Institutions, Inequality and Societal Transformations

Moricz, Sara

2019

Institutions, Inequality and Societal Transformations

Sara Moricz

Lund Economic Studies

Number 215

Lund University

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Lund University
Department of Economics
ISBN 978-91-7895-084-3
ISSN 0460-0029

Institutions, Inequality and Societal Transformations

Institutions, Inequality and Societal Transformations

Sara Moricz

LUND UNIVERSITY

DOCTORAL DISSERTATION
By due permission of the School of Economics and Management,
Lund University, Sweden.
To be defended at Holger Crafoords Ekonomicentrum EC3:207
on Wednesday June 5, 2018, at 14.15 pm.

Faculty Opponent
Andreas Kotsadam,
University of Oslo

Organization
LUND UNIVERSITY
Department of Economics
P.O. Box 7082
S-220 07 Lund, Sweden

Document name
DOCTORAL DISSERTATION

Date of issue
June 5, 2019

CODEN:
ISSN 0460-0029

Author
Sara Moricz

Sponsoring organization

Title and subtitle
Institutions, Inequality and Societal Transformations

Institutions matter for economic development. This thesis consists of three self-contained articles which provide different contributions to institutional economics.

The first article studies short-run changes in gender norms. It takes advantage of recent developments in machine learning algorithms to study changes in norms in Swedish tweets. More specifically, gender norms are defined in a comprehensive data-driven manner by training an LSTM neural network model. The model functions as a tool to create a measure on gender norms experienced by people in their everyday life, which is used to study norm-changes in relation to the Metoo movement. After an American actress encouraged women to share their experiences of sexual harassment under the Metoo hashtag, the Metoo movement spread to Sweden the 17th of October 2017. The event's date of occurrence is unrelated to the Swedish gender norm environment and, thus, provides a credible context for identifying an effect. Tweets with the Metoo-hashtag and related hashtags are removed from the analysis to ensure not to capture the effect of a more intensive gender debate. The article shows that Swedish tweets reflect gender norms less six months after the Metoo event, compared to five months before. The norm changes of the previous year function as a comparison group. The result is robust to the inclusion of calendar day and Twitter-user fixed effects. The result is also supported with placebo tests on an unrelated norm which is not expected to change. The article provides an example of norms changing rapidly, raising questions about previous literature's conceptualization of norms as being constant.

The second article investigates the effect of changing from an indirect (parliamentary) to a direct (presidential) democratic system on economic growth in Indonesia. In an indirect system, the local leader is appointed by the local parliament, whereas, in a direct system, the local leader is elected by the people. The main contribution of the article is to have a credible identification strategy to investigate the effect of a change in one specific political institution. The date of switching from the indirect to the direct democratic system differed for the various districts. The date of the switch is dependent on when the former dictator Suharto installed district heads, and therefore one expects no inherent differences in the growth rates of the districts other than from the change of system. The article relies on comparing districts that had direct elections earlier to districts that had them later. It shows no effect on economic growth. The results suggest that the form of democracy has a limited impact on governance, which is confirmed by examining a large number of indicators on governance (such as, the district head being corrupt). The result points to no effects of the forms of democracy on growth, which can inform policy regarding the choice of institutional design.

The third and last article presents new estimates of municipality-level income inequality for Sweden in 1871 and 1892. There exists scant information on the evolution of inequality during the nineteenth century, especially on income inequality, and the article contributes with a descriptive piece of evidence from the early phase of the Swedish industrialization process. The new municipality-level estimates are based on data on the vote distribution. In the electoral system at this time, votes were allocated in proportion to people's income, which renders it possible to retrieve income figures from the data source. The article shows that the income share of the top one percent richest in the industrial sector in the median municipality increases between 1871 and 1892, whereas inequality in the agricultural sector is stable. The article complements prior research on national-level wealth inequality by strengthening the interpretation of increasing inequality in the second half of the nineteenth century. Researchers can use the new estimates, available under an open-access license online, in further studies.

Key words
Gender Norms, Metoo, Sweden, LSTM Neural Network, Asia, Indonesia, Democracy, Elections, Economic Growth, Income Inequality, Economic Development

Classification system and/or index terms (if any)
JEL Classification: B52, J16, Z13, H11, O10, O43, N33, D31, O15

Supplementary bibliographical information

Language
English

ISSN and key title
0460-0029 Lund Economic Studies no. 215

ISBN
978-91-7895-084-3 (print)
978-91-7895-085-0 (pdf)

Recipient's notes

Number of pages
182

Price

Security classification

Distribution by: Department of Economics, P.O. Box 7082, S-220 07 LUND, Sweden

Signature____________________ Date April 25, 2019

Institutions, Inequality and Societal Transformations

Sara Moricz

LUND UNIVERSITY

LUND ECONOMIC STUDIES NUMBER 215

Lund University School of Economics and Management, Department of Economics

ISBN: 978-91-7895-084-3 (print)
ISBN: 978-91-7895-085-0 (pdf)
ISSN: 0460-0029 Lund Economic Studies no. 215

Printed in Sweden by Media-Tryck, Lund University, Lund 2018

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is an environmentally
certified and ISO 14001:2015 certified
provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Contents

Swedish Abstract

Institutioner har betydelse för ekonomisk utveckling. Denna avhandling består av tre fristående artiklar som bidrar till olika grenar inom institutionell ekonomi.

I den första artikeln undersöks kortsiktiga förändringar av könsnormer. Artikeln tar tillvara på senare tids utveckling av maskininlärningsalgoritmer för att studera normförändringar i svenska tweets. Mer specifikt så tränas en *LSTM neural network*- modell till att definiera könsnormer på ett omfattande och datadrivet sätt. Modellen används som ett verktyg för att ta fram ett mått på könsnormer som människor möter i vardagslivet, vilket används för att studera normförändringar i samband med Metoo-rörelsen. Metoo startade den 17 Oktober 2017 efter att en amerikansk skådespelerska uppmuntrade kvinnor att dela med sig av sina erfarenheter av sexuellt ofredande. Datumet för händelsen är orelaterat till svenska könsnormer och därmed tillåter kontexten att en effekt skattas utan jämförelseproblem. Tweets som innehåller Metoo-hashtagen eller relaterade hashtags tas bort från analysen för att inte effekten av en mer intensiv könsdebatt ska mätas. Artikeln finner att svenska tweets reflekterar könsnormer i en mindre utsträckning sex månader efter Metoo-händelsen i jämförelse med fem månader innan. Föregående års normförändringar används som en jämförelsegrupp. Resultatet är robust gentemot att använda fixa effekter för kalenderdagar och Twitter-användare. Resultatet får även stöd genom att istället undersöka en orelaterad norm som inte borde förändras med placebo-test. Studien ger ett exempel på att normer kan förändras snabbt, vilket visar på att tidigare litteratur borde omvärdera sitt synsätt på normer som relativt oföränderliga.

