia` (Honaker et al., 2011), which includes a bootstrapping based multiple imputation program

applicable to cross-sectional, time-series or panel data. Our data categorizes as panel data, since it comprises cross-sectional bilateral country linkages over the time span 1996-2018. `Amelia` uses a bootstrapping expectation-maximization algorithm, which means that the conventional expectation-maximization algorithm (EM, Moon, 1996) is applied to multiple bootstrapped samples of the observed, incomplete data. On each bootstrapped sample the parameters are estimated and imputations for the missing values generated. The result are multiply imputed values for each missing entry, the variation of which reflect the uncertainty inherent to the imputations. The possibility of taking uncertainty into account distinguishes multiple imputation from single imputation.

An underlying assumption of `Amelia` is that the input data are multivariate normal. What differentiates `Amelia` from other imputation algorithms is that it mixes theories of inference by combining a Bayesian approach with a bootstrapping approach. More precisely, instead of drawing the first and second moments of the multivariate distribution describing the input data from their posterior density, they are estimated using a bootstrapping approach. The EM algorithm is applied to each bootstrapping sample drawn from the observed data with replacement, so that point estimates of the first and second moments are retrieved for each of these samples. For each sample, and thereby each set of moment point estimates, the observed data is used to impute the missing values. Moreover, `Amelia` allows for observation specific priors and bounds, with bounds being fixed minima and maxima for the imputations. Observation level priors can be used to improve imputation accuracy when observation specific information is available that is not straightforward to include as a covariate. An example for this sort of information is expert knowledge. Similarly, bounds impose restrictions on the imputations in cases in which it is clear that the missing values can not lie outside a certain range.

We add candidate covariates to the Comtrade data in order to improve imputation accuracy by incorporating additional potentially relevant variation. Specifically, we include importer-sided country-year level variables, whose relevance we test with a machine learning approach. Using only complete cases, i.e., rows of observations without missing values, we impose varying random patterns of missing values for which we then impute. This allows us to compare average mean squared imputation errors for different parameterizations of the algorithm and different sets of added covariates as well as the use of leads and lags of such variables.

For feature selection, i.e., for the selection of covariates to include in the imputation, we first compute the average mean squared error of traded robot quantities using all candidate variables. We then omit single candidate variables and observe changes in the mean squared error compared to the benchmark including the full set. We thus depart from the more standard approach to select variables based on improvements of the prediction error and rather choose amongst candidate variables based on negative selection. More specifically, if omitting a variable on average decreases the mean squared error by more than one standard deviation compared to the benchmark average mean squared error, we consider it detrimental to the imputation and omit it in the final imputation. We employ this negative

selection strategy since the `Amelia` imputation algorithm in general benefits from more available information and the authors advocate adding all potentially relevant data. Thus, our approach is more of an assurance that the variables we add in fact do not degrade imputation accuracy.

None of our candidate covariates significantly worsens the mean squared imputation error and thus all are kept for the final imputation. In addition to the Comtrade trade flow variables in current USD, weight in kg and units, we employ the importer's average yearly unit price, which is also derived from the Comtrade data. Moreover, we use a set of importer-year level covariates that we expect to provide relevant variation for explaining robot installations and imports. As described above, `Amelia` imputes over the whole data set and thus also imputes the 10.1% missing observations in the IFR installations as shown by Table 4.A1. Besides the importers' gross domestic product, their value added in manufacturing as a percentage of GDP and the importers robot installations from the IFR data, we add importers' GDP to labor force ratio, which we name labor productivity.

Moreover, including leads and lags of all candidate covariates generally improves the imputation accuracy and therefore is a preferred option in any specification. Apart from the option to increase leads and lags of covariates, `Amelia` also has an option to include binary covariates in the imputation, which are specified separately and handled accordingly. We include a set of dyadic binary covariates relating to international trade between countries, like joint membership in a currency unit, common free-trade-agreements, etc. An overview is presented in Table 4.A2 in the appendix.

Additionally, we include exporter-time fixed effects, which account for exporter idiosyncratic variation over time and are thus chosen to improve the estimation. Moreover, `Amelia` includes the option to impose polynomials of cross-section specific time trends. Time dynamics are a common obstacle in the imputation of panel data. As Honaker et al. (2011) demonstrate, imposing fixed polynomials for time trends can greatly reduce the uncertainty of the imputed values. Since `Amelia` does not provide any built-in optimization for choosing the best fitting order of polynomial, we again use a machine-learning optimization approach to choose amongst the available options ranging from 1, linear, to 3, cubic. The chosen polynomial is imposed for all cross-sectional time trends equally. While `Amelia` offers an option to estimate individual time trends for each cross-section, in our case this renders the imputation computationally impossible leading to an unfeasibly long run-time. Table 4.1 summarizes the hyperparameter choices.

`Amelia` requires each input variable to be normally distributed. In order to ensure that this assumption holds, we separately run all covariates through a normalization algorithm provided by the R-package `BestNormalize` (Peterson, 2021), which minimizes the difference of a covariate's distribution to a normal distribution measured by a Pearson P-Statistic comparing various common transformations such as center and scale, logarithmic, orderNorm, Yeo-Johnson, etc. After the imputation, all variables are back-transformed. Figure 4.A1 shows the densities of candidate variables before and after normalization. `Amelia`

provides us with m sets of imputed values, which we average to arrive at single imputations in order to create the descriptive statistics discussed in the following section. As the authors of `Amelia` recommend, regression analyses based on the imputed data should be run on the individual sets of imputed data sets individually and the results combined afterwards. We check the plausibility of the imputations first by comparing the densities of back-transformed imputed quantities to the density of observed quantities. Moreover, the multiple imputations can be used to calculate confidence intervals for the averaged imputed values, which give an idea of the certainty with which the imputations are conducted.

Table 4.1: Hyperparameter choices of `Amelia` algorithm.

Option	**Choice**	**Description**
Polynomial of time trends	2	Assuming quadratic time trends for all exporter-importer flows leads to the lowest mean squared imputation error.
Individual time trends	False	Fitting individual time trends for all exporter-importer pairs renders the runtime unfeasible.
Leads	All covariates	Future values are used to explain missing values.
Lags	All covariates	Past values are used to explain missing values.
Bounds:		
- Installations	$(0,\infty)$	Installations are lower bounded to zero, since negative installations are per definition not possible.
- Exported units	$(0,\infty)$	Exported units are lower bounded to zero, since negative exports are per definition not possible and reflected by imports also contained in the data matrix.
Priors:		
- Installations, mean	$\tilde{\mu}_{it}$	We assume that the sum of total reported imports of a country in a given year is a good prior for its total installations in that year. We expect this to hold especially in the case of countries that are no robot producers themselves.
- Installations, S.D.	$\tilde{\sigma}_{it}$	In order to account for the uncertainty of the mean installation prior in the case of robot producing countries, we use an importer specific prior for the standard deviation of installations in combination with the mean prior. Since we expect a larger deviation of installations in a year from total imports for robot producers, we observe the standard deviation of the difference between installations and total imports in years for which complete data is available. For robot producers, this standard deviation will be larger than for non-robot producers.

Where $\tilde{\mu}_{it} = \sum_{eit}^{e} x_{eit}$, with x_{eit} depicting exports from exporter e to importer i in year t and $\tilde{\sigma}_{it} = \min[1, sd(\mathbf{D})]$, where $\mathbf{D}$ is a vector containing the year-wise differences of summed imports of importer i and robot installations of importer i for all years without missing entries.

4.5. Descriptives of completed origin-destination-matrix

In this section we provide descriptive statistics of the completed origin-destination-matrix and visualize the unit trade flows. Figure 4.A2 illustrates pooled inter-country flows of industrial robots from 1996 to 2018. It becomes apparent that Japan is by far the largest exporter of industrial robots over the 1996-2018 time period. This is further substantiated by Figures 4.5 and 4.A5, which show the revealed comparative advantages of the 20 countries with the highest RCA over the whole sample period and the market shares of the overall largest exporters of industrial robots over time, respectively. Figures 4.A3 and 4.A4 illustrate total exports and imports as world maps over the whole time period under consideration separately.

4.5.1. Market concentration, stability and specialisation

We compute countries' market shares in the world robot market as their share in exports in a given year over total exports in that year

$$\omega_{et}^{E} = \frac{x_{et}^{E}}{\sum_{e=1,\dots,N_t} x_{et}^{E}}, \tag{4.3}$$

where ω_{et}^{E} is the market share of exporter e in year t on the export market, x_{et}^{E} are the export robot trade flows of exporter e in year t and N_t is the number of active exporters in year t. Using these market shares, we can compute the normalized Herfindahl-Hirschman-Index (Rhoades, 1993), which is a standard measure for market concentration and given by

$$HHI_{t}^{E} = \frac{\left(\sum_{e=1}^{N_t} \omega_{et}^{2} - 1/N_t\right)}{1 - 1/N_t}, \tag{4.4}$$

where N_t is the number of exporters active on the market in a given year t, and ω_{et} is the market share as defined by equation 4.3. The normalized HHI ranges from 0 to 1, with an HHI_{t}^{E} of 1 indicating that a single exporter is the sole supplier in a given year.

No clear trend emerges over the whole time span under consideration. While market concentration amongst importers and exporters seems to have rather declined from 1996 to 2009, it seems to have been rather increasing since 2009 before starting to decline again between 2017 and 2018, again amongst importers and exporters. In other words, from 2009 onward, both, few exporters and importers, possibly one respectively, increased their shares in world robot trade vis-a-vis other exporters and importers. To examine this trend reversal in more detail, we compare the evolution of market shares of the twelve countries exporting most robot units over the whole time span as depicted by Figure 4.A5.

We would expect market concentration over time measured by the normalised HHI based on import shares to resemble the same measure based on installation data. Figure 4.3 depicts the latter. Several differences emerge. First, the sharp decline before 2000 in the installation

Figure 4.2: Normalized Herfindahl-Hirsch-Index by imports and exports, 1996-2018, robot unit trade flows. Source: Comtrade and own imputation.

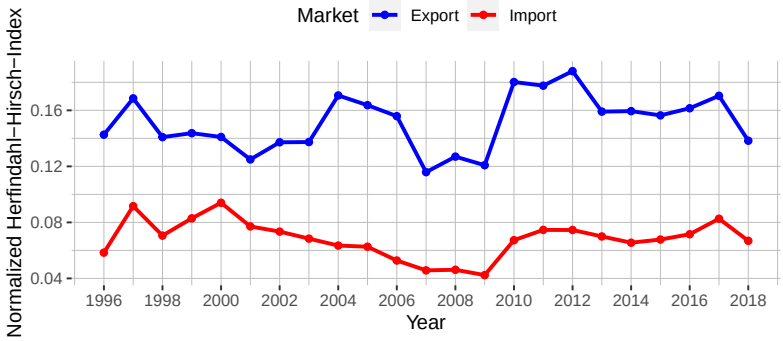
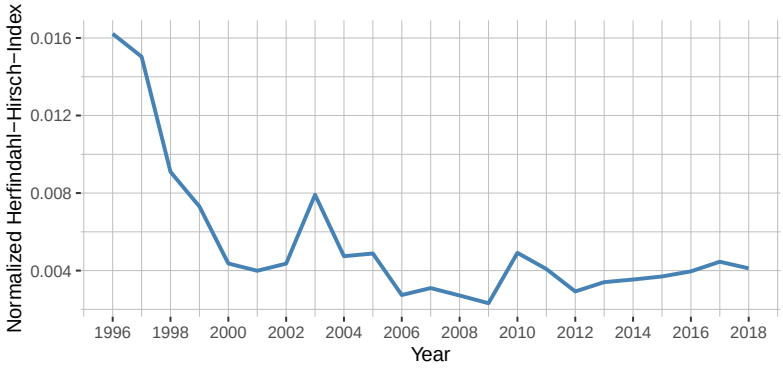


Figure 4.3: Normalized Herfindahl-Hirsch-Index, 1996-2018, robot installations. Source: IFR and own imputation.

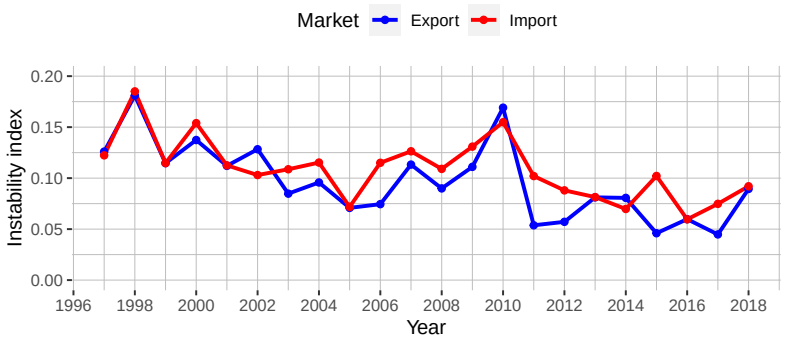


data is not visible in the import data. This might be because the IFR data collection had only started shortly before and only a relatively small fraction of robot producers had been part of the survey at this time. Thus, the high perceived concentration might be due to small sample bias in the IFR data. Second, compared to the smooth HHI time series based on import data, the installation based HHI time series is subject to higher year to year volatility. Both Figures however depict an increase in market concentration since 2009.

Japan has seen the most striking increase in market share from about 26% in 2009 to about 39% in 2010 and in direct comparison to other large robot exporters seems to mainly have driven the sharp increase in HHI observed in the aggregate data. Moreover, Figure 4.A5 illustrates Japan's consistent dominance on the export market. Besides Japan, only Korea and China are subject to clearly increasing market shares over time. A few large exporters show declining trends. The USA stands out the most being subject to a clearly negative trend in market shares. Although weaker in comparison, France and Germany are subject to declining market shares in the considered time span as well.

Similarly, we compute a market instability index, which illustrates the overall fluctuations

Figure 4.4: Instability-Index by imports and exports, 1996-2018, robot unit trade flows. Source: Comtrade and own imputation.



in market shares over time. It is given by

$$II_t^E = \frac{1}{2}\left(\sum_{e=1}^{N} |\omega_{et} - \omega_{et-1}|\right). \tag{4.5}$$

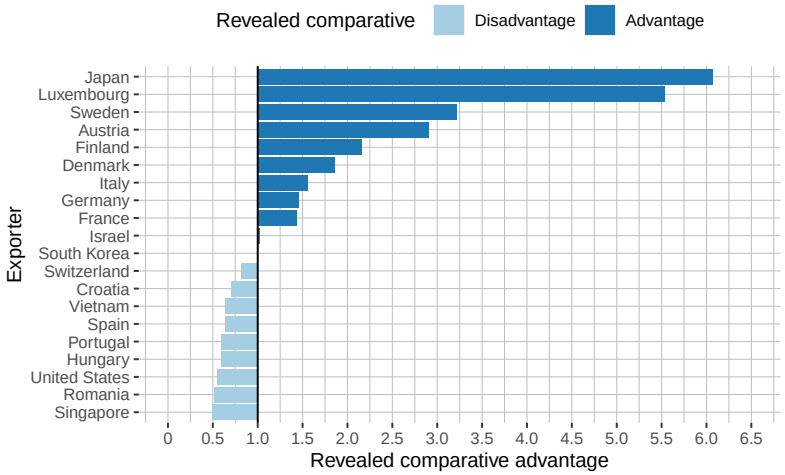
Figure 4.4 depicts the index based on importers' and exporters' market shares. The gap between importers and exporters observable in market concentration as depicted by Figure 4.2 is not observable for market instability, which reflects that the gap remains consistent over time. Moreover, instability amongst importers and exporters seems to be correlated over time. The sharp increase in Japan's market share between 2009 and 2010 is also reflected by the instability index. Interestingly, instability amongst importers increased as well between 2009 and 2010. This might be due to Japan having established trade partners that absorbed large shares of Japan's increased exports in that year.

Another metric of interest is the so called revealed comparative advantage (RCA), which is defined as

$$RCA_{es} = \frac{x_{es} / \sum_{s=1}^{S} x_{es}}{\sum_{e=1}^{N} x_{es} / \sum_{e=1}^{N} \sum_{s=1}^{S} x_{es}}. \tag{4.6}$$

A country is considered to have a revealed comparative advantage, when the ratio of its share in robot export is greater than the share of robot exports in world exports. This is the case when the nominator in equation 4.6 is larger than the denominator and the RCA thus is larger one. Figure 4.5 depicts the RCAs for the 20 countries with the largest RCAs based on pooled export data from 1996-2018 and thus reveals which countries have been specialised on producing robots over the time span under consideration. Besides Japan and Israel, all countries with an RCA larger than one are part of the EU-27. While other countries are large industrial robot producers in absolute terms, they seem to be less specialised in the robotics-industry, meaning that robot exports make up a smaller share of their total exports. This applies to China and the US for example.

Figure 4.5: Revealed comparative advantage in robot exports, trade volumes in USD, 1996-2018, shown for the 20 countries with highest RCA. Source: Comtrade and own imputation.



4.5.2. Derivation of unit prices

The combination of the number of units per shipment and the traded volume in current USD renders the calculation of unit prices possible. In general, time series price data on industrial robots is hardly available. The only time series on unit prices provided by the IFR ends in 2009. We therefore consider the provision of unit price data as a valuable contribution to the robot data landscape. Furthermore, the composition of our origin-destination-matrix allows for various break-downs of prices to geographical regions or single countries. Thus, we can observe average prices over time by exporters, which might indicate countries' specialisation on certain types of industrial robots. In general, industrial robots can differ quite substantially in size and weight. While the software provided with the robot is an important determinant of its price, larger, heavier robots require more raw materials in production and thus tend to be more expensive. Unfortunately, the data available to us does not allow for differentiating software costs from material and production costs. However, using trade volumes in weight (kg) in combination with the number of units shipped, we obtain the average weight of a robot shipped. This allows for the analysis of unit prices over time for different weights of robots. Figure 4.6 depicts yearly average robot unit prices in nominal and real, i.e. USD producer price index (PPI) adjusted, terms. In real prices, the average shipped robot unit has become less costly over time. Additionally, one would expect the productivity of a robot to have increased over the years so that the price decline would be more pronounced in efficiency units. Unfortunately, the average productivity is difficult to quantify given the data available.