Den andra artikeln undersöker effekten på ekonomisk tillväxt av att byta från ett indirekt (parlamentariskt) till ett direkt (presidentiellt) demokratiskt valsystem i Indonesien. Under indirekta val väljs den lokala ledaren av distriktsparlamenten och under direkta val väljs den lokala ledaren av befolkningen. Denna artikel avhjälper de jämförelseproblem som existerat i den tidigare litteraturen genom att studera förändringen i en specifik politisk institution. Förändringen av valsystemet skedde vid olika år för olika distrikt. Vilket år ett distrikt bytte valsystem beror på när den förre diktatorn Suharto installerade lokala ledare, och därmed förväntas inte distrikten ha några skillnader mellan sig som påverkar tillväxt utöver förändringen i valsystem. I artikeln jämförs tillväxt i distrikt som bytte valsystem tidigt gentemot de som bytte senare. Artikeln finner inga tillväxteffekter av förändringen. En möjlig förklaring till resultatet är att valsystem har ringa betydelse för förekomsten av pålitliga, opartiska och effektiva samhällsinstitutioner (*governance*). Artikeln finner stöd för detta genom att undersöka ett stort antal *governance*-indikatorer (t.ex. om den lokala ledaren anses korrupt). Resultatet pekar på att det inte finns några skillnader i tillväxt mellan indirekta och direkt valsystem, vilket kan användas vid framtida beslut angående institutionella ramverk.

Den tredje och sista artikeln estimerarar mått på inkomstojämlikhet i svenska kommuner för 1871 och 1892. Utvecklingen av ojämlikhet under 1800-talet är relativt okänd, speciellt med avseende på inkomst, och artikeln bidrar med en deskriptiv pusselbit från den tidiga fasen i den svenska industrialiseringsprocessen. De nya estimaten på kommunal nivå konstrueras från data på distributionen av röster vid kommunala val. I det kommunala valsystemet fick människor röster efter hur mycket de betalade i skatt, vilket gör det möjligt att härleda personers inkomst från datakällan. Artikeln finner att den rikaste procentenhetens andel av totalinkomsten inom den industriella sektorn ökar i mediankommunen, medan densamma inom jordbrukssektorn är oförändrad. Artikeln kompletterar den tidigare forskningen på förmögenhetsojämlikhet på nationell nivå genom att stärka resultatet angående ökande ojämlikhet under den senare delen av 1800-talet. Forskare kan använda de nya estimaten, tillgängliga under en open-source licens online, i vidare studier.

Keywords: Könsnormer, Metoo, Sverige, LSTM neural network, Asien, Indonesien, Demokrati, Val, Ekonomiskt tillväxt, Inkomstojämlikhet, Ekonomisk utveckling

JEL Classification: B52, J16, Z13, H11, O10, O43, N33, D31, O15

English Abstract

Institutions matter for economic development. This thesis consists of three self-contained articles which provide different contributions to institutional economics.

The first article studies short-run changes in gender norms. It takes advantage of recent developments in machine learning algorithms to study changes in norms in Swedish tweets. More specifically, gender norms are defined in a comprehensive data-driven manner by training an LSTM neural network model. The model functions as a tool to create a measure on gender norms experienced by people in their everyday life, which is used to study norm-changes in relation to the Metoo movement. After an American actress encouraged women to share their experiences of sexual harassment under the Metoo hashtag, the Metoo movement spread to Sweden the 17th of October 2017. The event's date of occurrence is unrelated to the Swedish gender norm environment and, thus, provides a credible context for identifying an effect. Tweets with the Metoo-hashtag and related hashtags are removed from the analysis to ensure not to capture the effect of a more intensive gender debate. The article shows that Swedish tweets reflect gender norms less six months after the Metoo event, compared to five months before. The norm changes of the previous year function as a comparison group. The result is robust to the inclusion of calendar day and Twitter-user fixed effects. The result is also supported with placebo tests on an unrelated norm which is not expected to change. The article provides an example of norms changing rapidly, raising questions about previous literature's conceptualization of norms as being constant.

The second article investigates the effect of changing from an indirect (parliamentary) to a direct (presidential) democratic system on economic growth in Indonesia. In an indirect system, the local leader is appointed by the local parliament, whereas, in a direct system, the local leader is elected by the people. The main con-

tribution of the article is to have a credible identification strategy to investigate the effect of a change in one specific political institution. The date of switching from the indirect to the direct democratic system differed for the various districts. The date of the switch is dependent on when the former dictator Suharto installed district heads, and therefore one expects no inherent differences in the growth rates of the districts other than from the change of system. The article relies on comparing districts that had direct elections earlier to districts that had them later. It shows no effect on economic growth. The results suggest that the form of democracy has a limited impact on governance, which is confirmed by examining a large number of indicators on governance (such as, the district head being corrupt). The result points to no effects of the forms of democracy on growth, which can inform policy regarding the choice of institutional design.

The third and last article presents new estimates of municipality-level income inequality for Sweden in 1871 and 1892. There exists scant information on the evolution of inequality during the nineteenth century, especially on income inequality, and the article contributes with a descriptive piece of evidence from the early phase of the Swedish industrialization process. The new municipality-level estimates are based on data on the vote distribution. In the electoral system at this time, votes were allocated in proportion to people's income, which renders it possible to retrieve income figures from the data source. The article shows that the income share of the top one percent richest in the industrial sector in the median municipality increases between 1871 and 1892, whereas inequality in the agricultural sector is stable. The article complements prior research on national-level wealth inequality by strengthening the interpretation of increasing inequality in the second half of the nineteenth century. Researchers can use the new estimates, available under an open-access license online, in further studies.

Keywords: Gender Norms, Metoo, Sweden, LSTM Neural Network, Asia, Indonesia, Democracy, Elections, Economic Growth, Income Inequality, Economic Development

JEL Classification: B52, J16, Z13, H11, O10, O43, N33, D31, O15

Acknowledgments

Fredrik Sjöholm encouraged me to apply to the PhD program when he was my master's thesis supervisor. I thought something along the lines of, "Well, it's probably like having any other job," and sent in my application. At the end of my master's studies, I also applied for other jobs but did not even manage to get any interviews before being offered a PhD position, which I accepted. I was quite excited about the opportunity; I would be able to work on whatever topic I thought was important for five years! In hindsight, I can see that I took full advantage of that opportunity. I have been in an environment that allows intellectual freedom and discovery, to a degree that I think is rare also in academic environments. I am deeply grateful to the people in our department for contributing to such an environment through informal discussions and interesting seminars. I constantly come up with new ideas, and during the years I have come to see Fredrik's task as my supervisor to keep checks on me. My friend and colleague Anders once said that the title of my thesis should be "Things that Sara thought were interesting at some point in time," which I thought was a very accurate description until two weeks ago. To my surprise, I realized when I wrote the introduction that there is a common thread to my papers, and I can thus conclude that Fredrik really did manage to keep me on track. To spend my time wisely has been a constant struggle, and I still greatly dislike the long soft deadlines that exist in a research environment. I like the feeling of getting things done, and as you read this, dear reader, it means that I have finally met the only hard deadline I have had during the past five years!