Figure 4.7 depicts PPI adjusted prices for weight quantiles. Interestingly, robots of different weight quantiles seem to converge in prices over time. While the gaps in prices between the respective weight quantiles were largest at the beginning of the period under study, there seems to be convergence to the same price per unit in 2018, of about 12,000 PPI

Figure 4.6: Average robot units price in 1000 USD, nominal and real (PPI adj.), 1996-2018. Source: Comtrade and own imputations.

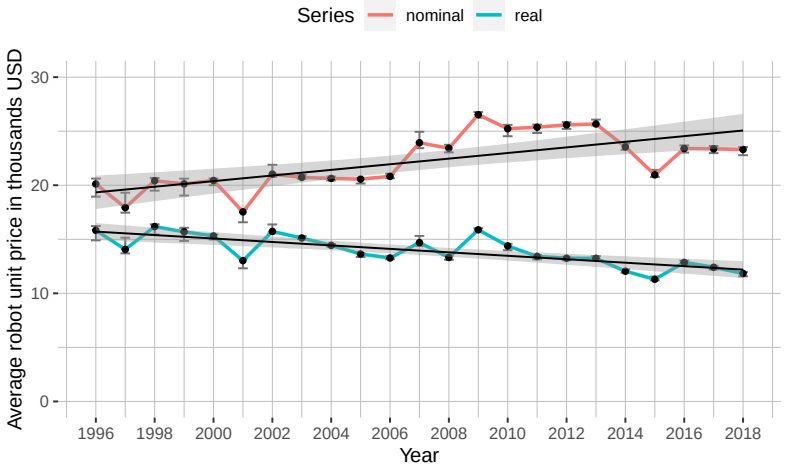
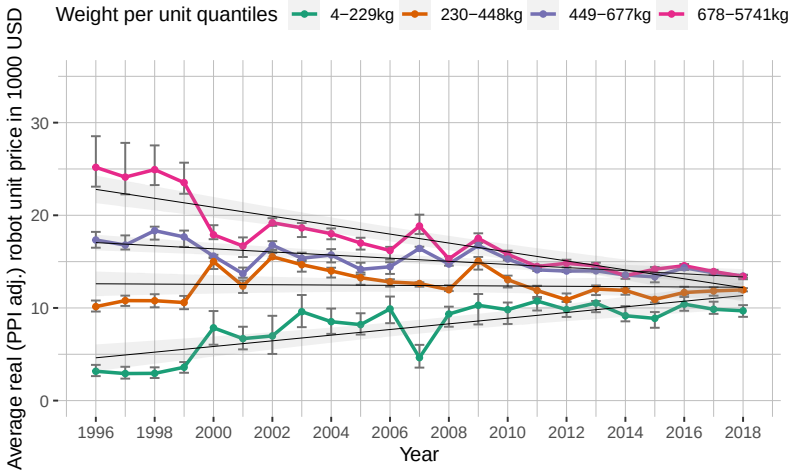


Figure 4.7: Average real (PPI adj.) robot units price in 1000 USD per weight quantile, 1996-2018. Source: Comtrade and own imputations.

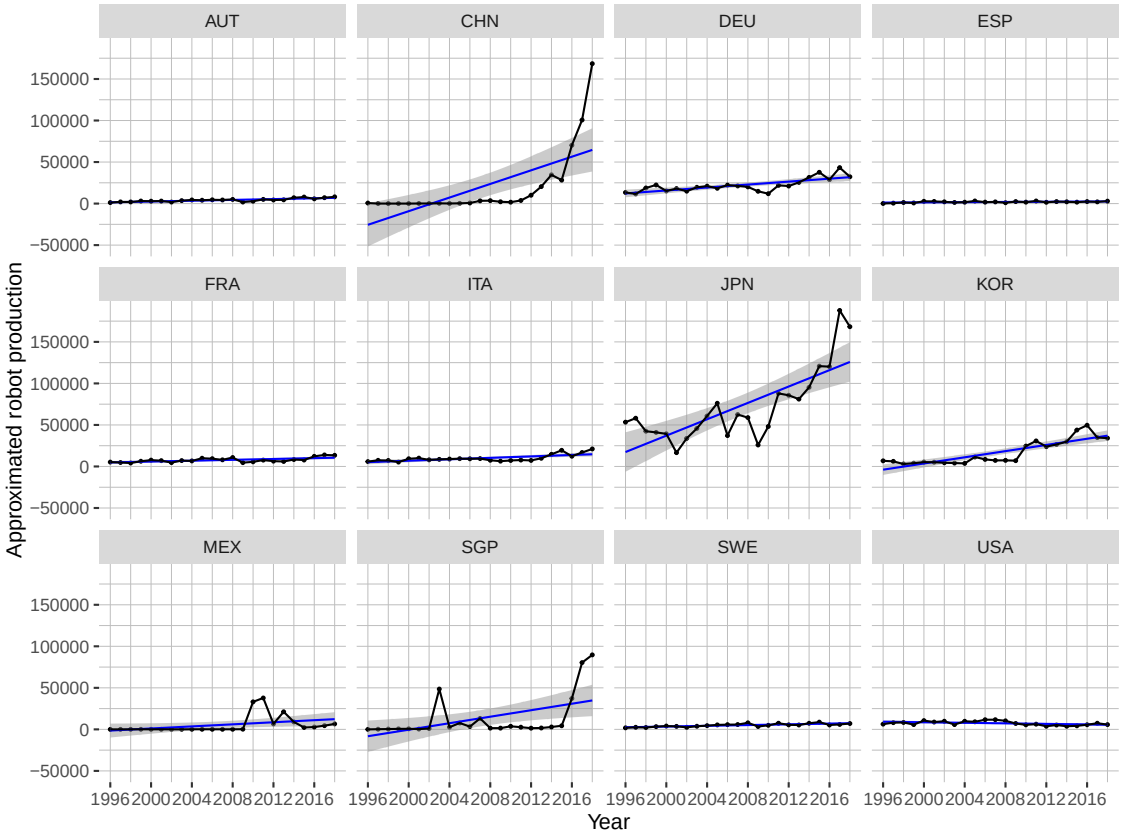


adjusted USD. One hypothesis for the convergence is that software might have become increasingly more important over time and therefore takes a larger part in the unit price. Since software development costs are probably little affected by robot weight and payload, some convergence in prices would be expected.

Figure 4.A7 shows differences in average PPI adjusted prices amongst the twelve largest exporters over time. For some countries we see clear negative trends over time. Especially in the cases of Japan and Singapore we observe strong unit price declines. While Germany and France also show a negative trend, it is less pronounced. To explore the question of whether lower average unit prices are related to lighter shipped units, we depict the median weight of a shipped robot unit in Figure 4.A8.

In the case of Japan and Singapore we do in fact see that the median shipped robot unit over time has become lighter as time goes by. While the median shipped robot unit from

Figure 4.8: Derived number of produced robot units by country, 1996-2018. Source: IFR, Comtrade, own imputations and calculations.



Japan weighed just under 600kg in 1996, it weighed about 300kg in 2018. Similarly, in the case of Singapore, the median has halved from about 400kg in 1996 to about 200kg in 2018. We observe clear upward trends for some of the other major robot exporters. Robot producers in Korea, Sweden and the US seem to have sold heavier robots over time, although there is substantial volatility across all three series and in the case of the US the trend seems to have reversed since 2011. In summary, these findings may hint to an ongoing specialisation of robot producers located in the respective countries. However, more accurate data are needed to investigate this hypothesis further.

4.5.3. Derivation of Production

The simultaneous availability of country level annual exports and installations, allows us to make inferences about country-year level production of countries. Exports represent the share of production that leaves the country and installations that exceed a country's imports represent the share of production that is installed locally. Taken together, these parts add up to a country's production in a given year. The values derived in this way for the production of robots are, of course, only rough. In some cases, the IFR reports installations in countries for which no robot imports were documented in that year in the

Comtrade database. These installations unexplained by Comtrade would be considered production according to our derivation approach. For some country-year combinations however it might appear more likely that an export reported by a robot manufacturer reported to the IFR is not included in the national trade statistics and thus the Comtrade database. This is a short-coming that needs to be considered when interpreting or making further use of the data. Figure 4.8 illustrates the derived production of robot units over the time span under consideration for the twelve countries subject to the largest production.

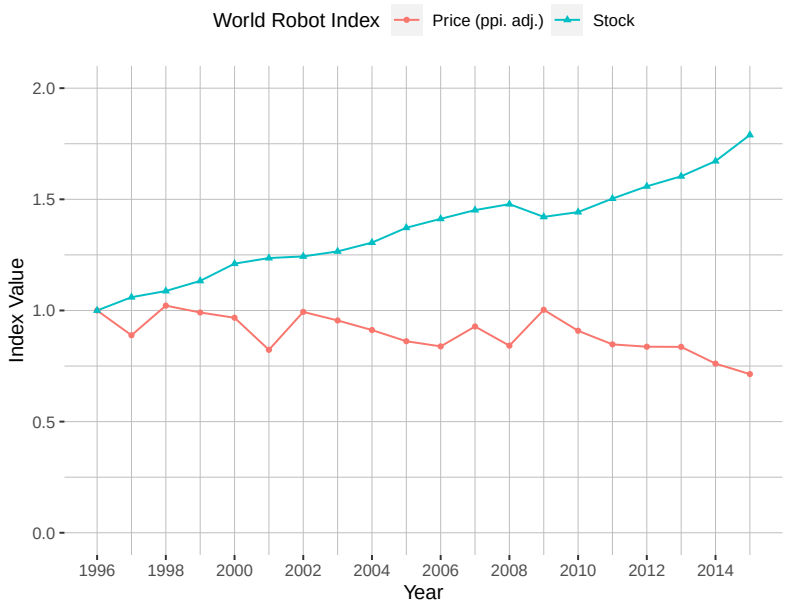
4.5.4. Sectoral analysis

As outlined above, we impose a sectoral dimension on the importer side of the origin-destination-matrix by deriving weights from OECD input-output tables. Using the sectorally distributed trade flows of robots, we can calculate descriptive statistics and indexes along the sectoral dimension. Figure 4.A9 shows the HHI index over time by sector. It suggests distinctive differences in market concentration across sectors. The sectors most prone to the use of industrial robots tend to be less concentrated than the less prone sectors. The sectors subject to the highest market concentration are the manufacture of electronic, computer and optical products and the manufacture of electrical equipment. Concentration varies greatly in the education and research sector, which may be due to the low number of robots inducing small sample bias leading to a noisy market concentration index.

4.6. Application

As outlined above, by providing consecutive time series price data for industrial robots we intend to close the gap of missing prices in the robot data universe. The general adoption of industrial robots has been studied along many dimensions. In related theory, the decision of a profit-maximizing firm to automate is usually expressed as a function of the costs to automate amongst other determinants. However, due to the lack of price data, a direct empirical implementation of such models has not been possible so far. Thus, to illustrate how the data we provide here can be used in estimations common in the automation literature, we replicate a set of estimations of related work in which the lack of price data prompts the authors to use a robot stock based proxy for robot adoption. Haarburger et al. (2023) investigate the effect of increasing robot adoption on markups combining sectoral data on the use of robots with firm-level data. The theoretical model they present suggests that firms increasingly automate due to decreasing robot prices. The empirical analysis uses a Bartik-type instrumental variable (Bartik, 1991), which consists of an exogenous time series component and an endogenous cross-sectional component. As the exogenous time series component the authors choose the world stock of robots, which they interact with the labor share in a given country and sector in 1995 and a proxy for that sector's technological advancement in 1990 from Archibugi & Coco (2004). Since large-scale industrial automation is generally regarded to have begun in the beginning of the 1990s, the authors argue that the

Figure 4.9: World robot stock and price indexes as time series components of interacted instrumental variables over time.



endogenous cross-sectional component is most likely little affected by the rise of industrial robots.

With the price data obtained in the scope of the present analysis, we can construct a similar instrumental variable, which uses producer-price-index adjusted world robot prices instead of the world robot stock as the time series component. Figure 4.9 depicts both time series as normalised indexes, i.e., as changes with respect to the first observation in 1996 to establish comparability. As expected, both time series are subject to opposing trends overall. However, a shortcoming of the price data when used as a proxy for the likelihood of a firm to adopt robots is that it does not account for the performance of robots. Based on the assumption that technological progress has made robots more performant over time, the attractiveness of robots from a firm's perspective should have increased even more than the pure decrease in prices would indicate. Due to the diversity of industrial robots, deriving a performance measure to adjust prices with is not straightforward.

The Bartik instrumental variable is constructed as

$$R_{cst}^{IV} = R_t^{WI} \frac{O_{cs}^{1995}}{L_{cs}^{1995}} I_c^{1990}, \tag{4.7}$$

where R_t^{WI} depicts a robot world index which is either the global stock of robots or global robot prices. The fraction on the right refers to the output-labor-ratio of a sector in a country in 1995. The term I_c^{1990} refers to the technological capacity of a country in 1990.

In the following we repeat the estimations from Haarburger et al. (2023) which involve the instrumental variable and contrast the results obtained using the world robot stock as the

Table 4.2: Comparing instrumental variables, first stage, manufacturing sectors only - replication of Table 3.2 in Chapter 3.

Dependent Variable:	Stock of robots p.w.	
Model:	S	P
Variables		
World stock robot IV	1.1***	
	(0.20)	
World price robot IV		-0.79***
		(0.11)
Fixed-effects		
Country-Sector	Yes	Yes
Country-Year	Yes	Yes
Sector-Year	Yes	Yes
Controls		
log(Production)	Yes	Yes
log(Net Exports)	Yes	Yes
Fit statistics		
Observations	3,388	3,388
R^2	0.96954	0.96364
BIC	4,980.5	5,580.5
F-test	137.93	114.84

Columns marked by S refer to estimations using the world robot stock instrument, columns marked by P to estimations using the world price instrument.
Clustered (Country & Sector) standard-errors in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.*

time series component with using real world robot prices instead.

Table 4.2 depicts the first stage of the following instrumental variable estimation. Columns indicated with the letter S refer to estimations using the world stock of robots, while columns indicated with the letter P analogously refer to estimations using real robot world prices. Both instrumental variables are significant at the 1% level. In line with economic theory, the sign of the price IV is negative.

Table 4.3 shows the results of the instrumental variable estimations with sectoral markups as the dependent variable. In terms of effect size and significance of the coefficients, the results are very similar across IVs. The statistic of the Wald first stage test, a test which is also referred to as the test of weak instruments, indicates the strength of the instrument used in the estimations (Wald, 1943). In general, the smaller the Wald test statistic, the weaker the instrument. Columns S.1 and P.1 show estimation results for all sectors, whereas columns S.2 and P.2 show the results for manufacturing sectors only. The Wald test statistic suggests that both instruments are stronger for manufacturing sectors. Moreover, in this specification, it suggests the price IV is stronger than the stock IV. Over all specifi-

Table 4.3: Comparing instrumental variables, IV estimation, by all sectors and manufacturing sectors only - replication of Table 3.3 in Chapter 3.

Dependent Variable:	Markups			
	All sectors		Only manufacturing	
Model:	S.1	P.1	S.2	P.2
Variables				
Stock of robots p.w.	-0.25**	-0.28*	-0.18***	-0.18***
	(0.10)	(0.15)	(0.02)	(0.03)
Fixed-effects				
Country-Sector	Yes	Yes	Yes	Yes
Country-Year	Yes	Yes	Yes	Yes
Sector-Year	Yes	Yes	Yes	Yes
Controls				
log(Production)	Yes	Yes	Yes	Yes
log(Net Exports)	Yes	Yes	Yes	Yes
log(Capital)	Yes	Yes	Yes	Yes
log(Patents)	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,676	3,676	2,863	2,863
R^2	0.94470	0.94197	0.90082	0.90122
BIC	848.25	1,025.7	-363.93	-375.63
F-test	66.082	64.009	27.441	26.102
Wald (1st stage)	3.0316	2.4884	57.521	83.444
F-test (1st stage)	712.53	116.36	1,074.1	169.20

Columns marked by S refer to estimations using the world robot stock instrument, columns marked by P to estimations using the world price instrument.
Clustered (Country & Sector) standard-errors in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.*

cations an increase in the stock of robots per worker is associated with a decrease in markups.

Table 4.4 shows results for interactions of the stock of robots p.w. with firm quintiles. Columns S.1 and P.1 show results for the interaction with sales quintiles and columns S.2 and P.2 show results for interactions with markup quintiles. Remarkably, results are similar in terms of effect size and significance between the different instrumental variables. The Wald statistic does not clearly suggest that one of the instruments is stronger.

Table 4.5 shows results for IV regressions with logarithmised production and net exports as the outcomes respectively. Again, the results are similar across both instrumental variables. In the case of production, the coefficient of the price IV regression is 50% larger compared to the stock IV regression. The signs of the coefficients for the stock of robots p.w. in columns S.2 and P.2, where logarithmised net exports are the outcome, differ. However,

due to their statistical insignificance an interpretation is not sensible.

Table 4.6 shows the estimation results for three other alternative outcomes, namely the logarithmised number of firms, logarithmised output prices and the logarithmised operating margin. Results again are consistent across IVs. According to the Wald statistic, the price IV is stronger than the stock IV in the estimation where output prices are the dependent variable. The opposite is the case in columns S.3 and P.3, where an alternative measure to markups, the operating margin is the dependent variable.

4.7. Conclusion

In this paper, we present novel data derived from the Comtrade database and data provided by the IFR. We construct a complete origin-destination-matrix of industrial robots covering the period 1996-2018 and 64 countries. Trade flows of robot units are provided in current USD, robot units and weight. Our data allows to derive robot unit prices at different levels of geographical aggregation, over time and by weight group. In addition, we derive the production of industrial robots at the country-year-level. To address the problem of incomplete data, especially for traded robot units and weight, we employ a sophisticated multiple imputation algorithm tailored to our problem at hand, which we also use to impute missing installation and stock IFR data.

Moreover, we examine the resulting data along several dimensions and find that market concentration in the global market for industrial robots has been relatively stable over time and that the market is still dominated by a few robot-producing countries. This finding is substantiated by countries' revealed comparative advantages and the development of countries' market shares over time. We also observe a trend towards specialisation across countries in terms of robot unit weights. While the robotics industry in some countries seems to specialise increasingly in lighter robots, the industry in other countries seems to take the opposite position, specialising in heavier robots. Surprisingly, however, we observe a clear conversion of prices across different weight quantiles of traded robot units.