Not only am I one of those people who would not have gone into a PhD program without encouragement, but also, I quit it to move to my then Norwegian boyfriend's home village, putting me in the quite unique position of having experienced both quitting the program and re-starting it after a year's absence. I was

quite nervous about telling people around me that I was quitting. I had been given a great opportunity and giving it up obviously revealed that I put more value on family life. However, to my surprise, the junior and senior faculty members were very understanding and supportive, which allowed me to see my colleagues not only as experts but as nice, caring people too. Fredrik also had the privilege of spreading the rumor that I was quitting to start goat farming (I did not), which opened up many interesting discussions about goats. Living in a rural Norwegian village was very different from anything I could possibly have imagined, the relationship ended, and I wanted to re-start the PhD program. Again, I was very nervous; it is one thing to quit, but quite another to re-start! Not only had I been disloyal, it was also clear that I had made the wrong decision. Again, junior and senior faculty members surprised me; most people just seemed to think that it was nice that I had come back! Fredrik said something along the lines of, "Yeah, I don't want to tell you I told you so," with a smile in his eyes. I deeply appreciate my colleagues for being nice and friendly.

During the first and second years of the PhD program, a PhD student-supervisor matching process takes place. It was not at all obvious that Fredrik would become my supervisor. We are in different fields; I decided early on to work on inequality and Fredrik mostly does research on trade. Our common denominator was an interest in Indonesia, which was not a great overlap academically. In the end, I decided that I wanted Fredrik as my supervisor, and he accepted. My main reason for this decision was that Fredrik is a caring and supportive person, which I, in the end, decided was more important than academic research interests. Today, I am very happy that I made that decision. Fredrik, I am very grateful for you being you over the years!

I have also had the opportunity to make some great friends during my PhD. Lina, John, Karl, Margaret, and Anna, I greatly appreciate us keeping up our board game nights and our non-board game nights. I am also happy that Simon will move back to the region soon, so he can join us again! Anders, I'm very happy that we discovered our common interest in machine learning, and in drinking beer. Josefin, thank you for sharing your stories with such emotional vigor–it is so refreshing! Hampus, thank you for having heated discussions about causality/relevance with me. Polina, Pol, Yana, Dominika, Kristoffer, Hjördis and Caglar, thank you for many interesting discussions over the years.

When you become an older PhD student, your peers start moving on. Although it

was many years ago now, I still miss our discussions around the lunch table driven by Ross's knowledge of English wildlife names, allowing us to discuss–well–weird animals and plants. I also miss Jörgen's extensive knowledge of historical language developments, allowing us to discuss–well–exactly that. And I greatly miss Yana sitting around the lunch table, always looking very innocent, but letting slip small, nasty, yet funny comments.

My gratitude also goes out to the rest of the PhD students at the department, past and present: Claes, Jim, Osmis, Viroj, Aron, Usman, Valeria, Veronika, Milda, Björn, Erik, Thomas, Dinh-Vinh, Dominice, Sara, Sanna, Danial, Duc-Hong, Sandra, Matthew, Jonas, Zahra, Demid, Marco, Ovi, Linn, Marcus, Philipp, Devon, Adrian and Olga.

I had the great privilege of spending half a year of my PhD studies as a visiting PhD student at the Paris School of Economics. This was made possible by generous support from the Hedelius Foundation. I also have deep gratitude to Daniel Waldenström, who hosted me during my stay.

I thank Björn Tyrefors for providing excellent comments at my final seminar. The papers in this thesis have also benefited from comments from participants at many conferences and workshops and financial support from Stiftelsen för främjande av ekonomisk forskning vid Lunds universitet and The Siamon Foundation.

I am thankful for the support of Tommy Andersson, who served as Study Director for the PhD program during most of my studies. The wonderful administrative staff at the department also deserves great gratitude. Anna, Azra, Jenny, Marie, Nathalie, Mariana, and Rikke, thank you for your open-door policy, which fixes any issue that might arise!

Swedish Acknowledgments Sist men inte minst, jag vill verkligen tacka vänner och familj för att ni funnits där för mig genom åren. Jakke, jag vet att du anser att jag oftast hjälper dig. Därför hoppas jag att du kan ta fram det här stycket när du tänker det: Du är den enda personen som stampar in, motsätter dig vad jag säger och berättar saker om mig själv som jag inte vet! Det är ovärderligt, och du har oftast rätt. Du är oerhört bra på att se och tolka människor, inte bara mig. Jag är så glad för att jag har en livstid kvar med dig i mitt liv som kan guida mina beslut! Mina föräldrar, Lena och Robban, jag är också så otroligt glad för att ni finns där för mig. Ni har naturligtvis varit bra föräldrar, men efter att ha varit vuxen

en längre tid vid det här laget vill jag bara utrycka min tacksamhet över att ni är vettiga människor som det är roligt att umgås med! Viktor och Mia, Eva, Emma, Alex och Tom, Rebecca, Else-Britt och Peik, Kristian, Mia och Shuai, Maria och Kristofer, Jens och Emma, Amelie, Cristel och Ricky, tack för att ni funnits med på resan.

Lund, April, 2019
Sara

Introduction

Introduction

This thesis contains three self-standing empirical papers related to new institutional economics (NIE).

1 New Institutional Economics

In his landmark work, North (1990) argued in favor of incorporating institutions into the neoclassical economics framework, since they drive economic development. North (1990) defines institutions as 'the humanly devised constraints that shape human interaction' (p.3). Institutions can be informal, such as norms, or formal, such as laws.

Following Putnam et al.'s (1994) contribution, one research line within institutional economics focuses on interpersonal trust. The literature highlights that markets do not exist in a vacuum: it is easier for an economic exchange to take place if people trust each other. Countries where the average person is of the opinion that "most people can be trusted" exhibit higher economic development and growth (e.g. Knack and Keefer, 1997; Algan and Cahuc, 2010; Tabellini, 2010). Norms of trust and cooperation are upheld by individuals (e.g. Herrmann et al., 2008), but various formal institutions, such as an impartial juridical system, can facilitate economic exchanges (e.g. Knack and Keefer, 1995; La Porta et al., 1997; Acemoglu et al., 2001; Easterly and Levine, 2003; Rodrik et al., 2004).

The rise of institutional economics favors the concept that, to deliver development in low income countries, one first needs to get the institutions right. In the last 20 years, policy institutions, such as the Word Bank, have refocused policy with this aim in mind (e.g. the World Development Reports of 2002; 2004; 2011).