Equipped with the new price data, we show how the empirical analysis of papers in the automation literature can now be more directly linked to economic theory, when the cost of automation is a determinant in such models. We replicate some of the empirical analysis of related work where the cost of automation had to be expressed by a proxy variable due to the lack of price data. We argue that the instrumental variable we construct based on robot prices satisfies the exclusion restriction better than the previous proxy-based IV. As this proxy is commonly used in the literature, we suggest using a price-based instrumental variable may be beneficial for related empirical analysis in the future.

We make all derived data available to other researchers.

Table 4.4: Comparing instrumental variables, IV estimation, manufacturing sectors only, by sales quintiles (S.1 and P.1) and markup quintiles (S.2 and P.2) - replication of Table 3.4 in Chapter 3.

Dependent Variable:	Markups			
Model:	S.1	P.1	S.2	P.2
Variables				
Stock of robots p.w. × Sales Quintile 1	-0.01**	-0.01**		
	(0.005)	(0.005)		
Stock of robots p.w. × Sales Quintile 2	-0.009***	-0.010**		
	(0.003)	(0.003)		
Stock of robots p.w. × Sales Quintile 3	-0.005**	-0.005**		
	(0.002)	(0.002)		
Stock of robots p.w. × Sales Quintile 4	0.004	0.005		
	(0.003)	(0.003)		
Stock of robots p.w. × Sales Quintile 5	0.01**	0.02**		
	(0.006)	(0.007)		
Stock of robots p.w. × Markup Quintile 1			-0.04**	-0.04**
			(0.01)	(0.01)
Stock of robots p.w. × Markup Quintile 2			-0.02**	-0.02*
			(0.006)	(0.008)
Stock of robots p.w. × Markup Quintile 3			-0.003	-0.0008
			(0.002)	(0.003)
Stock of robots p.w. × Markup Quintile 4			0.01**	0.02**
			(0.005)	(0.006)
Stock of robots p.w. × Markup Quintile 5			0.04***	0.04**
			(0.01)	(0.01)
Fixed-effects				
Country-Sector	Yes	Yes	Yes	Yes
Country-Year	Yes	Yes	Yes	Yes
Sector-Year	Yes	Yes	Yes	Yes
Controls				
log(Production)	Yes	Yes	Yes	Yes
log(Net Exports)	Yes	Yes	Yes	Yes
log(Capital)	Yes	Yes	Yes	Yes
log(Patents)	Yes	Yes	Yes	Yes
Fit statistics				
Observations	10,408	10,407	11,284	10,383

Columns marked by S refer to estimations using the world robot stock instrument, columns marked by P to estimations using the world price instrument.
Clustered (Country & Sector) standard-errors in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.*

Table 4.5: Comparing instrumental variables, IV estimation, manufacturing sectors only, Production and Exports - replication of Table 3.5 in Chapter 3.

Dependent Variables:	log Production		log Net Exports	
Model:	S.1	P.1	S.2	P.2
Variables				
Stock of robots p.w.	0.16***	0.24***	1.2	-0.82
	(0.02)	(0.03)	(1.5)	(2.0)
Fixed-effects				
Country-Sector	Yes	Yes	Yes	Yes
Country-Year	Yes	Yes	Yes	Yes
Sector-Year	Yes	Yes	Yes	Yes
Controls				
log(Production)	Yes	Yes	Yes	Yes
log(Net Exports)	Yes	Yes	Yes	Yes
log(Capital)	Yes	Yes	Yes	Yes
log(Patents)	Yes	Yes	Yes	Yes
Fit statistics				
Observations	2,863	2,863	2,863	2,863
R^2	0.99661	0.99622	0.84012	0.83953
BIC	185.57	501.21	24,750.7	24,761.3
F-test	1,007.7	991.67	17.058	17.052
Wald (1st stage)	59.663	49.797	58.291	84.811
F-test (1st stage)	1,148.0	193.54	1,076.0	168.70

Columns marked by S refer to estimations using the world robot stock instrument, columns marked by P to estimations using the world price instrument.
Clustered (Country & Sector) standard-errors in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.*

Table 4.6: Comparing instrumental variables, IV estimation, manufacturing sectors only, alternative outcomes: number of firms, output prices, operating margin - replication of Table 3.6 in Chapter 3.

Dependent Variables:	log Nr. of firms		log Output prices		log Operating margin	
Model:	S.1	P.1	S.2	P.2	S.3	P.3
Variables						
Stock of robots p.w.	0.42	-0.04	-0.14***	-0.14*	-0.82***	-0.74***
	(0.27)	(0.25)	(0.04)	(0.07)	(0.10)	(0.07)
Fixed-effects						
Country-Sector	Yes	Yes	Yes	Yes	Yes	Yes
Country-Year	Yes	Yes	Yes	Yes	Yes	Yes
Sector-Year	Yes	Yes	Yes	Yes	Yes	Yes
Controls						
log(Production)	Yes	Yes	Yes	Yes	Yes	Yes
log(Net Exports)	Yes	Yes	Yes	Yes	Yes	Yes
log(Capital)	Yes	Yes	Yes	Yes	Yes	Yes
log(Patents)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,303	1,303	2,231	2,231	1,960	1,960
R^2	0.89961	0.89897	0.69610	0.69556	0.87029	0.87128
BIC	5,832.9	5,841.1	1,123.7	1,127.7	9,218.0	9,203.0
F-test	23.178	23.165	5.9311	5.7984	18.327	18.216
Wald (1st stage)	22.461	20.299	81.884	137.88	24.649	12.558
F-test (1st stage)	119.67	32.573	1,236.6	59.301	381.18	123.48

Columns marked by S refer to estimations using the world robot stock instrument, columns marked by P to estimations using the world price instrument.
Clustered (Country & Sector) standard-errors in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.*

4.A. Appendix

Table 4.A1: Descriptives of non-binary covariates used in imputation and installations explained by imports in percentages

	Pre imputation (N=19500)	Post imputation (N=19500)
Trade volume in current USD		
Mean (SD)	3010 (16700)	3010 (16700)
Median [Min, Max]	248 [0, 674000]	248 [0, 674000]
Trade volume in kg		
Mean (SD)	68700 (352000)	67600 (350000)
Median [Min, Max]	6020 [0, 9310000]	5480 [0, 9370000]
Missing	5356 (27.5%)	0 (0%)
Trade volume in units		
Mean (SD)	185 (1560)	162 (1390)
Median [Min, Max]	12.0 [0, 71500]	11.0 [0, 71500]
Missing	4181 (21.4%)	0 (0%)
Average yearly price of importer		
Mean (SD)	28.7 (20.4)	28.6 (20.2)
Median [Min, Max]	24.5 [0.116, 459]	24.5 [0.116, 341]
Missing	162 (0.8%)	0 (0%)
GDP, expenditure approach		
Mean (SD)	1750000 (3280000)	1590000 (3090000)
Median [Min, Max]	560000 [3790, 20500000]	522000 [3790, 20500000]
Missing	2549 (13.1%)	0 (0%)
Value added in manufacturing as percentate of GDP		
Mean (SD)	16.3 (5.59)	16.3 (5.54)
Median [Min, Max]	15.7 [3.89, 34.9]	15.7 [3.89, 34.9]
Missing	781 (4.0%)	0 (0%)
IFR robot installations of importer (units)		
Mean (SD)	6700 (16800)	6270 (16100)
Median [Min, Max]	1380 [0, 190000]	1430 [1.00, 190000]
Missing	1960 (10.1%)	0 (0%)
Labor productivity		
Mean (SD)	58.6 (40.6)	54.8 (39.8)
Median [Min, Max]	53.6 [1.07, 264]	48.7 [1.07, 264]
Missing	2549 (13.1%)	0 (0%)
Installations explained by imports, %		
Mean (SD)	0.805 (0.292)	0.879 (0.235)
Median [Min, Max]	1.00 [0, 1.00]	1.00 [0, 1.00]
Missing	1987 (10.2%)	0 (0%)

Table 4.A2: Descriptives of binary covariates used in imputation

	Overall (N=19500)
Currency Unit	
Mean (SD)	0.0257 (0.158)
Median [Min, Max]	0 [0, 1.00]
Free-trade-agreement	
Mean (SD)	0.151 (0.358)
Median [Min, Max]	0 [0, 1.00]
Partial scope agreement	
Mean (SD)	0.0554 (0.229)
Median [Min, Max]	0 [0, 1.00]
Economic integration agreement	
Mean (SD)	0.305 (0.461)
Median [Min, Max]	0 [0, 1.00]
Tariff on HS-847950	
Mean (SD)	0.728 (0.445)
Median [Min, Max]	1.00 [0, 1.00]
Missing	2239 (11.5%)

Figure 4.A1: Transformed data compared to benchmark normal distribution with equal first and second moments.

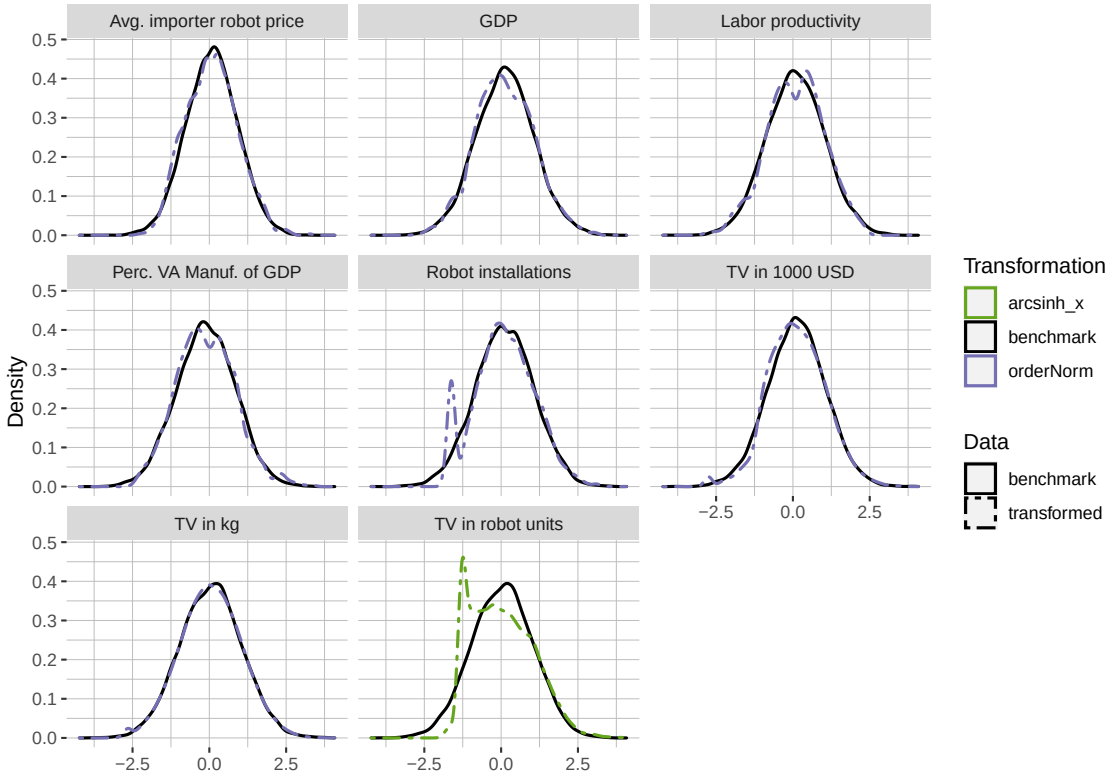


Figure 4.A2: Robot unit trade flows, pooled data for 1996-2018, intra-country trade excluded, 10 largest exporters out of 64 separate. Source: Comtrade and own imputation.

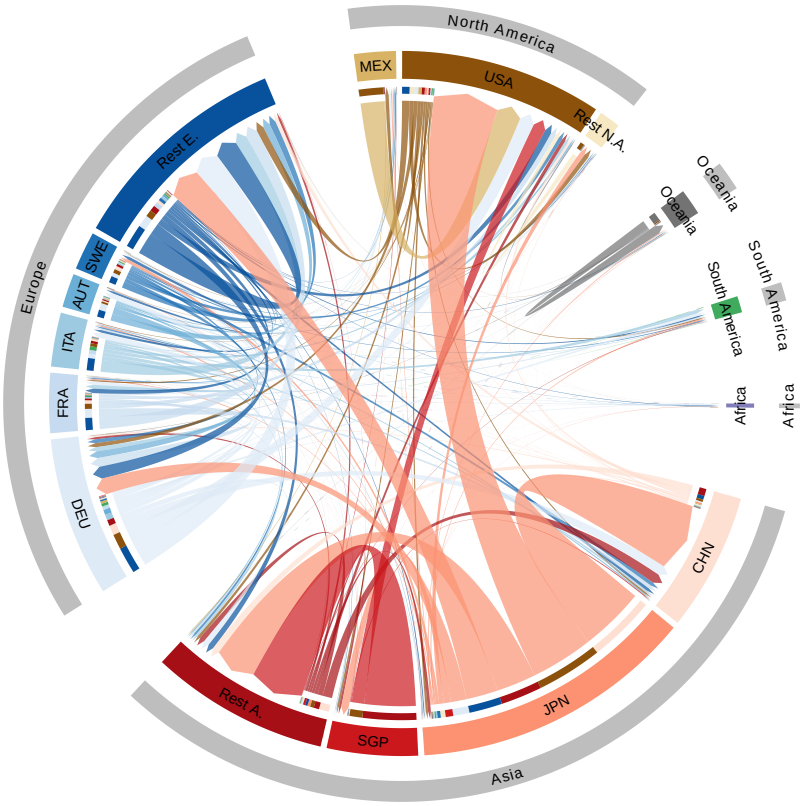


Figure 4.A3: Robot exports in billion USD, 1996-2018. Source: Comtrade and own imputation.

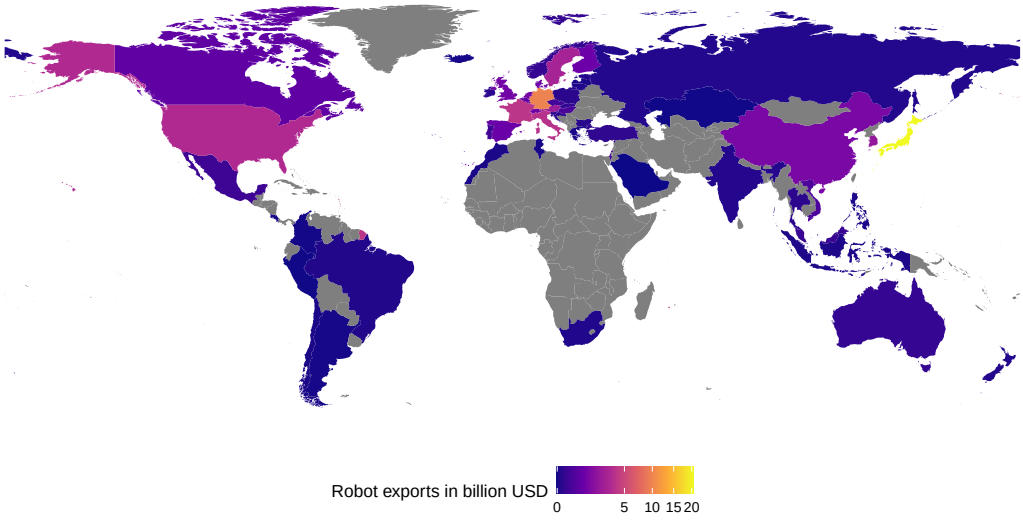


Figure 4.A4: Robot imports in billion USD, 1996-2018. Source: Comtrade and own imputation.

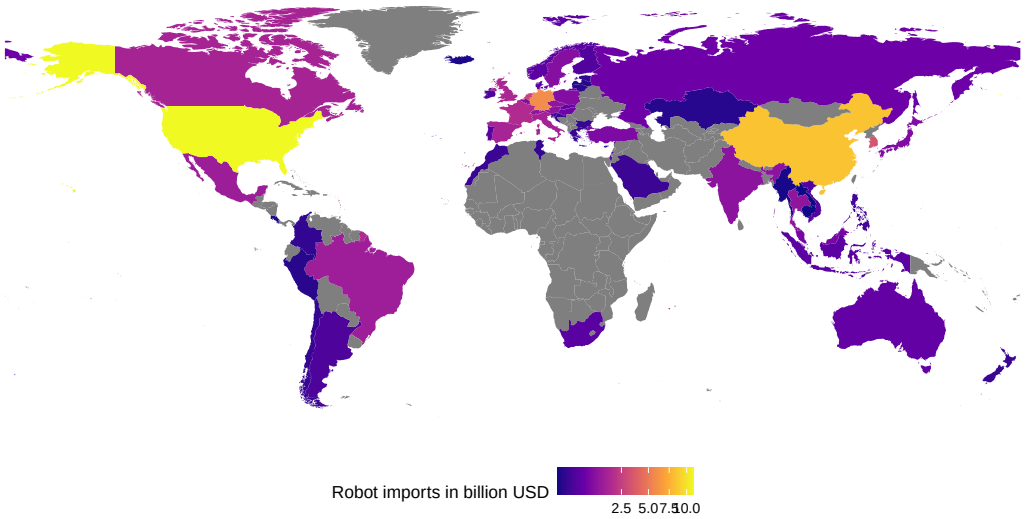


Figure 4.A5: Market shares of selected twelve countries with highest total exported units over 1996-2018 time span. Source: IFR and own imputation.

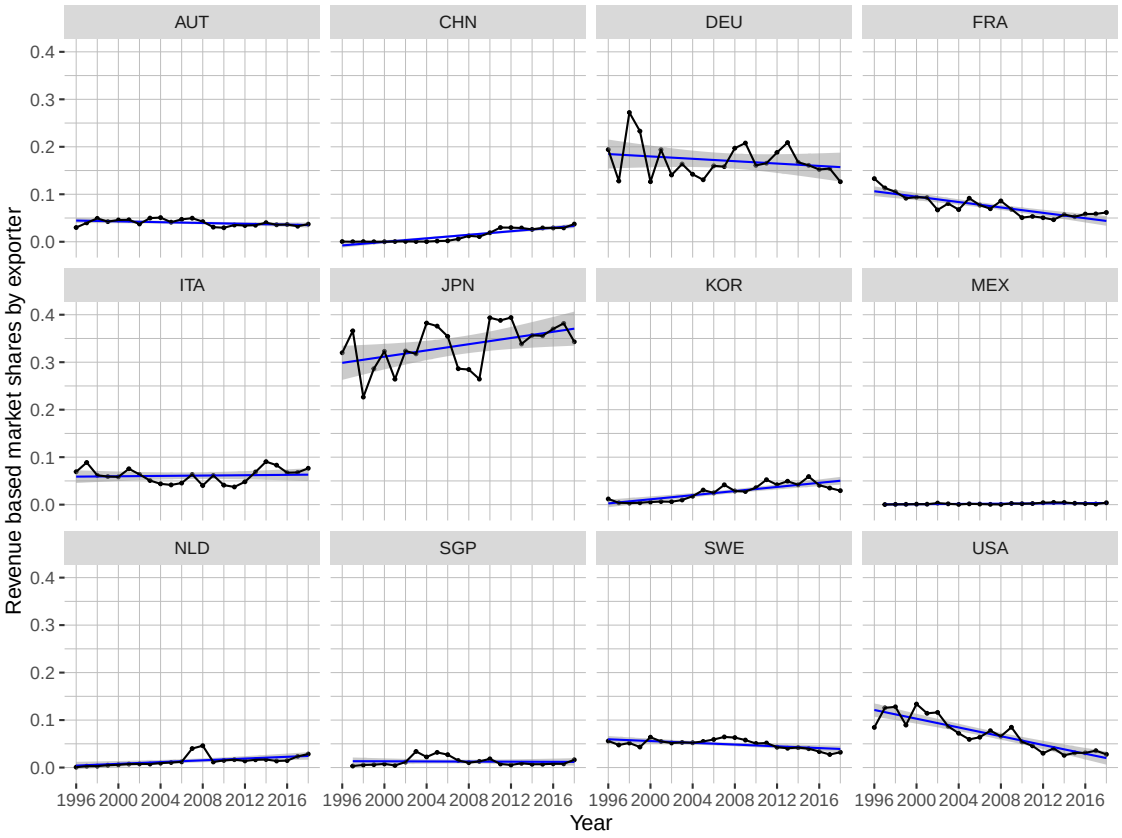


Figure 4.A6: Average exported robot unit weight in kg, 1996-2018. Source: Comtrade and own imputations.

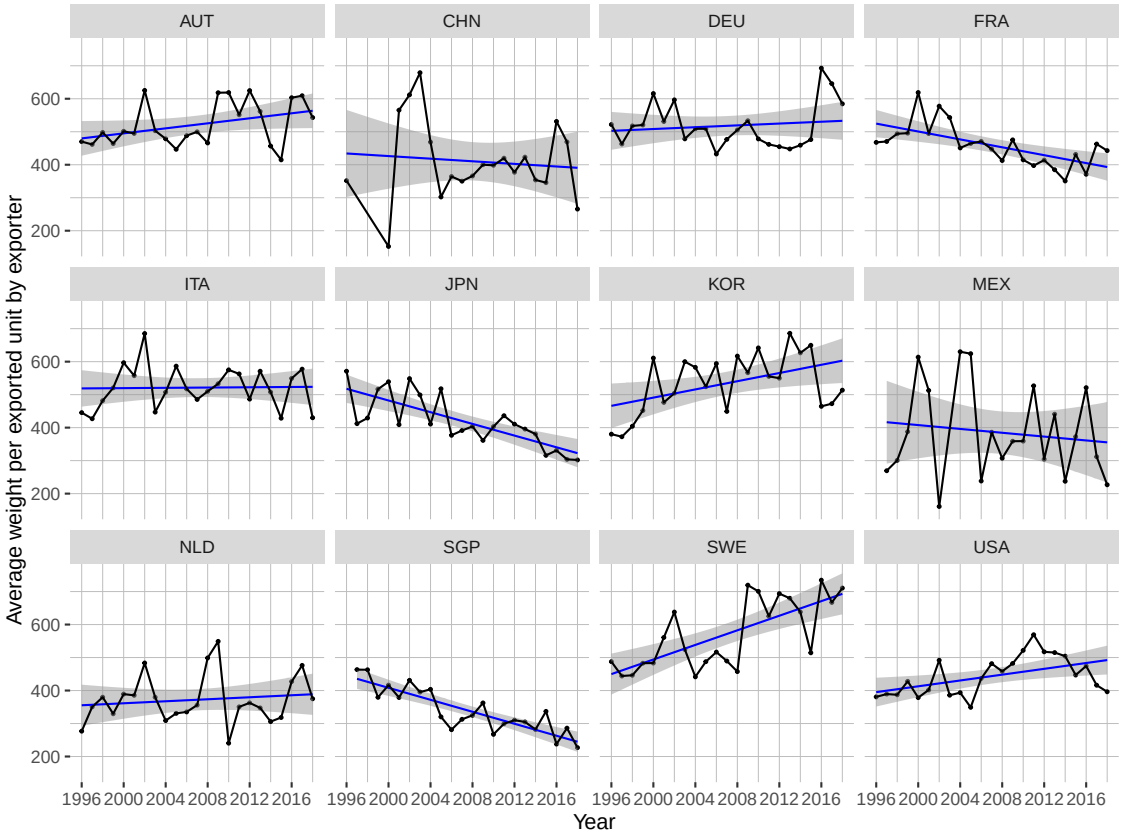


Figure 4.A7: Average real (PPI adj.) robot units price in 1000 USD per exported unit, 1996-2018. Source: Comtrade and own imputations.

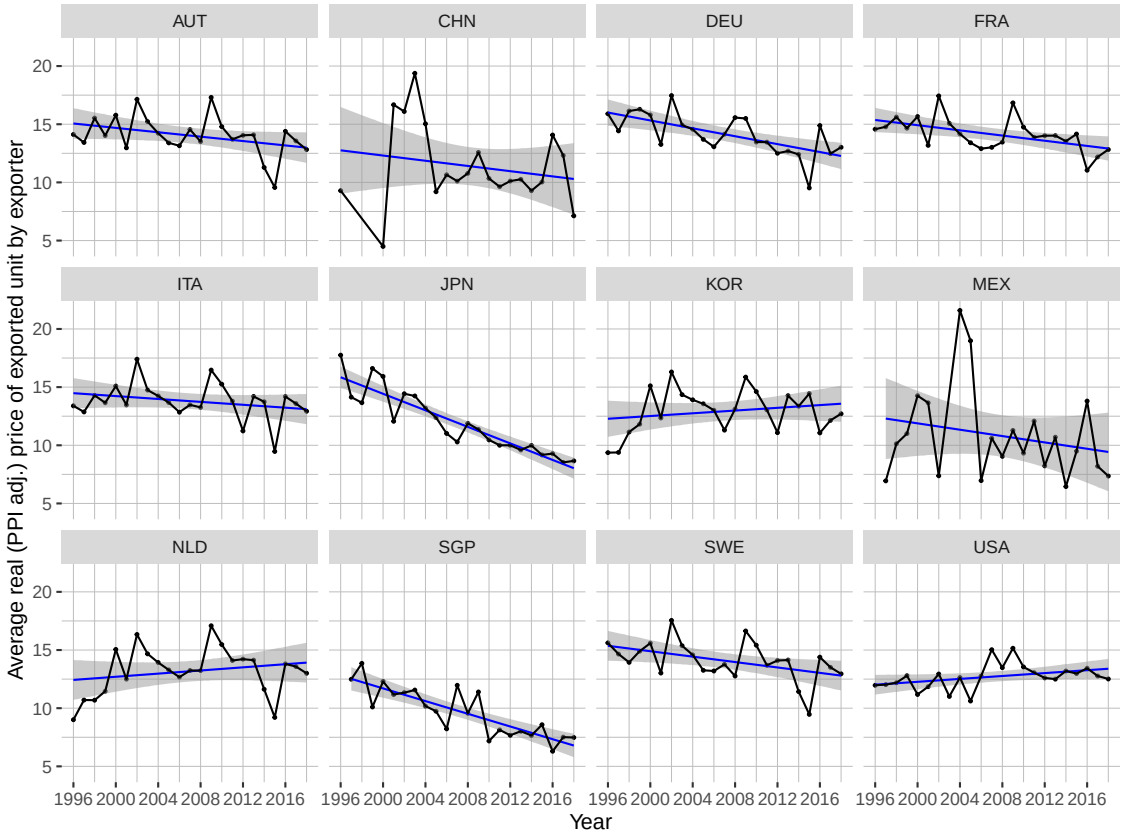


Figure 4.A8: Median exported robot unit weight in kg, 1996-2018. Source: Comtrade and own imputations.

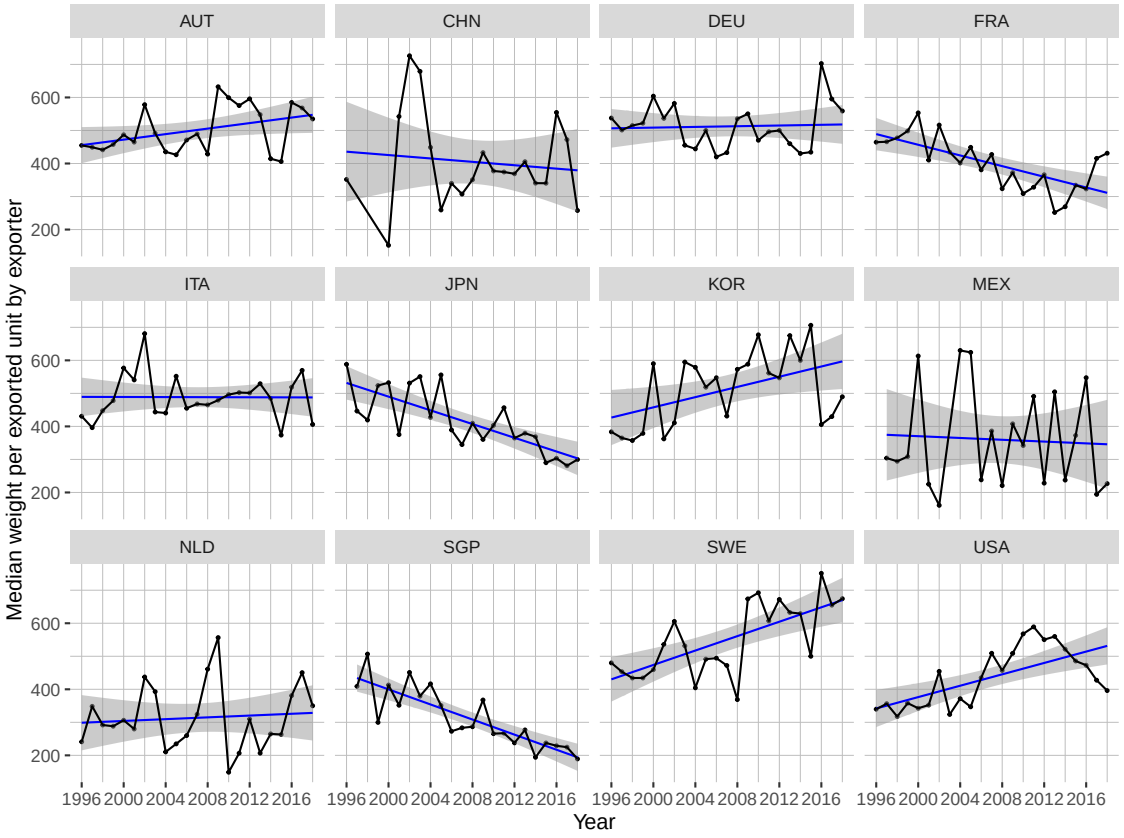
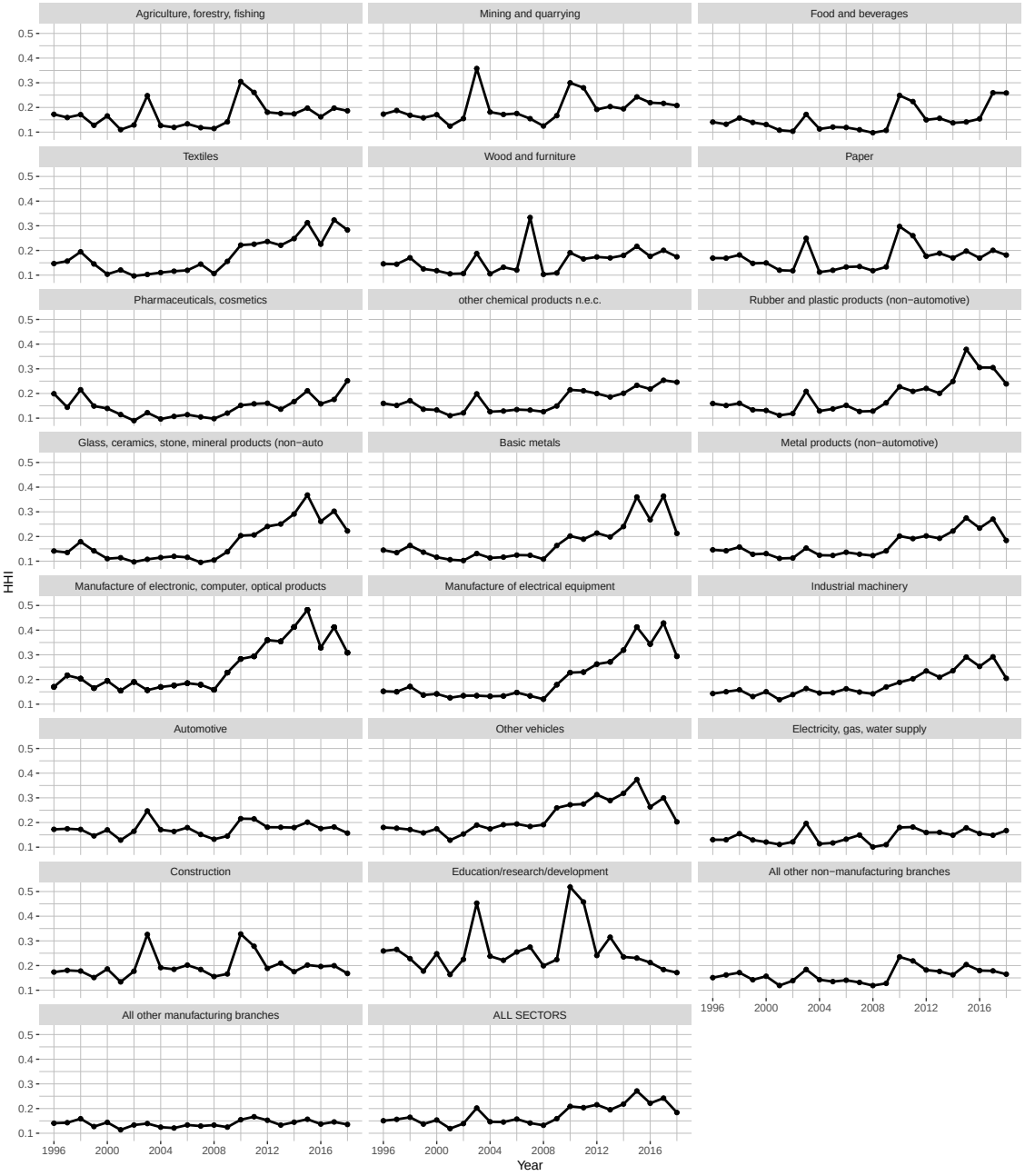


Figure 4.A9: Herfindahl-Hirsch-Index, IFR sector classification, weights derived from OECD input-output tables, 1996-2018



Chapter 5

Interviewer Biases in Medical Survey Data: The Example of Blood Pressure Measurements

Abstract

Health agencies rely upon survey-based physical measures to estimate the prevalence of key global health indicators such as hypertension. Such measures are usually collected by non-healthcare worker personnel and are potentially subject to measurement error due to variations in interviewer technique and setting, termed "interviewer effects". Using blood pressure as a case study, we aimed to determine the relative contribution of interviewer effects on the total variance of blood pressure measurements in three large nationally-representative health surveys from the Global South. In a linear mixed model, we modeled systolic blood pressure as a continuous dependent variable and interviewer effects as random effects alongside individual factors as covariates. To quantify the interviewer effect-induced uncertainty in hypertension prevalence, we utilized a bootstrap approach comparing sub-samples of observed blood pressure measurements to their adjusted counterparts. Our analysis revealed that the proportion of variation contributed by interviewers to blood pressure measurements was statistically significant but small: approximately 0.24-2.2% depending on the cohort. Thus, hypertension prevalence estimates were not substantially impacted at national scales. However, individual extreme interviewers could account for measurement divergences as high as 12%. Thus, highly biased interviewers could have important impacts on hypertension estimates at the sub-district level.

This chapter is joint work with Pascal Geldsetzer (Stanford University, Stanford, CA, USA), Andrew Young Chang (Stanford University, Stanford, CA, USA), Erik Meijer (Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA), Nikkil Sudharsanan (Technical University of Munich, Munich, Germany), Vivek Charu (Stanford University, Stanford, CA, USA), and Peter Kramlinger (University of California Davis, Davis, CA, USA).

5.1. Introduction

Global health indicators such as blood pressure, weight, and height are critical for monitoring both national and international health system performance. Such markers are largely collected through household surveys, which are often seen as the gold standard methodology due to their population-representative nature (Ties Boerma & Sommerfelt, 1993; Corsi et al., 2012; Boerma et al., 2003; Clark & Sanderson, 2009; Mbondji et al., 2014).

Interviewer-collected physical measures such as heart rate or body mass index (BMI) may appear to hold greater "objectivity" than self-reported indicators or subjective social indicators. Self-reported data is frequently prone to not only random measurement error, but also systematic measurement error due to interviewee attitudes such as recall bias and social desirability bias (Althubaiti, 2016). Nevertheless, physical measures are still subject to a substantial degree of random measurement error due to administrator technique and environmental context during acquisition (Ulijaszek & Kerr, 1999; Cernat & Sakshaug, 2020; Ali & Rouse, 2002; Svensson & Theorell, 1982). This phenomenon may possibly be magnified in the case where medical measurements are taken by non-clinician interviewers who may not routinely perform such measures outside of the research setting.