Institutions are conceptualized as evolving slowly. Engerman and Sokoloff (2002) argue that geographic factors give rise to various levels of inequality, which in turn give rise to institutions that retain those levels of inequality. For most countries, such institutions also hinder development. A complementary perspective is put forward by North et al. (2009) and Acemoglu and Robinson (2012). The authors argue that elites obstruct development in developing countries. They also find that the reason behind the development of Western European countries lies with the elites gradually managing to implement impersonal organizations in Britain – such as political parties and limited holding companies – as an intra-elite bargaining solution. The gradual transformation spurred economic development and further extension of elite rights to non-elites.

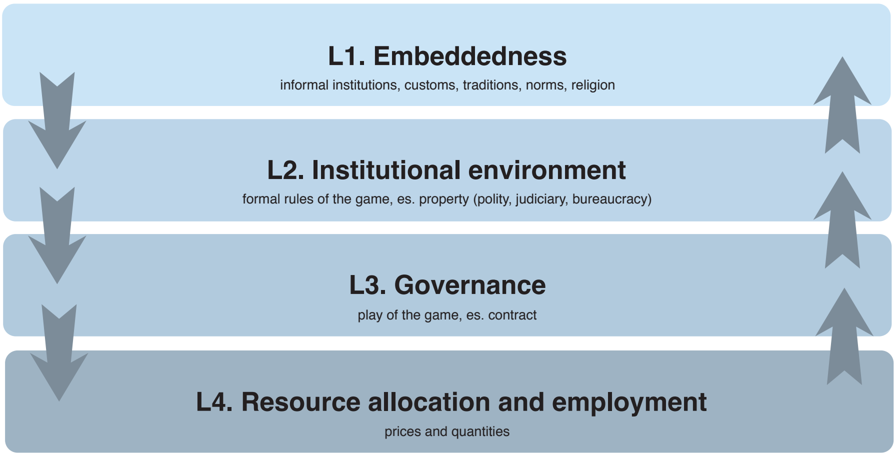


Figure 1: Williamson's conceptualization of institutions (2000, p. 597)

A conceptualization of institutions, development and change is provided by Williamson (2000), as illustrated in figure 1. He divides institutions into four levels, with feedback loops between them: the first level (*L1. Embeddedness*), which includes norms, changes over centuries or millennia, whereas the fourth level (*L4. Resource allocation and employment*) changes continuously. Williamson (2000) regards neoclassical economics as mostly dealing with the fourth level, whilst taking the other levels for granted. Since publishing his article, institutions have become more incorporated into the economic utility framework (e.g. Akerlof and Kranton, 2000; Bisin and Verdier, 2001). Furthermore, Williamson (2000) conceptualizes institutional economics as mostly dealing with *L2. The institutional environment* and

L3. Governance. Generally, institutional economics overlap with other sub-fields in economics, such as cultural economics and political economy. Institutions and change are wide-spanning concepts and the following paragraphs present the thesis papers' relationship to Williamson's (2000) framework.

The first paper, *Using Artificial Intelligence to Recapture Norms: Did #Metoo Change Gender Norms in Sweden?* investigates gender norms, an example of an institution (level 1). There might be norms surrounding what men and women are supposed to do, for example, men might be expected to be bread-winners while women might be expected to care for children. The paper investigates whether gender norms can change quickly, although existing literature usually finds support for institutional persistence. One related paper is Nunn and Wantchekon (2011), which shows that the slave trade caused lower levels of trust in modern-day Africa. Moreover, Voigtländer and Voth (2012) find that pogroms in relation to the Black Death are linked to Nazi support in the 1920s and 1930s. Finally, Guiso et al. (2016) investigates a part of Putnam et al.'s (1994) original hypothesis: that northern Italy is more developed than its southern counterpart, due to its stronger civic society, which arose by historically having free city states. They find that free cities in northern Italy, established in medieval times, have higher civic capital today.

The second paper, *Democracy and Economic Growth: Results from a Natural Experiment in Indonesia*, investigates whether the change from an indirect to a direct democratic system enhances economic growth (indicated by the arrows from levels 2 to 4 on the right-hand side). In an indirect system, the executive (the president) is indirectly elected by the legislature (the parliament); however, in a direct system, both the executive and the legislature are elected by the people. The paper isolates the effect of one formal institutional change on economic growth. Some authors argue that informal institutions affect formal ones (Greif, 1994), others that political institutions affect economic ones (Acemoglu and Robinson, 2012); however, few papers manage to isolate the effect of one specific institutional change in a credible way. One exception is Olken (2010), who shows that satisfaction and willingness to contribute to development projects increases with a direct decision-making mechanism at the village-level, compared to that of a representative one, in Indonesia. Moreover, Tyrefors and Pettersson-Lidbom (2014) find that welfare spending was lower under a direct democratic system, compared to that of a representative one, in Swedish municipalities in the 1920-40s. They attribute the result to the former being more prone to local elite capture.

The third and final paper, *Explorative Analysis of Municipality-level Inequality in Late Nineteenth Century Sweden*, investigates the change in income inequality in Swedish municipalities between 1871 and 1892. The feedback loop from higher institutional levels to lower ones (indicated by the arrow from level 4 to level 1 on the right-hand side) is rarely investigated; however, as mentioned previously, one potential factor that may explain why various growth-enhancing institutions arise in a country is the level of inequality. To further investigations of this hypothesis, the third paper contributes with descriptive evidence and a new dataset.

2 Applied Microeconometrics

Economics can be divided into two branches: theoretical and empirical. Compared to other subjects, empirical economics investigates questions related to aggregated human behavior and has a tradition of using observational (real-world) data, as opposed to using experimental data. Where other disciplines use experiments to identify causal mechanisms, empirical economists typically cannot, as it is hard to find an experimental design which mimics aggregate group behavior. The last 20 years, the methodological toolbox in Economics has grown to incorporate various methods striving to emulate experimental designs using observational data. Such estimation strategies are commonly referred to as applied microeconometrics and the first and second papers of the thesis are examples thereof. The push for using methods striving to emulate experimental designs leads to a greater focus on internal validity, ensuring that the effects measured are without bias. At the same time, this push leads to a lesser focus on external validity, ensuring that the effects measured are relevant to another context. Therefore, one can also discern a discussion on how to prioritize among internal and external validity in the profession. It might be the case that economics will always balance between internal and external validity, due to the form of research questions that define the discipline. Most likely, empirical economics will not be able to completely move over to using experimental-like research designs but will continue to use more descriptive methods: the third paper is an example thereof.

3 Overview of the thesis

Using Artificial Intelligence to Recapture Norms: Did #Metoo Change Gender Norms in Sweden?

The first paper contributes to the literature on institutional change through studying short-run changes in gender norms. Gender norms, although hard to prove empirically, likely affect the overall economy (Spencer et al., 1999; Mueller and Plug, 2006; Bowles et al., 2007; Guiso et al., 2008; Fryer and Levitt, 2010; Pope and Sydnor, 2010; Leibbrandt and List, 2014; Bertrand et al., 2015; Mazei et al., 2015; Nollenberger et al., 2016). Prior research portrays norms as changing slowly (Williamson, 2000) and provide evidence of gender norms being constant over generations (Fernández et al., 2004; Alesina et al., 2013; Nollenberger et al., 2016).