Nevertheless, many household surveys make the implicit assumption that, after their training, interviewers all perform to the same standard as one another (Jaszczak et al., 2009). Subsequent analyses therefore assume that the interviewers are not a source of measurement error and that uncertainty estimates are purely based on the sampling strategy. At the national level, these "interviewer effects" may average out from the large number of interviewers contributing both positive and negative measurement error. At finer geographic divisions, however, the relatively smaller number of interviewers may lead to greater variation or even potential bias in the measurement of a target indicator. This is particularly important because estimates from small areas are increasingly being used in public health decision making and for mapping disease prevalence at subnational levels, sometimes in resolutions as fine as 5 x 5 km (Dwyer-Lindgren et al., 2019; Reiner Jr et al., 2018; Osgood-Zimmerman et al., 2018; N. Graetz et al., 2018).

Prior analyses have queried the intra- and inter-observer reproducibility of specific physical measures, but such investigations have tended to focus on the reliability of these markers for clinical situations (Ali & Rouse, 2002; Svensson & Theorell, 1982; Schulze et al., 2000). Furthermore, most such studies have utilized healthcare workers like nurses and medical trainees as the measurement-takers given their applicability to the medical setting, and have examined high-income country populations (Bogan et al., 1993; Dickson & Hajjar, 2007). Large-scale empirical analyses of non-clinician interviewers' reliability for physical measures for public health purposes, especially in low- and middle-income countries (LMICs), remain sparse. The amount of random measurement error found in such global health indicators varies, with some exhibiting relatively low degrees (e.g., controlled laboratory-based tests) while others with increased operator inputs suffer from potentially greater degrees of interviewer-introduced measurement error. For example, anthropometry for newborns,

adult waist circumference, and blood pressure measurements require interviewers to make subjective decisions about how and where to place the instruments and in what settings to do so (Cernat & Sakshaug, 2020, 2021).

Here we assess the magnitude of interviewer-induced measurement error in large-scale global health surveys using the case study of high blood pressure. High blood pressure is an ideal case study because it is already a disease of considerable importance in low- and middle-income countries (LMICs) (Zhou et al., 2017; Yusuf et al., 2020). Blood pressure is readily and frequently measured noninvasively, and non-clinician study personnel can be taught how to collect blood pressure assessments (Jaszczak et al., 2009). This is particularly important as community health workers and other non-nurse/non-physician healthcare workers are increasingly being called upon to care for noncommunicable diseases in primary care in poor countries, and they are also frequently called upon for survey data collection as well (Jeet et al., 2017; Singh & Sachs, 2013; Otieno et al., 2012).

As such, in the present analysis (assuming that interviewers are randomly allocated to households within primary sampling units) we examine the magnitude of uncertainty attributable to interviewer effects on blood pressure measurements and hypertension (systolic blood pressure $\geq 140 mmHg$) in three large longitudinal health surveys from the Global South.

5.2. Materials and Methods

5.2.1. Data Sources

We demonstrate the implications of interviewer measurement biases using three common longitudinal health surveys. Besides waves 4 and 5, as well the east extension of the Indonesia Family Life Survey (IFLS), we use all five waves of the National Income Dynamics Study (NIDS), and the first wave of the Longitudinal Aging Study in India (LASI) in our analysis. All three data sets were collected with the purpose to document socioeconomic and health outcomes over time. Moreover, they were designed to provide sufficient sample size and adequate sampling schemes to be nationally representative. Thus, they are generally considered suitable to estimate prevalences of diseases for whose documentation adequate examinations were conducted as part of the survey, such as hypertension.

5.2.2. Sampling strategy

NIDS: The NIDS data were collected in five waves between February 2008 and December 2017 (Southern Africa Labour and Development Research Unit, 2018a,b,c,d,e). Since the NIDS data are of longitudinal nature, the households interviewed in the first wave were re-contacted for the following waves. However, the sample was topped up throughout the following waves to account for under-sampled socioeconomic groups and attrition. A two-stage stratified cluster sample design was applied in the data generation process of the

first wave.

The underlying 2003 master data used to generate NIDS were provided by Statistics South Africa, comprised 3000 primary sampling units (PSUs), and were stratified with respect to 53 district councils. The NIDS data depict a subset of 400 PSUs which were randomly drawn within the strata, whilst conserving proportionality. Within each PSU, 8 non-overlapping samples of dwelling units had been drawn for the creation of the master data, which are referred to as clusters in the NIDS documentation. The majority of clusters were assigned various household surveys before the creation of NIDS. Two clusters in each PSU however had never been involved in surveys, and became the base for NIDS. For further details see (Leibbrandt et al., 2009). NIDS wave 1 comprises completed surveys of 7,296 households from the aforementioned sub-sampled 400 PSUs. In order to establish national representativeness, different sets of weights were constructed as described in (Wittenberg, 2009). Since our analysis does not aim for national representativeness, but focuses on interviewer effects only, we do not apply the weights provided within the NIDS data and thus do not further discuss the computation of the weights here. Thus, they do not account for the interviewer effects, but merely the sampling weights.

After cleaning and preprocessing the NIDS data as outlined above, 87,658 observations remain, which we use throughout our analysis.

IFLS: The IFLS data used in the scope of this analysis comprise waves four, five and the east extension (Strauss et al., 2009; Sikoki et al., 2013; Strauss et al., 2016). As is the case with NIDS, due to the IFLS data being a longitudinal survey, the households interviewed during the first wave were re-contacted for all following waves. Thus, the sampling scheme of the first wave determined the sample composition of all following waves. IFLS1 stratified on provinces and urban versus rural locations within which simple random sampling was applied. Out of a total of 27 Indonesian provinces, only 13 are included in the sample, which however represented 83% of the population in 1993 (Strauss et al., 2016). Within the selected 13 provinces, 321 enumeration areas (EAs) were randomly chosen, with proportions being selected to cause oversampling of urban EAs and smaller provinces to ensure the comparability of rural and urban EAs. While within each urban EA 20 households were selected, 30 were selected within each rural EA, resulting in a total of 7,224 completed household interviews in IFLS1. For a more detailed description of the sampling scheme please refer to (Strauss et al., 2016). IFLS East includes most of the provinces not covered by the main IFLS. Within each selected province, 14 villages or urban villages were randomly drawn. These were then subdivided into units/areas with about 100-150 households, from which one was drawn at random. Within each of these, again 20 households were drawn if urban and 30 if rural. See (Sikoki et al., 2013) for more details. After initial data cleaning and processing 26,554 individual level observations from IFLS 4, 5, and East remain, which we use in the scope of this analysis.

LASI: We use the first wave of LASI data which was collected between 2017 and 2019 (International Institute for Population Sciences (IIPS), MoHFW, Harvard T. H. Chan School of Public Health (HSPH) and the University of Southern California (USC), 2020). The sampling scheme applied throughout the LASI data collection followed the 2011 census, and implemented a multistage, stratified cluster sample design. While in the case of urban areas three sampling stages were conducted, four stages were conducted in the case of rural areas. The first stage consisted in the selection of PSUs within states. In the second stage villages were selected in the rural PSUs and wards within the urban PSUs. Stage three included the selection of households in rural areas and the selection of Census Enumeration Blocks (CEBs) in wards. The final and fourth stage applied in urban areas comprised the selection of households. The LASI data used in the scope of this analysis comprise 55,469 observations post pre-processing and cleaning.

5.2.3. Interviewer Training, Characteristics, and Monitoring

NIDS: Interviewer training was held at the same time as the pre-test was conducted, specifics on the training of blood pressure measurements are not documented. The NIDS documentation does not mention specially trained health professionals taking the health measurements as is common in similar surveys. Thus, health measurements have been taken by the interviewer conducting the rest of the household surveys.

With wave five a set of interviewer demographics and experience variables were added to the available data.

The use of paradata was implemented to oversee interviewers and thereby reduce interviewer effects. Precisely, paradata are used to monitor questionnaire duration, refusal rates, magnitude of anthropometric measurement differences between current waves and previous waves, flag extreme BMI measures, and run other similar checks. The checks were taken periodically from about 6 weeks into fieldwork or when there were enough data to estimate meaningful averages. When interviewers' performance measures were conspicuous they were investigated, retrained, moved to different teams for closer supervision or removed. In some cases the respective households were re-interviewed. The Southern Africa Labour and Development Research Unit (SALDRU) carried out a range of pattern searches and consistency checks on the data during fieldwork to identify interviewer effects and potential general cases of mis-capture.

The NIDS sample used in our analysis comprises a total of 513 distinct interviewers taking blood pressure measurements.

IFLS: Supervisory training was held for all senior personnel. In the case of IFLS5 this training of trainers included reviewing all parts of the survey: household, community-facility, health, computer-assisted personal interview system (CAPI) tracking, and the management information systems used in the scope of the data collection. Household interviewer training was conducted in two phases. Training sessions were divided into two parts, classroom training and field practice. Household interviewers received 19 days of classroom training and 4 days of field practice. The collection of health data was conducted by regular interviewers, i.e., no health professionals were involved in the data collection on site during the interviews. Training for health-related measurements was part of the regular interviewer training. In the case of IFLS4 and IFLS East, the CAPI system had not been implemented yet and blood pressure measurements were conducted by nurses, i.e. professional health workers, and non-professional interviewers, respectively.

The combined IFLS data contains a total of 409 distinct interviewers taking blood pressure measurements.

LASI: A series of manuals were designed to standardize different aspects of surveys conducted in the scope of the LASI data collection. These manuals were instrumental in the training of interviewers. One of the manuals specifically focuses on the physical measures section of LASI and thus includes instructions for the measurement of blood pressure. The training duration of interviewers and health investigators was 35 days, of which five took place in the field. Even though the interviewers were employed via sub-contractors, they were trained by trainers, who themselves were trained by the International Institute for Population Sciences (IIPS). After training was completed, investigators were individually assessed to assure that their work met the requirements previously defined by the manuals.

The LASI sample used in our analysis comprises a total of 504 distinct interviewers taking blood pressure measurements.

5.2.4. Definition of Hypertension, Blood Pressure Measurement

Multiple systolic blood pressure measurements were taken in the scope of all surveys included in this study. In the case of the IFLS and LASI data, three measurements were taken per individual, in the case of the NIDS data only two. In order to mitigate the white coat effect and to average out idiosyncratic fluctuations in measurements, we average the second and third measurement, while disregarding the first in the case of IFLS and LASI. In the case of NIDS, we only consider the second measurement, disregarding the first. Following this procedure, we obtain a single systolic blood pressure value for each

interviewee. We consider interviewees to be hypertensive if their resulting single systolic blood pressure measurement is equal to or greater than 140 mm Hg.

Measurements were conducted using an Omron HEM 7121 BP monitor in the case of LASI and an Omron HEM 7203 in the case of IFLS. Information on the exact device used for blood pressure measurement throughout NIDS data collection is not part of the publicly available documentation.

5.2.5. Definition of Covariates

We add covariates to the model, which we consider potential determinants of blood pressure. To keep the results comparable, we use mostly the same set of covariates across all data sets. Besides using interviewees' sex, age, BMI, and smoking status, we proxy interviewees' socioeconomic background with income and education. The variables we choose in the respective data sets to compose our income proxy refer to monthly salaries and wages or monthly profits from entrepreneurship for NIDS and IFLS, and the logarithm of total household income for LASI. While the resulting income variables are hardly comparable across data sets, we assume comparability within data sets. To align the information on the education of individuals, we re-code education into three categories, namely less than primary schooling, primary and/or secondary schooling and tertiary education, except in LASI, where we added a fourth category for no schooling. In order to proxy for the possible use of blood pressure lowering medication we include a variable which depicts whether an interviewee has ever been diagnosed with hypertension before.

5.2.6. Statistical Analysis

We model a linear relationship of systolic blood pressure and available covariates. Formally, systolic blood pressure for individual i in household j at location k measured by interviewer l is denoted as Y_{ijkl}, so that

$$Y_{ijkl} = \beta_0 + \sum_{d=1}^{p} x_{ijkld}\beta_d + u_j + v_k + w_l + \varepsilon_{ijkl}, \tag{5.1}$$

where $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$ is a Gaussian error term. Furthermore, $x_{ijkl1}, \ldots, x_{ijklp}$ are the available co-variates, $\beta_0 \in \mathbb{R}$ a common intercept and u_j, v_k, w_l the respective level-effects of household, location and interviewer, for $j = 1, \ldots, J$, $k = 1, \ldots, K$, $l = 1, \ldots, L$. These level-effects, as well as the parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$ are unknown and have to be estimated given a sample of independent measurements.

Since the main objective lies in investigating systolic blood pressure, this model includes a selection of socio-economic control covariates. The separately modeled level-effects include a household effect, the interviewer effect, and the maximum number of geographical level-

effects supported by the respective data set. In the following we will motivate the use of the individual level-effects. We suspect that the interviewer effect significantly influences systolic blood pressure measurements, and is at the core of our analysis, as described above. Of note, due to the inability to trace interviewers across waves of the datasets, we treat all observations individually and ignore the time dimension.

We motivate the use of geographical level-effects based on the assumption that geographical cultural clusters, geographical differences in the availability of food, geographical differences in health care access, and similar factors might affect systolic blood pressure spatially.

It is common practice to assign interviewers to households and not to interviewees directly. An interviewer then interviews all eligible individuals belonging to an assigned household. Variation in systolic blood pressure on the household level therefore potentially confounds the estimation of the interviewer effect. Thus, we include household effects to absorb household level variation.

We are interested in investigating $Y_{ijkl} - w_l$, which is the systolic blood pressure adjusted for the true measurement error induced by interviewers, which we are estimating with our approach. Accordingly, we consider $Y_{ijkl} - \widehat{w}_l$, where $\widehat{w}_l$ is a suitable estimator for w_l. In regression problems with multiple dimensions such as the present case outlined in 5.1, the question arises as to which effects are best modelled as random versus modelled as fixed. In general, with a large number of coefficients to be estimated, the potential loss in degrees of freedom associated with modelling fixed effects is considered an argument in favor of random effects. In the large surveys considered in this article several hundred interviewers were involved in taking measurements. Estimating a fixed effect for each interviewer is thus prohibitively expensive in terms of degrees of freedom. We therefore proceed in line with common practice and assume that the household effect u_j, and interviewer effect w_l are stochastic (Hodges, 2013; Hsiao, 2014; Fielding, 2004). In case of the location effect v_k the optimal choice is less clear. The potential loss of degrees of freedom is lower due to the lower number of coefficients to be estimated, especially at the highest level of geography. However, in order to maintain maximum comparability of the level-effects, we consider it sensible to model all as random.

These random level-effects are assumed to be independently drawn from underlying normal distributions (Hodges, 2013). As part of the estimation procedure we obtain estimates for the respective second moments of these distributions, which then can be used for simulation exercises or the calculation of reliability ratios. With the assumption of random effects, equation 5.1 constitutes a linear mixed model (LMM), that is:

$$u_j \sim N(0, \sigma_u^2), j = 1, \ldots, J; \qquad v_k \sim N(0, \sigma_v^2), k = 1, \ldots, K; \qquad w_l \sim N(0, \sigma_w^2), l = 1, \ldots, L.$$

5.2.7. Omitted variable bias

An individual's blood pressure depends on various factors, only some of which can be fully captured in large-scale surveys. Genetic preconditions for example are practically impossible to capture sufficiently in survey settings. Thus, we are agnostic about facing omitted variable bias in explaining systolic blood pressure independent of the particular survey data set considered. However, depending on the survey, some essential predictors of blood pressure are missing, which in principle could be recorded in a survey setting.

Recalling that our main interest lies in investigating interviewer effects, we are mostly concerned about falsely attributing variation in systolic blood pressure measurements to interviewers. Confounding is most likely to occur if an interviewer's specific subset of individuals substantially differs from the overall population, along a dimension relevant for variation in systolic blood pressure.

The risk of confounded interviewer intercept estimates caused by small samples is mitigated by using the best linear unbiased predictor (BLUP) for random effects (Henderson, 1975; Rao & Molina, 2015). This estimator is a weighted average of the pooled sample and the sample from the level-specific subgroup, i.e., all measurements taken by one specific interviewer. The former exhibits a bias and small variance, whereas the latter is unbiased but has a large variance. It is constructed so that the more observations there are in the level-specific subgroup, the more weight is attributed to it. Conversely, if the level-specific subgroup sample is very small, the BLUP relies more heavily on the pooled sample. The estimation procedure therefore amounts to a variance-bias trade-off in which the BLUP is optimal in terms of the mean squared error (MSE). Consequently, the potential small sample bias that leads to confounded interviewer intercept estimates is small, and its impact negligible.

5.2.8. Testing for the Presence of Interviewer Effects

We are interested in investigating the presence and significance of interviewer effects. This relates to the formal test of the hypothesis H_0: $\sigma_w^2 = 0$ vs. H_1: $\sigma_w^2 > 0$. This test is performed by evaluating the likelihood-ratio statistic

$$LRT = 2\left(\ell_{H_1} - \ell_{H_0}\right),$$

where ℓ_{H_0} is the log-likelihood of the model under the null and ℓ_{H_1} for the alternative. In our concrete case, ℓ_{H_1} nests ℓ_{H_0} and additionally includes interviewer random effects. As fundamental problem, the null lies at the boundary of the parameter space. The asymptotic distribution of the Likelihood-Ratio-Test (LRT) has the inconvenient distribution of a point-mass on zero with weight 0.5 and χ_1^2-distribution elsewhere. The finite sample distributions

however may severely differ from the asymptotic distribution (Crainiceanu & Ruppert, 2004, 2005). For multiple random effects as in the present model, a parametric bootstrap can approximate the finite sample distribution well enough (Crainiceanu, 2008; Greven et al., 2008). In particular,

$$LRT \overset{d}{\approx} aU\chi^2_1,$$

where $\overset{d}{\approx}$ denotes approximate equality in distribution, $U \sim Bern(1-p)$. Both a and p are unknown and have to be estimated by bootstrap replications. Eventually, p-values for the LRT under the null can be provided.

5.2.9. Adjusting for Interviewer Effects

Once we have established the presence and significance of interviewer effects, we adjust blood pressure measurements for these interviewer effects. Since we obtain not only an estimate of the second moment of the interviewer effect distribution, but also intercepts for all individual interviewers, we can individually adjust systolic blood pressure measurements. A simple adjustment then takes the form

$$\widehat{Y}^{\text{adj}}_{ijkl} = Y_{ijkl} - \widehat{w}_l, \tag{5.2}$$

where $\widehat{w}_l$ are the interviewer intercept effects (the BLUPs).

5.2.10. Assessing Uncertainty in Sample Hypertension Prevalence

In order to quantify the uncertainty in hypertension prevalence induced by interviewer measurement error we use a non-parametric bootstrap approach. Precisely, for this approach we repeatedly take sub-samples of observed systolic blood pressure measurements and their corrected counterparts and compare resulting prevalences of hypertension. We depict the two generated sets of prevalences as densities, which allows for a straightforward comparison.

Bootstrap

We employ a non-parametric cluster bootstrap approach to infer about the uncertainty of hypertension prevalence given the corrected observations. We refer to this approach as non-parametric, since we do not use estimated parameters from the estimated model to generate new data, but only use the predicted interviewer effects to create adjusted measurements post estimation. Thus, we compare the density of hypertension prevalences based on corrected observations to the density of prevalences based on uncorrected observations. In order to account for the clustered structure of our data, we fix the coarsest geographic level (e.g. provinces) in the data and within these levels we draw from the second coarsest geographical level (e.g. municipalities).

The location level effects depict multiple levels of granularity and thus can also be represented as distinct effects. Let $p = 1, ..., P$ indicate the coarsest geographical level effect (e.g. province), and $m = 1, ..., M(p)$ represent the second coarsest geographical level effect (e.g. municipality).

Formally, let $y_{ipm}, m = 1, ..., M(p)$ be the ith individual measurements in province p and y_{ipm}^{adj} the adjusted measurements respectively. Then, R bootstrap replications are generated via:

1. for $r = 1, \ldots, R$:
2. for $p = 1, \ldots, P$:
3. Draw $M(p)$ municipalities with replacement
4. Obtain composite sample $B(p) \subset \{y_{ipm} | \text{for individual } i \text{ in muncipality } m\}^{M(p)}$
5. Pool random samples to obtain $B = \underset{p}{\cup} B(p)$
6. Calculate $p_r(B) = |B|^{-1} \sum_{y \in B} \mathbb{I}(y > 140)$, and p_r^{adj} analogously

The bootstrap prevalences $(p_r)_{r=1,...,R}$ and $(p_r^{adj})_{r=1,...,R}$ allow for inferring about the effect of adjustment.

5.3. Results

5.3.1. Sample Characteristics

Table 5.1 shows descriptive statistics for the data sets used in this study after pre-processing. Data from 169,681 total encounters were utilized, with 26,554 from the Indonesia Family Life Survey (IFLS), 55,469 from the Longitudinal Aging Study in India (LASI), and 87,658 from the National Income Dynamics Study (NIDS) of South Africa, respectively.

5.3.2. Variation shares in hypertension prevalence

To interpret the effect sizes of the interviewer level-effects, we compare their shares in total variation to the shares of other level-effects and the residual from the same estimations. Table 5.2 presents the variance components of the fitted linear mixed models (LMM) for the IFLS, NIDS, and LASI datasets.
The bootstrap likelihood ratio test (LRT) tests give p-values of $p < 0.0001$ for all three datasets. This strongly suggests the presence of interviewer effects in all three datasets, although they are numerically small.

5.3.3. Uncertainty in Sample Hypertension Prevalence

Figure 5.1 displays the non-parametric bootstrap densities for hypertension prevalence, based on the original data (blue, dashed), and the corrected measurements (red, dotted). The vertical line represents the observed prevalence by data source.

5.3.4. Effect Study

In order to illustrate the interviewer-introduced uncertainty in hypertension prevalences, we perform an effect study. Using the set of observed systolic blood pressure measurements and the measurements corrected for the estimated interviewer effects, we can compare observed interviewer-specific prevalences of hypertension to the respective corrected interviewer-specific prevalences. Alternatively, we can also illustrate differences in prevalences for geographic areas, such as sub-districts.

Interviewer-Specific Prevalences: Observed and Corrected

Figure 5.2 illustrates a sub-sample of the interviewer-specific observed and adjusted prevalences of hypertension for the IFLS dataset. The sub-sample is created based on the distribution of differences in observed and adjusted prevalences. For example, to focus on the most extreme cases, we depict the prevalences for all interviewers for whom the difference between observed and adjusted prevalence lies above the 70th-percentile of these differences. In other words, we show the 30 percent of cases subject to the most drastic adjustment effects. The top 50%, 30%, 10%, and 1% cases are presented.

The analogous findings for NIDS and LASI are presented by Figures 5.3 and 5.4

Sub-district specific prevalances: observed and corrected

Analogously to the interviewer specific prevalances, we can also depict changes in prevalances for geographical units, as illustrated in Figure 5.5. The higher the granularity in geographical division, the larger the influence of single interviewers. We thus depict adjustment induced changes in prevalences on the most granular level available for each respective data set. In case of LASI and IFLS the most granular geographical level are sub-districts. In the case of NIDS less granular level data is available, so that we are limited to the cluster level.

5.4. Discussion

In the present analysis, we found that interviewer effects in blood pressure measurements were statistically significant, although numerically trivial, in three large longitudinal health surveys from Indonesia, India, and South Africa. This was achieved by calculating the

proportion of total variance attributable to various sources, one of which was the interviewer. Nevertheless, both the absolute and relative contribution of the interviewer to blood pressure measurement variation was not particularly high, especially when compared to geographic/community-level effects. In the IFLS cohort, interviewer-level effects comprised 0.5% of the variance, while in NIDS, 2.2%, and in LASI, 0.2%. In fact, household effects (13.6%, 12.1%, 6.6%, respectively) dominated the variance of all three datasets, with residential effects (i.e., province, state, subdistrict, municipality) higher than interviewer effects except for in NIDS.

On the population level, however, the combined interviewer effect could potentially impact the uncertainty in hypertension prevalence. As such, we generated non-parametric bootstraps of prevalence estimates unadjusted and adjusted for the interviewer effect, which show very small but consistently lower point estimates of hypertension prevalence in all three datasets on the order of a fraction of a percent. This may have minor implications for public policy targeting hypertension and suggest slight present overestimation of true hypertension prevalence in these settings.

Nevertheless, the magnitude of the discrepancies is not exceedingly high at these larger scales — where we found the interviewer effect to carry the greatest possibility of influencing hypertension estimation was at smaller geographic divisions. Taking the most "extreme" individual interviewers responsible for the greatest adjustment effects in each dataset and comparing their observed and adjusted hypertension prevalences revealed divergences as high as 12% in NIDS. We therefore assessed their impacts by comparing the observed and interviewer-effect adjusted sub-district specific hypertension prevalences subject to the greatest adjustment effects. These revealed up to 5-7 percentage points (p.p.) prevalence differences between observed and corrected values at sub-district levels for the top 1% of cases subject to adjustment effects. The substantial degree of bias that these may introduce at the local level compared to the population (or whole sample) level are well-visualized in the resultant cluster-specific blood pressure density plots. For example, in LASI, the modal systolic blood pressure signed difference between subdistrict and total population was nearly $25 mmHg$.

Our study represents the largest empirical estimation of interviewer effects on blood pressure. We also believe it to be the first of its kind involving low- and middle-income country populations. Thus, it contributes to the growing body of work examining and quantifying interviewer-based sources of measurement error for survey-based global public health indicators. The results are reassuring that the present strategy of utilizing non-clinician study interviewers is likely not generating a critical degree of variation in blood pressure measurement for populations, and we propose one possible method by which analysts may adjust for these small interviewer effects.

Because our investigation is, to our knowledge, the first to assess interviewer-effects for blood pressure in household surveys from low- and middle-income countries, we are only able to compare our findings to those from much smaller samples in two surveys from the UK. Cernat and Sakshaug found that there are interviewer effects on measurement error from both nurses and trained non-clinician interviewers in these two UK-based surveys (Cernat & Sakshaug, 2020, 2021). For non-clinician interviewers, they noted that in measures such as height, weight, blood pressure, and pulse, interviewer effects similarly comprised only a small fraction of the variance—for blood pressure, less than 1%. Much like our findings, these studies also identified that area-level effects contributed a greater source of variation than the interviewer effect for many physical measures.

Nevertheless, our work further models the public health implications of the interviewer effect by estimating the impact of these forces on hypertension prevalence estimates at multiple geographic levels. In doing so, our analyses also identified that extremely biased interviewers could lead to markedly biased hypertension estimates, and that if there is disproportionate allocation of these "extreme" interviewers to a locale at the level of a sub-district or smaller, that there may be substantially biased hypertension prevalence estimates in these geographic units.

Strengths of our study include the size of the analytic cohort (total 169,681 observations), as well as the use of three different nationally-representative datasets from Africa, South Asia, and Southeast Asia. There is substantial heterogeneity in the resultant populations, not just by the distribution of gender, age, and urban/rural breakdown, but also the underlying true prevalence of hypertension. Blood pressure measurements from years 2008 through 2019 were included, further capturing time-related variation. The most important limitation of our analysis is that, analytically, our modeling strategy relies upon the assumption that all interviewers were quasi-randomly allocated to participants within the primary sampling units. Other limitations of the work include the inclusion of systolic blood pressure only, for reasons of statistical feasibility. Diastolic hypertension (both independently and in conjunction with systolic hypertension) may be a risk factor for adverse cardiovascular outcomes (Flint et al., 2019; Strandberg et al., 2002). In addition, the LASI cohort was substantially older than the IFLS and NIDS cohorts. Furthermore, the full dataset does not constitute a random sample of all household surveys in low- and middle-income countries. Lastly, all three survey cohorts involved interviewers who were highly trained using established, high-quality protocols and closely monitored by study administration. As biased interviewers have higher impact on measurement error in small geographic units, our results may underestimate the magnitude of interviewer effects for less-rigorously trained/observed interviewers in LMIC settings.

We conclude by noting that interviewer effects appear to be present, but small at best in household surveys of blood pressure in lower middle- and middle-income countries. Future

work could involve targeted empirical analyses of the influence of "extreme interviewers" on quantifying the local burden of disease, as well as replication of our methods in other cohorts from different continents and from low-income countries. Additionally, we recognize that blood pressure is but one physical measure from a large pool of monitored global health indicators. As prior research in other settings has suggested that interviewer effects vary with the type of measurement performed, independent analyses of these other markers such as weight and body mass index should be pursued to provide a more comprehensive understanding of the phenomenon.

5.5. Figures and Tables

Table 5.1: Descriptives of IFLS, NIDS and LASI data.

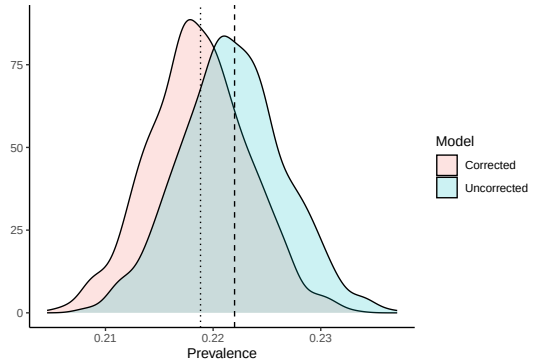
	IFLS (N=26554)	LASI (N=55469)	NIDS (N=87658)	Overall (N=169681)
Average SBP measurement				
Mean (SD)	129 (19.9)	129 (19.1)	121 (21.3)	125 (20.8)
Median [Min, Max]	126 [68.0, 241]	127 [60.0, 234]	117 [44.0, 240]	122 [44.0, 241]
Sex				
Male	16569 (62.4%)	25694 (46.3%)	36446 (41.6%)	78709 (46.4%)
Female	9985 (37.6%)	29775 (53.7%)	51212 (58.4%)	90972 (53.6%)
Age				
Mean (SD)	41.7 (13.1)	59.5 (10.4)	36.1 (17.1)	44.6 (18.0)
Median [Min, Max]	42.0 [15.0, 101]	58.0 [45.0, 108]	32.0 [14.0, 108]	46.0 [14.0, 108]
Education				
Less than primary	0 (0%)	6317 (11.4%)	8224 (9.4%)	14541 (8.6%)
Primary or secondary	22629 (85.2%)	42613 (76.8%)	67901 (77.5%)	133143 (78.5%)
Tertiary	3925 (14.8%)	6539 (11.8%)	11533 (13.2%)	21997 (13.0%)
Body-mass-index (BMI)				
Mean (SD)	23.3 (4.28)	22.8 (4.73)	26.1 (6.72)	24.6 (6.00)
Median [Min, Max]	22.7 [10.7, 57.1]	22.3 [10.5, 55.6]	24.6 [10.4, 60.0]	23.4 [10.4, 60.0]
Ever diagnosed with hypertension				
Not diagnosed	23406 (88.1%)	39600 (71.4%)	75814 (86.5%)	138820 (81.8%)
Diagnosed	3148 (11.9%)	15869 (28.6%)	11844 (13.5%)	30861 (18.2%)
Log income				
Mean (SD)	12.5 (4.04)	11.2 (1.66)	7.95 (0.942)	9.71 (2.72)
Median [Min, Max]	13.7 [0, 19.5]	11.4 [0, 18.8]	7.82 [4.25, 13.0]	9.13 [0, 19.5]
Smoking				
Non-smoker	14850 (55.9%)	47686 (86.0%)	75814 (86.5%)	138350 (81.5%)
Smoker	11704 (44.1%)	7783 (14.0%)	11844 (13.5%)	31331 (18.5%)
Urban/Rural				
Urban	15300 (57.6%)	19263 (34.7%)	43477 (49.6%)	78040 (46.0%)
Rural	11254 (42.4%)	36206 (65.3%)	44181 (50.4%)	91641 (54.0%)

Table 5.2: Variance components of the fitted LMMs by data set for IFLS, NIDS and LASI.

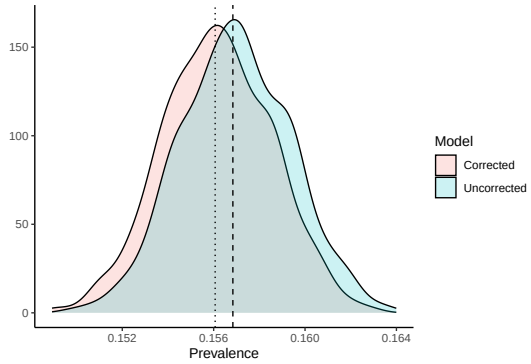
	Effect	Variance	Percentage
IFLS			
	Household	37.6	13.6%
	Interviewer	1.47	0.53%
	Province	2.96	1.07%
	Municipality	2.29	0.83%
	Subdistrict	2.78	1.01%
	Residuals	229	82.9%
	Total	276.1	$\approx$ 100%
NIDS			
	Household	39.6	12.1%
	Cluster	3.74	1.15%
	Interviewer	7.19	2.2%
	Province	3.06	0.94%
	Residuals	273	83.7%
	Total	328.17	$\approx$ 100%
LASI			
	Household	21.3	6.55%
	Interviewer	0.785	0.24%
	State	7.99	2.46%
	District	7.75	2.39%
	Village/ward	4.96	1.53%
	Residuals	282	86.8%
	Total	324.785	$\approx$ 100%

Figure 5.1: Bootstrap densities for hypertension prevalence, based on the original data (blue, dashed), and the corrected measurements (red, dotted). The vertical line represents the observed prevalence.

(a) IFLS



(b) NIDS



(c) LASI

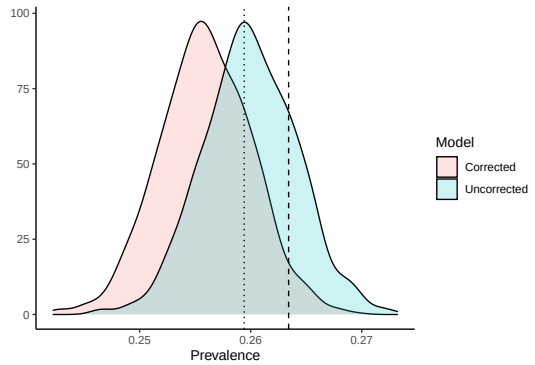
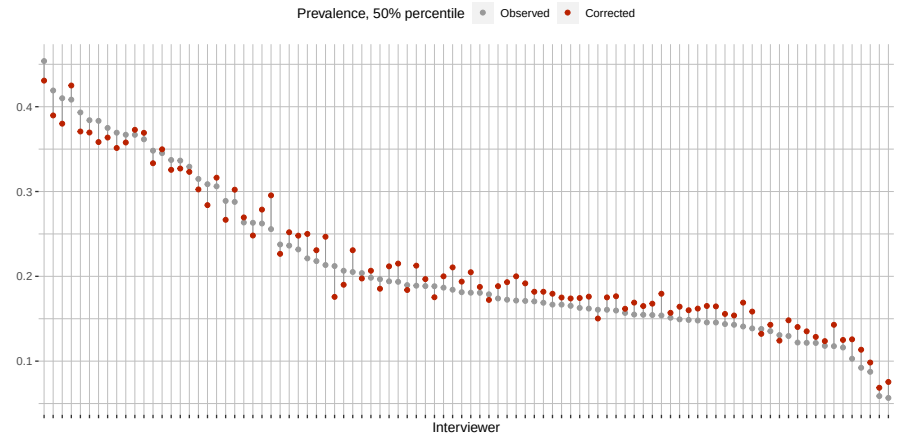
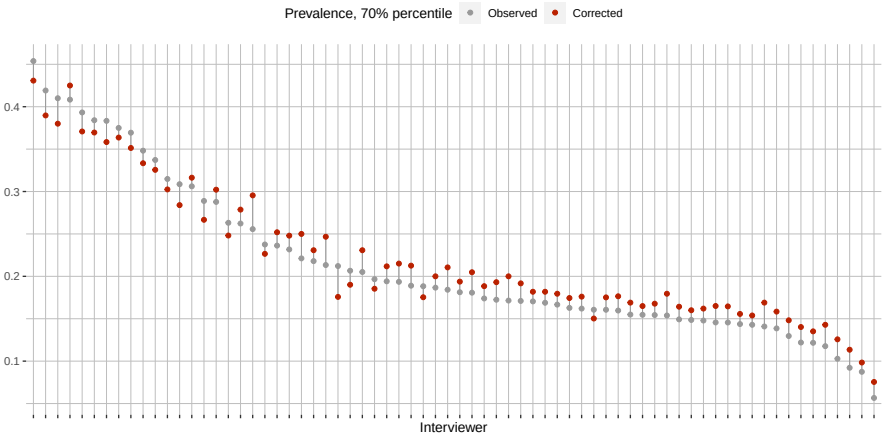


Figure 5.2: IFLS: Observed and adjusted interviewer specific prevalences of hypertension, 50%, 30%, 10%, 1% of cases subject to largest adjustment effects.