The paper takes advantage of the recent developments in machine learning algorithms to study norm changes in Swedish tweets. A more data-driven and comprehensive definition of gender norms is used, compared to prior research. The paper studies whether tweets reflect gender norms to a greater or smaller extent due to the Metoo event. On the 16th of October 2017, an American actress encouraged females to share their experiences of sexual harassment under the Metoo hashtag and, by the 17th of October 2017, the Metoo movement had spread to Sweden. The Metoo event is exogenous to the Swedish gender norm environment. The estimation strategy relies on comparing how gender norms change half a year before and after the Metoo event, while utilizing data from the previous year as a comparison group. The identification strategy resembles the one used in the difference-in-differences, or a regression discontinuity in time (RDIT), framework. Tweets that contain the Metoo hashtag and related hashtags are removed from the analysis. The paper shows that Swedish tweets reflect gender norms to a lesser degree in the six months after the metoo-event, compared to the five months before. The result is robust to, for example, including Twitter user fixed effect. The paper gives an example of gender norms changing in the short term, which raises the question of whether the literature's stance on norms being constant needs re-visiting.

Democracy and Economic Growth: Results from a Natural Experiment in Indonesia

The second paper investigates the effect of changing from an indirect to a direct democratic system on economic growth. Previous research does not provide any clear empirical results (Persson, 2005; Persson and Tabellini, 2006; Knutsen, 2011); however, Persson and Tabellini (2006) suggest that countries democratizing to become indirect (parliamentary) democracies grow, whereas direct (presidential) ones do not to the same degree.

It is difficult to untangle the relationships between political institutions, economic institutions, and development due to possible feedback loops between them. The major contribution of this paper is to have a credible identification strategy to investigate the effect of a change in one specific political institution. In Indonesia, the date for switching from an indirect to a direct democratic system differed for various districts. The date of the switch depends on when the former dictator, Suharto, historically installed district heads, which he did on a range of dates. The paper supports the date of the switch being exogenous, through testing pre-treatment variables related to governance and growth. It relies on comparing districts that have had direct elections, to those that have not yet had such elections, in the difference-in-differences framework. The paper shows that growth rates do not differ between districts with and without direct elections. The result is consistent with scant evidence of differences in governance between the districts. Hence, the paper suggests that the exact form of democracy not is crucial for economic growth.

Explorative Analysis of Municipality-level Inequality in Late Nineteenth Century Sweden

The third paper contributes to the Swedish development puzzle with one descriptive piece of evidence. Although the Scandinavian countries today have low levels of inequality compared to other countries, there is limited quantitative evidence to support that inequality was comparatively low historically. Bengtsson et al. (2018) show that wealth inequality in Sweden was lower than that in Great Britain and France, but higher than in the US, in 1750 and 1800 (the few countries where data exists). Wealth inequality increased relatively more in Sweden between 1850 and 1900, reaching the same levels as the other three countries in 1900. In contrast, during the 20th century, income and wealth inequality decreases from around 1900 to the 1970s and 80s (Roine and Waldenström, 2008, 2009).

The paper constructs inequality measures for Swedish municipalities in 1871 and 1892. The paper compares how the top one percent income share has changed within the municipalities between the two time points. The paper shows that inequality in the industrial sector increases in the median municipality, suggesting an upward trend in the early stages of the Swedish industrialization process. In contrast, it also demonstrates that inequality in the agricultural sector remains stable. The paper corroborates Bengtsson et al.'s (2018) finding, which supports the notion of Sweden being unusually equal historically. It also complements Roine and Waldenström's (2008) analysis in suggesting an upward trend in income inequality before the turn of the 20th century.

One difficulty with cross-country datasets used in the current inequality literature is the problem of isolating the effect of inequality in a credible way. Hopefully, the constructed inequality measures can be used in further studies on the effect of inequality on other municipality-level variables.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91(5):1369–1401.

Acemoglu, D. and Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. Crown Publishing Group.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Algan, Y. and Cahuc, P. (2010). Inherited Trust and Growth. *American Economic Review*, 100(5):2060–92.

Bengtsson, E., Missiaia, A., Olsson, M., and Svensson, P. (2018). Wealth Inequality in Sweden, 1750 - 1900. *The Economic History Review*, 71(3):772–794.

Bertrand, M., Kamenica, E., and Pan, J. (2015). Gender Identity and Relative Income within Households. *The Quarterly Journal of Economics*, 130(2):571–614.

Bisin, A. and Verdier, T. (2001). The Economics of Cultural Transmission and the Dynamics of Preferences. *Journal of Economic Theory*, 97(2):298–319.

Bowles, H. R., Babcock, L., and Lai, L. (2007). Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask. *Organizational Behavior and Human Decision Processes*, 103(1):84–103.

Easterly, W. and Levine, R. (2003). Tropics, Germs, and Crops: How Endowments Influence Economic Development. *Journal of Monetary Economics*, 50(1):3–39.

Engerman, S. L. and Sokoloff, K. L. (2002). Factor Endowments, Inequality, and Paths of Development Among New World Economics. NBER Working Paper 9259, National Bureau of Economic Research.

Fernández, R., Fogli, A., and Olivetti, C. (2004). Mothers and Sons: Preference Formation and Female Labor Force Dynamics. *The Quarterly Journal of Economics*, 119(4):1249–1299.

Fryer, R. G. and Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, 2(2):210–40.

Greif, A. (1994). Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies. *Journal of Political Economy*, 102(5):912–950.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, Gender, and Math. *Science*, 320(5880):1164–1165.

Guiso, L., Sapienza, P., and Zingales, L. (2016). Long-term Persistence. *Journal of the European Economic Association*, 14(6):1401–1436.

Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial Punishment Across Societies. *Science*, 319(5868):1362–1367.

Knack, S. and Keefer, P. (1995). Institutions and Economic Performance: Cross-country Tests Using Alternative Institutional Measures. *Economics & Politics*, 7(3):207–227.

Knack, S. and Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly journal of economics*, 112(4):1251–1288.

Knutsen, C. H. (2011). Which Democracies Prosper? Electoral Rules, Form of Government and Economic Growth. *Electoral Studies*, 30(1):83–90.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. W. (1997). Legal Determinants of External Finance. *The Journal of Finance*, 52(3):1131–1150.

Leibbrandt, A. and List, J. A. (2014). Do Women Avoid Salary Negotiations? Evidence from a Large-scale Natural Field Experiment. *Management Science*, 61(9):2016–2024.

Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., and Hertel, G. (2015). A Meta-analysis on Gender Differences in Negotiation Outcomes and Their Moderators. *Psychological Bulletin*, 141(1):85.