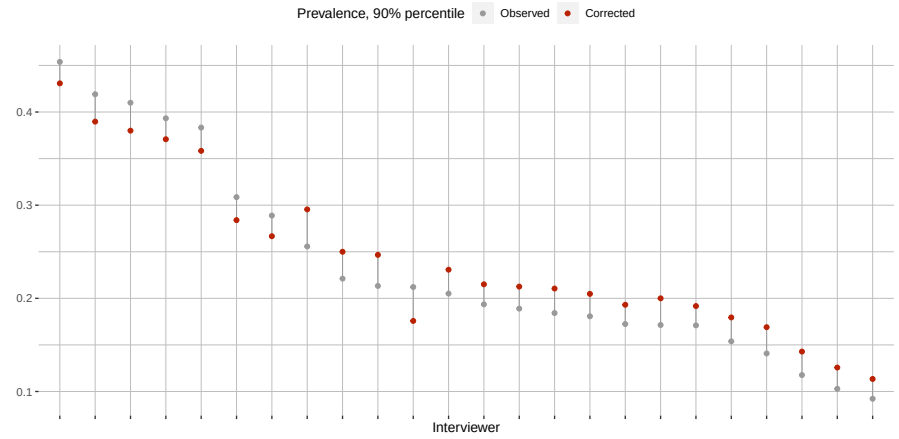
(a) 50th percentile



(b) 70th percentile



(c) 90th percentile



(d) 99th percentile

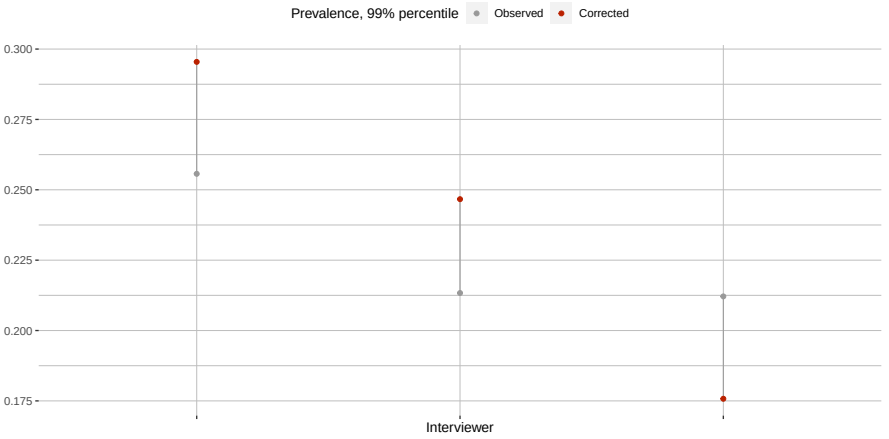
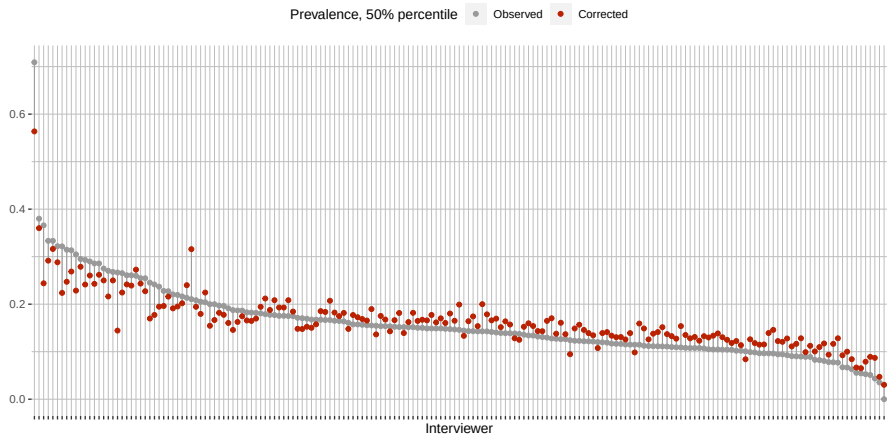
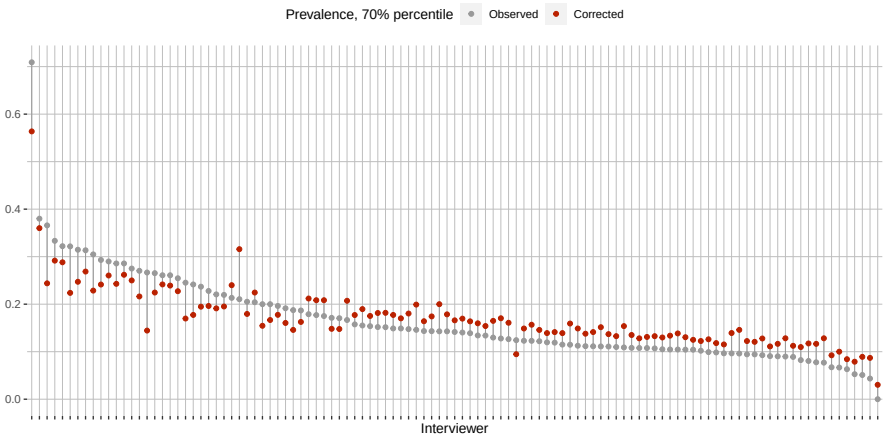


Figure 5.3: NIDS: Observed and adjusted interviewer specific prevalences of hypertension, 50%, 30%, 10%, 1% of cases subject to largest adjustment effects.

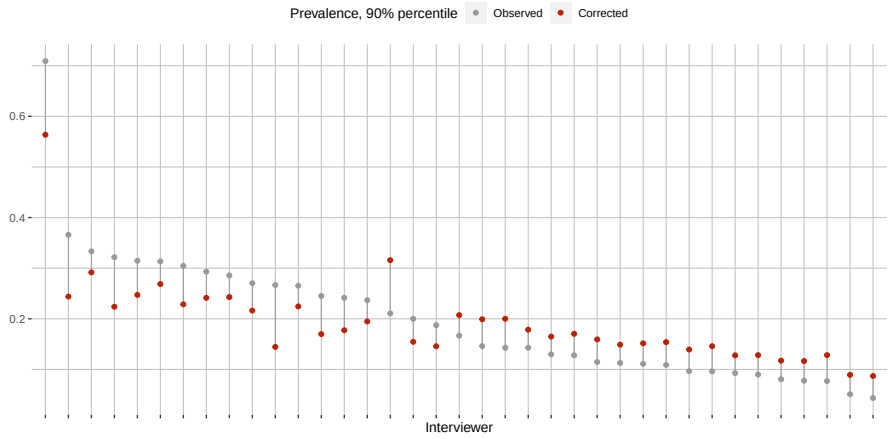
(a) 50th percentile



(b) 70th percentile



(c) 90th percentile



(d) 99th percentile

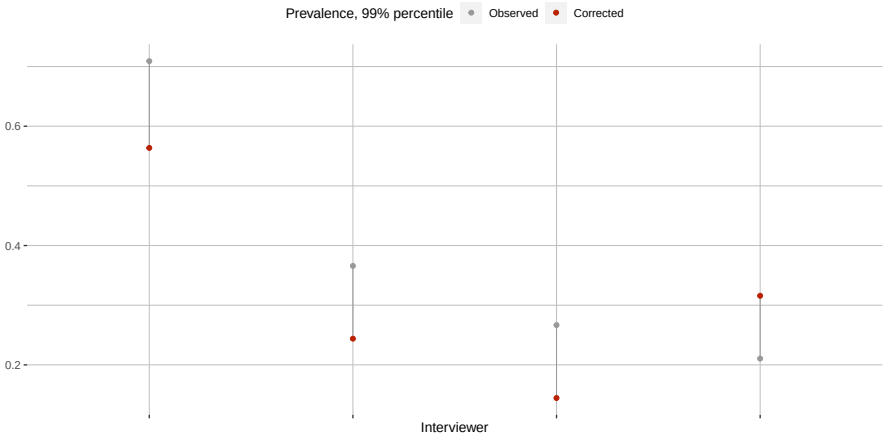
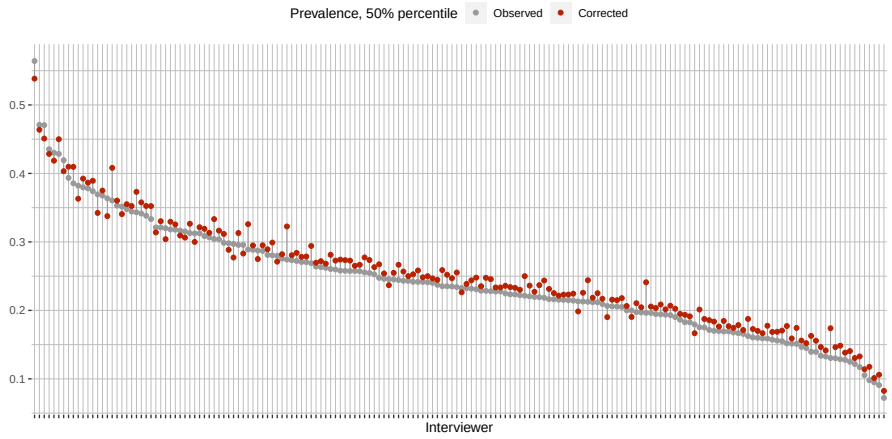
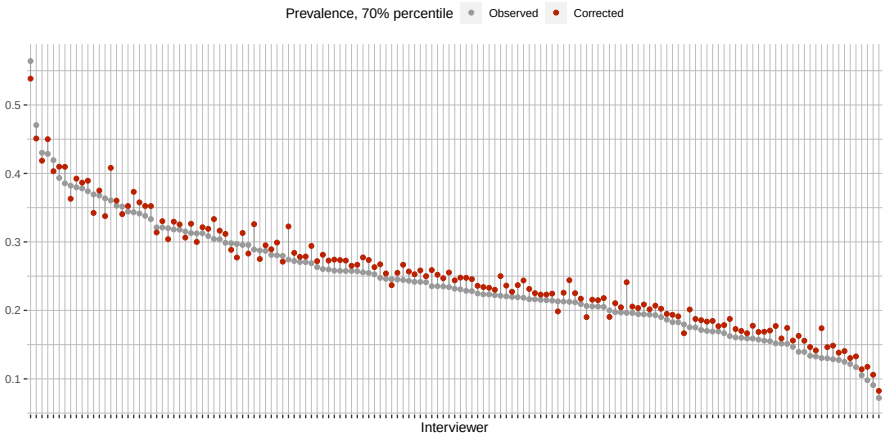


Figure 5.4: LASI: Observed and adjusted interviewer specific prevalences of hypertension, 50%, 30%, 10%, 1% of cases subject to largest adjustment effects.

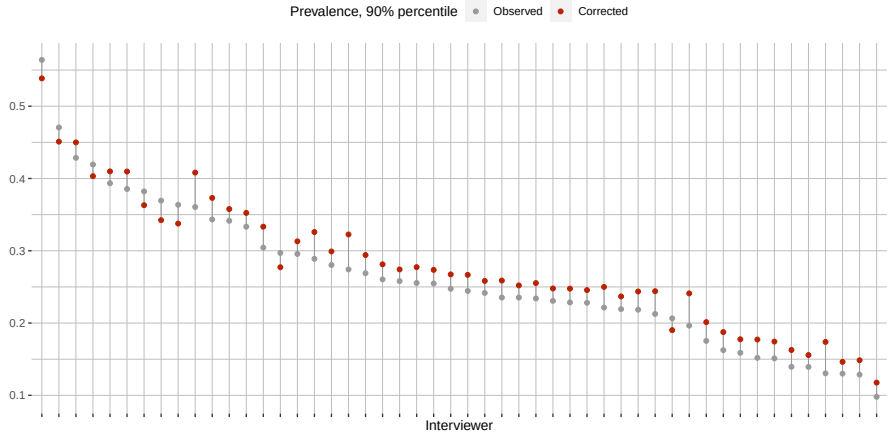
(a) 50th percentile



(b) 70th percentile



(c) 90th percentile



(d) 99th percentile

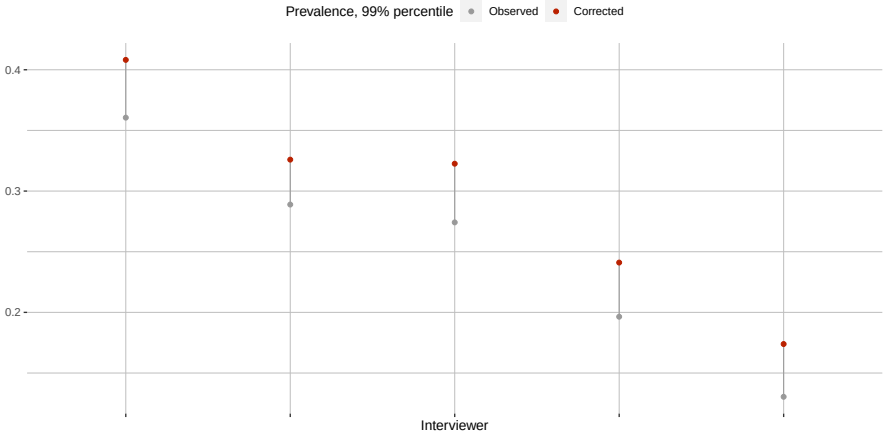
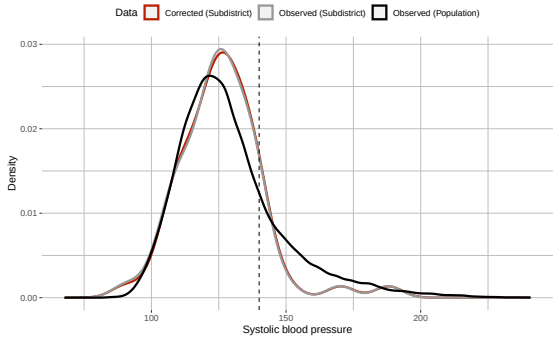
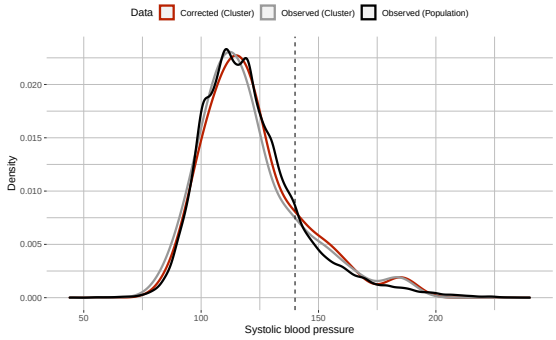


Figure 5.5: Systolic blood pressure densities, observed and adjusted for estimated interviewer effects, for selected subdistricts subject to large adjustment induced changes by data source. Population densities are added as comparison.

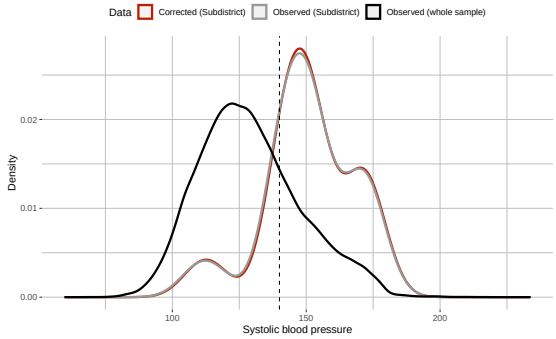
(a) IFLS



(b) NIDS



(c) IFLS



Bibliography

Acemoglu, D., Lelarge, C., & Restrepo, P. (2020). Competing with robots: Firm-level evidence from france. , *110*, 383–88.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, *128*(6), 2188–2244.

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, *83*(6), 2411–2451.

Aghion, P., Antonin, C., Bunel, S., & Jaravel, X. (2020). What are the labor and product market effects of automation? new evidence from france.

Aghion, P., Antonin, C., Bunel, S., & Jaravel, X. (2022). The effects of automation on labor demand: A survey of the recent literature.

Ali, S., & Rouse, A. (2002). Practice audits: reliability of sphygmomanometers and blood pressure recording bias. *Journal of human hypertension*, *16*(5), 359–361.

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, 211–217.

Andrews, D., Criscuolo, C., & Gal, P. (2016). The global productivity slowdown, technology divergence and public policy: a firm level perspective. *Brookings Institution Hutchins Center Working Paper*, *24*.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Archibugi, D., & Coco, A. (2004). A new indicator of technological capabilities for developed and developing countries (arco). *World Development*, *32*(4), 629-654. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0305750X04000051` doi: https://doi.org/10.1016/j.worlddev.2003.10.008

Artuc, E., Bastos, P., & Rijkers, B. (2019). Robots, tasks and trade. *World Bank Policy Research Working Paper*(WPS 8674).

Artuc, E., Bastos, P., & Rijkers, B. (2023). Robots, tasks, and trade. *Journal of International Economics*, 103828.

Atkeson, A., & Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, *98*(5), 1998–2031.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2020, May). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics*, *135*(2), 645–709. Retrieved 2020-09-11, from `https://academic.oup.com/qje/article/135/2/645/5721266` (Publisher: Oxford Academic) doi: 10.1093/qje/qjaa004

Autor, D., & Salomons, A. (2018). *Is automation labor-displacing? productivity growth, employment, and the labor share* (Tech. Rep.). National Bureau of Economic Research.

Autor, D. H., Dorn, D., & Hanson, G. H. (2016). The china shock: Learning from labor-market adjustment to large changes in trade. *Annu. Rev. Econ.*, *8*, 205–240.

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58.

Bai, J., & Ng, S. (2017). Principal components and regularized estimation of factor models. *arXiv preprint arXiv:1708.08137*.

Bai, J., Ng, S., et al. (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, *3*(2), 89–163.

Baker, S. R., & Bloom, N. (2013). *Does uncertainty reduce growth? using disasters as natural experiments* (Tech. Rep.). National Bureau of Economic Research.

Bartik, T. J. (1991). Who benefits from state and local economic development policies?

Berg, A., Bounader, L., Gueorguiev, N., Miyamoto, H., Moriyama, K., Nakatani, R., & Zanna, L.-F. (2021). *For the benefit of all: Fiscal policies and equity-efficiency trade-offs in the age of automation* (Tech. Rep.). International Monetary Fund.

Bloom, N. (2009). The impact of uncertainty shocks. *econometrica*, *77*(3), 623–685.