Mueller, G. and Plug, E. (2006). Estimating the Effect of Personality on Male and Female Earnings. *ILR Review*, 60(1):3–22.

Nollenberger, N., Rodríguez-Planas, N., and Sevilla, A. (2016). The Math Gender Gap: The Role of Culture. *American Economic Review*, 106(5):257–61.

North, D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

North, D. C., Wallis, J. J., Weingast, B. R., et al. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.

Nunn, N. and Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American Economic Review*, 101(7):3221–52.

Olken, B. A. (2010). Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia. *American Political Science Review*, 104(2):243–267.

Persson, T. (2005). Forms of Democracy, Policy and Economic Development. NBER Working Papers 11171, National Bureau of Economic Research.

Persson, T. and Tabellini, G. (2006). Democracy and Development: The Devil in the Details. *American Economic Review*, 96(2):319–324.

Pope, D. G. and Sydnor, J. R. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives*, 24(2):95–108.

Putnam, R. D., Leonardi, R., and Nanetti, R. Y. (1994). *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, N.J.

Rodrik, D., Subramanian, A., and Trebbi, F. (2004). Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development. *Journal of Economic Growth*, 9(2):131–165.

Roine, J. and Waldenström, D. (2008). The Evolution of Top Incomes in an Egalitarian Society: Sweden, 1903–2004. *Journal of Public Economics*, 92(1):366–387.

Roine, J. and Waldenström, D. (2009). Wealth Concentration over the Path of Development: Sweden, 1873–2006. *The Scandinavian Journal of Economics*, 111(1):151–187.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35(1):4–28.

Tabellini, G. (2010). Culture and Institutions: Economic Development in the Regions of Europe. *Journal of the European Economic association*, 8(4):677–716.

Tyrefors, B. H. and Pettersson-Lidbom, P. (2014). Democracy, Redistribution, and Political Participation: Evidence From Sweden 1919–1938. *Econometrica*, 82(3):961–993.

Voigtländer, N. and Voth, H.-J. (2012). Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany. *The Quarterly Journal of Economics*, 127(3):1339–1392.

Williamson, O. E. (2000). The New Institutional Economics: Taking Stock, Looking Ahead. *Journal of Economic Literature*, 38(3):595–613.

World Bank (2002). *World Development Report 2002 : Building Institutions for Markets*. Oxford University Press, New York.

World Bank (2004). *World Development Report 2005 : A Better Investment Climate for Everyone*. Oxford University Press, New York.

World Bank (2011). *World Development Report 2011 : Conflict, Security, and Development*. Oxford University Press, New York.

Paper I

Using Artificial Intelligence to Recapture Norms: Did #Metoo Change Gender Norms in Sweden?

Abstract

Norms are challenging to define and measure, but this paper takes advantage of text data and the recent development in machine learning to create an encompassing measure of norms. An LSTM neural network is trained to detect gendered language. The network functions as a tool to create a measure on how gender norms changes in relation to the Metoo movement on Swedish Twitter. This paper shows that gender norms on average are less salient half a year after the date of the first appearance of the hashtag #Metoo. Previous literature suggests that gender norms change over generations, but the current result suggests that norms can change in the short run.

Key Words: Gender norms, Metoo, Sweden, LSTM neural network

JEL Classification: B52, J16, Z13

1 Introduction

Norms are informal institutions thought to be passed down from parents to children through generations. But when do the children stop following the ways of their parents and start changing the norms? The Metoo movement might provide one such example, and maybe for the first time in history has such a popular movement left an easily accessible trace as more and more private conversations take place online on social media. This article contributes to institutional economics, firstly, by providing a method for automatically detecting norms in text data. Then, the article shows that Swedish Twitter reflects gender norms less in the period after the Metoo event.

Previous research conceptualizes norms as long-run phenomena and authors present evidence of gender norms being constant over generations (Williamson, 2000; Bisin and Verdier, 2001; Fernández et al., 2004; Alesina et al., 2013; Nollenberger et al., 2016). At the same time, gender norms seem to be important: gender norms explain a part of the gender gap in mathematical test scores (Spencer et al., 1999; Guiso et al., 2008; Fryer and Levitt, 2010; Pope and Sydnor, 2010; Nollenberger et al., 2016) and indications exist of them also affecting the gender gap in the propensity to negotiate (Bowles et al., 2007; Leibbrandt and List, 2014; Mazei et al., 2015).[1] Gender norms are hard to measure, which limits the research field. Bertrand et al. (2015) provide the most recent contribution to the measurement of gender norms. They ingeniously measure the norm "a man should earn more than the woman in a couple" directly from observational data. The usage of observational data overcomes social desirability bias: that participants in surveys or experiments act according to what they think is acceptable behavior. Unfortunately, the type of observational data they use only measures hard life choices, such as educational decisions and marriages. One major contribution of this article is to use observational data while comprehensively measuring gender norms. The comprehensive definition of gender norms includes the myriads of aspects defining the gender roles women and men face in their everyday life. The other major contribution of this article is to show that gender norms are less followed in Swedish tweets in the wake of the Metoo event, which complement existing studies by suggesting that gender norms can change in the short run.

[1]Some indication also exists of various personality traits, foremost agreeableness, being differently rewarded in the labor market depending on if it is a woman or man exhibiting it (Mueller and Plug, 2006).

Norms are informal institutions, and North defines institutions to be 'the humanly devised constraints that shape human interaction' (1990, p. 3). Akerlof and Kranton (2000) brought norms into the economic utility framework. In their identity framework, people belong to different social categories with different prescriptions for how they should behave. This paper postulates the following data generating process: unobserved gender norms give rise to observed gendered language, i.e. males and females are being depicted differently in texts. To obtain a tool for automatically detecting gender norms, I censor Swedish tweets by replacing all instances of the words *he* and *she* with a blank spot (___) and train a neural network algorithm to fill in the correct word. Neural networks are overarching machine learning algorithms which allow analysts to do broad searches for good models. They are state-of-the-art methods in natural language processing to translate between languages and capture the sentiment of a text. A person asked to fill in a blank spot with *he* or *she* would use previous knowledge of gender prescriptions to make a good guess. Instead, with my data setup, the training yields a model defining the gender prescriptions as the average of Twitter users' differential portrayals of males and females in the training tweets.

The current method entails an objective definition of gender norms. One dimension of gendered language (*he/she*) is used to infer the rest of the dimensions. The objective approach is important insofar that analysts can hold norms implicitly affecting their choices regarding what they count as a sub-norm. Bolukbasi et al. (2016) create an algorithm to remove unwanted gendered words in word vectors, which necessitates that an analyst provides a list of unwanted words. Word vectors are vector representation of words obtained through clustering, also used in this paper. Wu (2018) finds that words related to physical appearance are high-predictors for posts on the Economics Job Market Rumors forum and interprets such words as indicating the presence of gender norms. She uses a penalized logistic regression with the counts of words as the independent variables, which, again, is a model-technical building block in this paper. The main difference between the two other articles and this one is in the end goal: This article strives to encapsulate norms into a model to be used as an evaluation tool on a new sample in a research design.

I have downloaded all Swedish tweets from Twitters API; *he* and *she* are frequent words and the total sample size of the he/she dataset is large at 2 million tweets. Twitter data is written directly by people without an editor and norms are therefore more likely to be reflected in tweets than in, for example, newspapers. This paper contrasts an long short-term memory (LSTM) neural network algorithm to a baseline Naive Bayes algorithm. The main difference between them is that the neural network model has the ability to learn phrases, i.e. it takes word order into account, whereas the Naive Bayes model does not. In the main version of the models are highly-gendered words, such as *mum* and *dad*, censored to allow for a more nuanced definition of gender norms. The models are trained on 800 000 he/she tweets. On a new set of 200 000 out-of-sample tweets, the neural network model predicts 77 percent correct whereas the Naive Bayes model predicts 75 percent correct. Words related to sport are high predictors for males and words related to personal relationships are high predictors for females. Then, this model is used as a tool to measure the change in gender norms adherence in relation to the Metoo movement.

The Metoo movement started when an American actress asked other women to report misogynistic behavior. The hashtag peaked on Swedish Twitter on 17th of October 2017 and did not exist before that. The Metoo event provides a good context for evaluating if gender norms change rapidly, the exact date for its occurrence is exogenous while it spurred a collective response in the Swedish society. The movement lead to women sharing their experience of sexual harassment and it is likely that it leads to a norm shift towards more awareness and less acceptance of the sexualization of women. The movement also spurred an intensive public debate on gender issues and the norm shift may equally well include less acceptance towards stereotyping the genders in general. The current encompassing measure capture the possible norm shift. The model is used to investigate how much gender norms are followed by measuring its predictive accuracy, i.e. how often it predicts the correct word *he* or *she* in the blank spot in tweets. For example, let us assume that the model has encoded "struggle", "math" and "homework" as following a female gender stereotype and, before the event, tweets such as "my daughter she struggles with her math homework" are prevalent, then the model has high predictive accuracy indicating that gender norms are followed. If after Metoo many daughters are called "competing for good grades" instead and this is encoded as indicating a male gender stereotype, the model would detect the change by a decrease in its predictive accuracy.

The research design uses an additional 1 million out-of-sample he/she tweets from May 2017 to May 2018, data from five months leading up to the event to six months after, to evaluate the Metoo movement. The model is trained on data one year before, from May 2016 to May 2017, which defined gender norms to the average of this time period. Tweets that contain the hashtag #Metoo and correlated hashtags are filtered out to avoid capturing the effect of a more intensive gender debate. The Metoo movement is colinear with time and the measured effect always runs the risk of being driven by a third factor. To decrease such possibilities, the research design is such that the 200 000 out-of-samples tweets mentioned before (used to test model performance) match the year before and can be used as a comparison group. In my preferred specification, those tweets allow controlling flexibly for time trends by time fixed effects. The tweets include user identification numbers which allow for user fixed effects on the 35 000 users that tweets at least once on either side of the Metoo event as a robustness check. Before the Metoo event gender norms are on average more followed in tweets than the year before, but after the Metoo event my preferred estimate shows that norms on average are followed 1.9 percent less over the sample mean. The gender norms followed before the event on Swedish Twitter would be erased by 18 Metoo movements. The comparison is not normative since the data-driven definition of gender norms does not include judgment.

Swedish Twitter users are younger and slightly more male than the general population (ISS, 2016) rendering the norms and the effect on this demographic to be estimated. At the same time, younger persons are likely to have got a stronger experience of the Metoo event as it was largely an online movement. The result should be interpreted in the broader context of changing norms in the short run. The Metoo movement is an example of the newer kind of popular movement spread via social media, a form of bottom-up revolution. The current result of changes in the following of gender norms, therefore, points to online mass population movements affecting an overall norm in the short run. For the women and men participating in such movements, it is important to see that they do bring change.

The rest of the paper is organized as follows: Section 2 reviews articles investigating gender norms. Section 3 presents the proposed method for measuring norms and discusses how it differs compared to previous measures. Section 4 presents the data. Section 5 illustrates the neural network model, whereas Appendix A presents model selections and specifications in closer detail. Section 6 presents the model results. Section 7 describes the Metoo event in Sweden and Section 8 the research design behind the evaluation of the Metoo event. Section 9 presents the results on how Metoo have impacted gender norms and section 10 concludes.

2 Previous Literature

Norms are informal institutions. North define institutions to be 'the humanly devised constraints that shape human interaction' (1990, p. 3). Norms are informal since other people punish a person for breaking them, rather than a formal third-party, such as the police. In Akerlof and Kranton (2000)'s identity framework people belong to different social categories defined by prescriptions for how they should behave. Breaking the prescriptions affect one's own and other's utility, essentially imposing an externality, leading to people being punished for following a specific identity.

Norms are generally thought to be changing slowly, over the course of centuries or millennia (Williamson, 2000). Bisin and Verdier (2001) formalize theoretically how parents and the surrounding environment transmit norms to children. Alesina et al. (2013) show that gender roles transmit over long time periods; immigrant children of parents from societies in which the plow historically was used (a proxy for a norm of males working outside the household) think that a woman's place is in the household to a greater extent. Fernández et al. (2004) provide evidence of gender attitudes transferring within families; the wives of men whose mothers had been working during WW2 are more likely to participate in the American labor force.[2] At the same time, norms can change quickly by external shocks. Fortin (2015) shows how the evolution of gender norms and female labor force participation co-varies over time. She argues that the AIDS scare changed gender norm attitudes, explaining parts of the slow-down of the increasing trend in female labor force participation from the mid-1990s in the US. Goldin (2006) argues that young women in the 1960s anticipated a life in the labor market giving rise to a "revolutionary phase" observed in a multitude of variables from the late 1970s in the US. In this article, the newer method and the exogenous Metoo event makes it possible to investigate if gender norms can change in a shorter time span.

In the last twenty years, norms have been recognized as important in explaining economic phenomena, but the lack of available data on norms most likely hampers the research field in Economics. To my knowledge, Bertrand et al. (2015) is the first paper which investigates a gender norm directly from observational data. More specifically, they investigate how the norm "a man should earn more than his wife" affects a multitude of variables such as sorting into couples and labor mar-

[2] Nollenberger et al. (2016) corroborates transfers of attitudes within families.

ket participation. Strikingly, they show that the gender gap in non-market work increases with a rising share of female family income. Also, Wu (2018) uses observational data, in this case, blog posts from the Economics Job Market Rumors forum. She finds that words related to physical appearance frequently appear in posts related to females. A need for more observational data on norms, in general, exists to further the field. Observational data overcome social desirability bias, i.e. participants in a survey or an experiment change their answer or behavior due to being surveyed. In addition, it is difficult to obtain survey data. The lower half of Table 1 lists the two papers pertaining to gender norms reviewed which measure gender norms directly from surveys.