Bloom, N. (2014). Fluctuations in uncertainty. *Journal of economic Perspectives*, *28*(2), 153–176.

Boerma, J. T., Ghys, P. D., & Walker, N. (2003). Estimates of hiv-1 prevalence from national population-based surveys as a new gold standard. *The Lancet*, *362*(9399), 1929–1931.

Bogan, B., Kritzer, S., & Deane, D. (1993). *Nursing student compliance to standards for blood pressure measurement* (Vol. 32) (No. 2). SLACK Incorporated Thorofare, NJ.

Boivin, J., & Ng, S. (2005). *Understanding and comparing factor-based forecasts.* National Bureau of Economic Research Cambridge, Mass., USA.

Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, *132*(1), 169–194.

Bonfiglioli, A., Crinò, R., Fadinger, H., & Gancia, G. (2020). Robot imports and firm-level outcomes.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

Cepni, O., Guney, I. E., & Swanson, N. R. (2020). Forecasting and nowcasting emerging market gdp growth rates: The role of latent global economic policy uncertainty and macroeconomic data surprise factors. *Journal of Forecasting*, *39*(1), 18–36.

Cernat, A., & Sakshaug, J. W. (2020). Nurse effects on measurement error in household biosocial surveys. *BMC Medical Research Methodology*, *20*(1), 1–9.

Cernat, A., & Sakshaug, J. W. (2021). Interviewer effects in biosocial survey measurements. *Field Methods*, *33*(3), 236–252.

Clark, A., & Sanderson, C. (2009). Timing of children's vaccinations in 45 low-income and middle-income countries: an analysis of survey data. *The Lancet*, *373*(9674), 1543–1549.

Cockburn, N., Flood, D., Seiglie, J. A., Manne-Goehler, J., Aryal, K., Karki, K., ... others (2023). Health service readiness to provide care for hiv and cardiovascular disease risk factors in low-and middle-income countries. *PLOS Global Public Health*, *3*(9), e0002373.

Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. (2012). Demographic and health surveys: a profile. *International journal of epidemiology*, *41*(6), 1602–1613.

Crainiceanu, C. M. (2008). Likelihood ratio testing for zero variance components in linear mixed models. In D. B. Dunson (Ed.), *Random effect and latent variable model selection* (pp. 3–17). New York, NY: Springer New York. Retrieved from `https://doi.org/10.1007/978-0-387-76721-5_1` doi: 10.1007/978-0-387-76721-5_1

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, *66*, 165 – 185.

Crainiceanu, C. M., & Ruppert, D. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, *92*, 91 – 103.

Dauth, W., Findeisen, S., Suedekum, J., & Woessner, N. (2021). The adjustment of labor markets to robots. *Journal of the European Economic Association*, *19*(6), 3104–3153.

De Benedictis, L., & Tajoli, L. (2007a). Economic integration and similarity in trade structures. *Empirica*, *34*(2), 117–137.

De Benedictis, L., & Tajoli, L. (2007b). Openness, similarity in export composition, and income dynamics. *The Journal of International Trade & Economic Development*, *16*(1), 93–116.

De Loecker, J., & Eeckhout, J. (2018, June). *Global Market Power* (Working Paper No. 24768). National Bureau of Economic Research. Retrieved 2020-03-18, from `http://www.nber.org/papers/w24768` (Series: Working Paper Series) doi: 10.3386/w24768

De Loecker, J., Eeckhout, J., & Unger, G. (2020, May). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics*, *135*(2), 561–644. Retrieved 2020-09-11, from `https://academic.oup.com/qje/article/135/2/561/5714769` (Publisher: Oxford Academic) doi: 10.1093/qje/qjz041

De Loecker, J., & Warzynski, F. (2012, May). Markups and Firm-Level Export Status. *American Economic Review*, *102*(6), 2437–2471. Retrieved 2020-03-18, from `https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2437` doi: 10.1257/aer.102.6.2437

de Nigris, S., Haarburger, R., Hradec, J., Craglia, M., & Nepelski, D. (2022). *Ai watch: Ai uptake in manufacturing* (Tech. Rep.). Joint Research Centre (Seville site).

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Dickson, B. K., & Hajjar, I. (2007). Blood pressure measurement education and evaluation program improves measurement accuracy in community-based nurses: a pilot study. *Journal of the American Academy of Nurse practitioners*, *19*(2), 93–102.

Diez, M. F. J., Fan, J., & Villegas-Sánchez, C. (2019). *Global Declining Competition.* International Monetary Fund. (Google-Books-ID: tuKYDwAAQBAJ)

Dinlersoz, E., & Wolf, Z. (2018, September). *Automation, Labor Share, and Productivity: Plant-Level Evidence from U.S. Manufacturing* (Working Papers No. 18-39). Center for Economic Studies, U.S. Census Bureau.

Dorn, D., Katz, L. F., Patterson, C., Van Reenen, J., et al. (2017). Concentrating on the fall of the labor share. *American Economic Review*, *107*(5), 180–85.

Duch Brown, N., Gomez Losada, A., Miguez, S., Rossetti, F., & van Roy, V. (2023). *Ai watch-evolution of the eu market share of robotics* (Tech. Rep.). Joint Research Centre (Seville site).

Duch-Brown, N., & Haarburger, R. (2023). Expanding the industrial automation data universe: Prices, production, trade flows.

Duch-Brown, N., Rossetti, F., Haarburger, R., et al. (2021). *Evolution of the eu market share of robotics: Data and methodology* (Tech. Rep.). Joint Research Centre (Seville site).

Dwyer-Lindgren, L., Cork, M. A., Sligar, A., Steuben, K. M., Wilson, K. F., Provost, N. R., ... others (2019). Mapping hiv prevalence in sub-saharan africa between 2000 and 2017. *Nature*, *570*(7760), 189–193.

Edmond, C., Midrigan, V., & Xu, D. Y. (2015). Competition, markups, and the gains from international trade. *American Economic Review*, *105*, 3183–3221.

Egger, P., & Larch, M. (2008). Interdependent preferential trade agreement memberships: An empirical analysis. *Journal of International Economics*, *76*(2), 384–399.

Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., & Murray, C. J. (2002). Selected major risk factors and global and regional burden of disease. *the lancet*, *360*(9343), 1347–1360.

Fielding, A. (2004). The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed. *Multilevel Modelling Newsletter*, *16*, 3-9.

Flint, A. C., Conell, C., Ren, X., Banki, N. M., Chan, S. L., Rao, V. A., ... Bhatt, D. L. (2019). Effect of systolic and diastolic blood pressure on cardiovascular outcomes. *New England Journal of Medicine*, *381*(3), 243–251.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, *114*, 254–280.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, *89*(425), 122–127.

Graetz, G., & Michaels, G. (2018). Robots at work. *Review of Economics and Statistics*, *100*(5), 753–768.

Graetz, N., Friedman, J., Osgood-Zimmerman, A., Burstein, R., Biehl, M. H., Shields, C., ... others (2018). Mapping local variation in educational attainment across africa. *Nature*, *555*(7694), 48–53.

Greven, S., Crainiceanu, C. M., Küchenhoff, H., & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, *17*(4), 870–891. Retrieved from `http://www.jstor.org/stable/25651233`

Grossman, G. M., & Oberfield, E. (2021). *The elusive explanation for the declining labor share* (Tech. Rep.). National Bureau of Economic Research.

Gupta, R., & Kabundi, A. (2011). Forecasting macroeconomic variables using large datasets: dynamic factor model versus large-scale bvars. *Indian Economic Review*, 23–40.

Haarburger, R., Unger, F., & Stemmler, H. (2023). Taking over the world? on robots and market power.

Harrigan, J., & Reshef, A. (2015). Skill-biased heterogeneous firms, trade liberalization and the skill premium. *Canadian Journal of Economics/Revue canadienne d'économique*, *48*(3), 1024–1066.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2).

Hodges, J. S. (2013). *Richly parameterized linear models.* New York, NY: CRC Press.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia ii: A program for missing data. *Journal of statistical software*, *45*, 1–47.

Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). New York, NY: Cambridge University Press.

Humlum, A. (2019). Robot adoption and labor market dynamics. *Princeton University*.

ILOSTAT. (2022). *Ilo modelled estimates database,.* Retrieved 2022-31-01, from `https://ilostat.ilo.org/data/`

International Federation of Robotics. (2018). *World robotics* (Tech. Rep.).

International Institute for Population Sciences (IIPS), MoHFW, Harvard T. H. Chan School of Public Health (HSPH) and the University of Southern California (USC). (2020). *Longitudinal ageing study in india (lasi) wave 1, 2017-18, india report.* International Institute for Population Sciences, Mumbai.

Jaszczak, A., Lundeen, K., & Smith, S. (2009). Using nonmedically trained interviewers to collect biomeasures in a national in-home survey. *Field methods*, *21*(1), 26–48.

Jeet, G., Thakur, J., Prinja, S., & Singh, M. (2017). Community health workers for non-communicable diseases prevention and control in developing countries: evidence and implications. *PloS one*, *12*(7), e0180640.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *31*(3), 300–303.

Karabarbounis, L., & Neiman, B. (2014). The global decline of the labor share. *The Quarterly journal of economics*, *129*(1), 61–103.

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, *186*(2), 294–316.

Koch, M., Manuylov, I., & Smolka, M. (2019). *Robots and Firms* (SSRN Scholarly Paper No. ID 3377705). Rochester, NY: Social Science Research Network. Retrieved 2020-09-11, from https://papers.ssrn.com/abstract=3377705

Krenz, A., Prettner, K., & Strulik, H. (2021). Robots, reshoring, and the lot of low-skilled workers. *European Economic Review*, *136*, 103744.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, *54*(1-3), 159–178.

Lashkari, D., Bauer, A., & Boussard, J. (2018). Information technology and returns to scale. *Available at SSRN 3458604*.

Leibbrandt, M., Woolard, I., & de Villiers, L. (2009). Methodology: Report on nids wave 1. *Technical paper*, *1*.

Li, J., & Chen, W. (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, *30*(4), 996–1015.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mbondji, P. E., Kebede, D., Soumbey-Alley, E. W., Zielinski, C., Kouvividila, W., & Lusamba-Dikassa, P.-S. (2014). Health information systems in africa: descriptive analysis of data sources, information products and health statistics. *Journal of the Royal Society of Medicine*, *107*(1_suppl), 34–45.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, *71*(6), 1695–1725. Retrieved 2020-04-02, from https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00467 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00467) doi: 10.1111/1468-0262.00467

Moon, T. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47-60. doi: 10.1109/79.543975

Müller, C., & Kutzbach, N. (2019). World robotics 2019–industrial robots. *IFR Statistical Department, VDMA Services GmbH, Frankfurt am Main, Germany*, 45.

OECD. (2015). Key short-term indicators. Retrieved from https://www.oecd-ilibrary.org/content/data/data-00039-en doi: https://doi.org/https://doi.org/10.1787/data-00039-en

OECD. (2021). *Oecd inter-country input-output (icio) tables.* Retrieved from `http://oe.cd/icio`

OECD. (2023a). *National accounts of oecd countries, volume 2022 issue 2.* Retrieved from `https://www.oecd-ilibrary.org/content/publication/3e073951-en` doi: https://doi.org/https://doi.org/10.1787/3e073951-en

OECD. (2023b). *Oecd annual labor force statistics (alfs).* Retrieved from `https://stats.oecd.org/Index.aspx?DataSetCode=ALFS_EMP`

Osgood-Zimmerman, A., Millear, A. I., Stubbs, R. W., Shields, C., Pickering, B. V., Earl, L., ... others (2018). Mapping child growth failure in africa between 2000 and 2015. *Nature*, *555*(7694), 41–47.

Otieno, C., Kaseje, D., Ochieng', B., & Githae, M. (2012). Reliability of community health worker collected data for planning and policy in a peri-urban area of kisumu, kenya. *Journal of community health*, *37*, 48–53.

Peterson, R. A. (2021). Finding optimal normalizing transformations via best normalize. *R Journal*, *13*(1).

Porshakov, A., Ponomarenko, A., Sinyakov, A., et al. (2016). Nowcasting and short-term forecasting of russian gdp with a dynamic factor model. *Journal of the New Economic Association*, *30*(2), 60–76.

Prettner, K., & Strulik, H. (2017). The lost race against the machine: Automation, education, and inequality in an r&d-based growth model. *Education, and Inequality in an R&D-Based Growth Model (December 1, 2017). cege Discussion Papers*(329).

Rahim, N. E., Flood, D., Marcus, M., Theilmann, M., Aung, T. N., Agoudavi, K., ... others (2023b). Individual-level diabetes prevention activities in 44 low-and middle-income countries: a cross-sectional analysis of nationally representative, individual-level data in 145,739 adults. *Lancet Global Health*.

Rahim, N. E., Flood, D., Marcus, M. E., Theilmann, M., Aung, T. N., Agoudavi, K., ... others (2023a). Diabetes risk and provision of diabetes prevention activities in 44 low-income and middle-income countries: a cross-sectional analysis of nationally representative, individual-level survey data. *The Lancet Global Health*, *11*(10), e1576–e1586.

Rao, J. N., & Molina, I. (2015). *Small area estimation.* John Wiley & Sons.

Reiner Jr, R. C., Graetz, N., Casey, D. C., Troeger, C., Garcia, G. M., Mosser, J. F., ... others (2018). Variation in childhood diarrheal morbidity and mortality in africa, 2000–2015. *New England Journal of Medicine*, *379*(12), 1128–1138.

Rhoades, S. A. (1993). The herfindahl-hirschman index. *Fed. Res. Bull.*, *79*, 188.

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology* (Vol. 3). Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Schulze, M. B., Kroke, A., Bergmann, M. M., & Boeing, H. (2000). Differences of blood pressure estimates between consecutive measurements on one occasion: implications for inter-study comparability of epidemiologic studies. *European journal of epidemiology*, *16*, 891–898.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, *8 1*, 289-317.

Sikoki, B. S., Witoelar, F., Strauss, J., Meijer, E., & Suriastini, N. W. (2013). *Indonesia family life survey east 2012: User's guide and field report* (Tech. Rep.). SurveyMETER.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, 1–48.

Singh, P., & Sachs, J. D. (2013). 1 million community health workers in sub-saharan africa by 2015. *The Lancet*, *382*(9889), 363–365.

Sobczyk, P., Bogdan, M., & Josse, J. (2017). Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood. *Journal of Computational and Graphical Statistics*, *26*(4), 826-839. Retrieved from `https://doi.org/10.1080/10618600.2017.1340302` doi: 10.1080/10618600.2017.1340302

Southern Africa Labour and Development Research Unit. (2018a). Development research unit. national income dynamics study 2008, wave 1 [dataset]. *Southern Africa Labour and Development Research Unit (SALDRU), editor. Version*, *1*.

Southern Africa Labour and Development Research Unit. (2018b). Development research unit. national income dynamics study 2010-2011, wave 2 [dataset]. *Southern Africa Labour and Development Research Unit (SALDRU), editor. Version*, *1*.

Southern Africa Labour and Development Research Unit. (2018c). Development research unit. national income dynamics study 2012, wave 3 [dataset]. *Southern Africa Labour and Development Research Unit (SALDRU), editor. Version*, *1*.

Southern Africa Labour and Development Research Unit. (2018d). Development research unit. national income dynamics study 2014-2015, wave 4 [dataset]. *Southern Africa Labour and Development Research Unit (SALDRU), editor. Version*, *1*.

Southern Africa Labour and Development Research Unit. (2018e). Development research unit. national income dynamics study 2017, wave 5 [dataset]. *Southern Africa Labour and Development Research Unit (SALDRU), editor. Version*, *1*.

Stiebale, J., Suedekum, J., & Woessner, N. (2020, July). *Robots and the Rise of European Superstar Firms* (SSRN Scholarly Paper No. ID 3661423). Rochester, NY: Social Science Research Network. Retrieved 2020-10-01, from `https://papers.ssrn.com/abstract=3661423`

Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.

Stock, J. H., & Watson, M. W. (2011). Dynamic factor models.

Stock, J. H., & Watson, M. W. (2012). *Disentangling the channels of the 2007-2009 recession* (Tech. Rep.). National Bureau of Economic Research.

Strandberg, T. E., Salomaa, V. V., Vanhanen, H. T., Pitkälä, K., & Miettinen, T. A. (2002). Isolated diastolic hypertension, pulse pressure, and mean arterial pressure as predictors of mortality during a follow-up of up to 32 years. *Journal of hypertension*, *20*(3), 399–404.

Strauss, J., Witoelar, F., & Sikoki, B. (2016). *The fifth wave of the indonesia family life survey: overview and field report* (Vol. 1). Rand Santa Monica, CA, USA.

Strauss, J., Witoelar, F., Sikoki, B., & Wattie, A. M. (2009). *The fourth wave of the indonesia family life survey: Overview and field report.* RAND Labor and Population Working Paper WR-675/1-NIA/NICHD. Santa Monica, CA ....

Svensson, J. C., & Theorell, T. (1982). Cardiovascular effects of anxiety induced by interviewing young hypertensive male subjects. *Journal of Psychosomatic Research*, *26*(3), 359–370.

Ties Boerma, J., & Sommerfelt, A. E. (1993). Demographic and health surveys (dhs: contributions and limitations. *World health statistics quarterly 1993; 46 (4): 222-226*.

Ulijaszek, S. J., & Kerr, D. A. (1999). Anthropometric measurement error and the assessment of nutritional status. *British Journal of Nutrition*, *82*(3), 165–177.

UN, C. (1990). The united nations commodity trade statistics database. *http://comtrade.un. org/*.

UNCDAT. (2018). *Unctad trains: The global database on non-tariff measures.* Retrieved from `https://databank.worldbank.org/source/unctad-%5E-trade-analysis-information-system-(trains)`

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*(3), 426–482.

Wittenberg, M. (2009). Weights: report on nids wave 1. *NIDS Technical Paper*, *2*.

Yusuf, S., Joseph, P., Rangarajan, S., Islam, S., Mente, A., Hystad, P., ... others (2020). Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (pure): a prospective cohort study. *The Lancet*, *395*(10226), 795–808.

Zhou, B., Bentham, J., Di Cesare, M., Bixby, H., Danaei, G., Cowan, M. J., ... others (2017). Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19· 1 million participants. *The Lancet*, *389*(10064), 37–55.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*(2), 265–286.