Many papers are reactionary in their research design; researchers try to disprove that various observed differences between men and women are due to biology. One observed difference between men and women is the gender gap in mathematical test scores, where boys usually achieve higher grades. The gap varies over states/countries and is positively correlated with measures of gender inequality (Guiso et al., 2008; Fryer and Levitt, 2010; Pope and Sydnor, 2010). Biological differences are unlikely to vary over states/countries and the pattern points to the importance of gender norms. Nollenberger et al. (2016) show that the gender gap decreases in second-generation immigrant children if their parents' source country is more gender equal. Their research design rules out possible confounding state/country factors, thereby strengthening the interpretation. Alike research designs are inventive, but the outcome variable becomes the gender dimension investigated, for instance, mathematical test scores, which unfortunately is quite limited in scope.

Table 1: Papers which measures norms directly

Paper	Norm	Measured by
Bertrand et al. (2015)	”a man should earn more than his wife”	relative income between spouses
Wu (2018)	”the gender norm among economists”	gendered language on the Economics Job Market Rumors forum
Alesina et al. (2013)	”women should stay in the household”	historical plough usage, female labour force participation, share of female firm owners, proportion of women in parliament, two questions from World Values Survey*
Fortin (2015)	”women should stay in the household”	index variables created from the surveys General Social Survey (US) and National Longitudinal Survey (US)*

Note: * a list of the survey questions is found in appendix section B.1.

3 Presentation of My Method and How It Compares to Previous Measures of Norms

This paper assumes the following the data generating process: There exists an unobserved gender norm (or at least it is hard to quantify what it consists of and how strong it is) giving rise to observed gendered language. Gendered language is that people write differently depending on if they depict a male or female. As an illustrative example, a nurse of female sex would be referred to as a *nurse*, whereas a nurse of male sex would be referred to as a *male nurse*. The method used in this paper consists of training a model to detect gendered language. Table 2 illustrates it: The model learns to put in the correct word, *he* or *she*, in tweets where blank spots have replaced those two words. A binary he/she variable, made up of the correct words, is used as a dependent variable to infer gendered language. The independent variables are the sequence of all words in the training dataset vocabulary, including the blank spot.

A person that would guess if the tweets in Table 2 should be filled with *he* or *she* would use previous knowledge of gender prescriptions to make a good guess. The model does the same but uses previous knowledge from all the tweets it has been trained on. More specifically, the model learns the average of how Twitter users choose to depict women and men in separate ways in the training data. It learns to separate because the training setup renders the model to be discriminatory. For instance, it can only use words and phrases describing *he* more than *she* to predict to the he class. I argue this is equivalent to learning a definition of gender prescriptions. If there were no gender norms, the Twitter users would not systematically choose to portray females and males in specific ways, and the model would not learn any words and phrases that help it in predicting *he* and *she*. The stronger the gender norms, the more differential depictions of the genders exists in tweets, and the easier it is for the model to learn words and phrases that predict the genders. Thus, this paper defines the gender norms to be a set of boundaries of different intensity that defines “being a woman” and “being a man”, which follows North’s and Akerlof and Kranton’s definitions (1990; 2000). For example, in the tweets, women are overrepresented as pictured to be involved in personal relations, such as being friends and colleagues, and the model learns that this defines “being a woman”. The interpretation becomes that the social punishment for not being involved in such personal relations is higher for a woman than a man.

Table 2: Examples of Input Tweets

Tweets	Predicted Class
English	
>>... Don't know which reality ___ lives in. *Smiling Face With Open Mouth and Cold Sweat Emoji*	She
<user> I don't like Pippi either. Hate how ___ always buys <female friends/male friends> and all prudence.	She
<user> <user> Swedish nazi brags about that ___ has killed in Ukraine white Europe is a common utopia among SD:s … <url>	He
What was it ___ was called at FIFA? Infantilo ?	He
Swedish	
>>... Vet inte vilken verklighet ___ lever i. *Smiling Face With Open Mouth and Cold Sweat Emoji*	Hon
<user> jag gillar inte heller Pippi. Avskyr hur ___ ständigt köper sina <väninnor/ vänner> och all präktighet.	Hon
<user> <user> Svensk nazist skryter om att ___ har dödat i Ukraina vitt Europa är en utopi vanliga SD:are … <url>	Han
Vad var det ___ hette på FIFA? Infantilo ?	Han

Note: The dependent variable, if *she* or *he* should be placed in the blank spot, is binary. The model uses the sequence of words, the fill-in-the-blank tweet, to predict which class, he or she, it belongs to. In the table, the predicted class is equivalent to the true class. Words are masked both due to protect privacy and to reduce noise, e.g. <user> and <url>. Other words are masked to take away an "obvious" gender dimension. For example, there exists a female and a male version of the word *friends* in Swedish, *female friends* (vänninor) and *male friends* (vänner). The masking is illustrated by <female friend/ male friend> being considered a word instead of the original two. The translation from Swedish is made by the author.

Persons can on a qualitative basis observe the gender norms, but quantifying their changes is more difficult. A model capturing a definition of gender norms can be used to measure change over different samples, both by investigating sequences of words used and the overall predictive accuracy. This paper uses predictive accuracy, i.e. how many he and she tweets the model predicts correct, to measure if a new sample of tweets reflects the already-defined norms. The measure reflects the extent to which third-party persons will experience the past gender norms by reading the new tweets. Another possible interpretation of the measure is that it reflects the average following of the past gender norms by the users posting the new tweets. If a user chooses not to tweet due to a change in the following of the past norms, the missing tweet does contribute to the norm change by its absence, i.e. nobody experiences the tweet. According to North (1990), the humanly devised constraints (the norms) are shaped by human interaction (the posting and the reading of tweets). In sum, the measure indicates a change of norms.

The proposed method will yield a much more encompassing definition of norms than previously employed. It uses all the myriads of aspects that Twitter users choose to write to depict the genders in separate ways to create the definition. With the proposed method, the analyst only selects one group feature to infer the rest whereas with surveys, experiments or research designs, the researcher decides which sub-norm to investigate. I expect the data-driven method to take into account male gender roles to a much larger extent than what has previously been done in research. For example, by using surveys, researchers ask respondents questions about how they feel about women working in the labor market, but usually do not ask questions about males taking a larger responsibility for working in the household (see Appendix Table B.1). Also, the proposed method will yield an objective definition of gender norms. An alternative way of quantifying gendered language would be to let persons mark tweets as gendered. Using such training data would teach the model to capture an average of the marking persons' gender stereotypes. Since persons might hold norms subconsciously, the marking might be biased. The current method not only allows for the use of text data, but it also lets the data decide what the norm consists of, yielding an encompassing and objective definition.

4 Data

I have downloaded 100 million tweets marked with being in Swedish from May 2016 until May 2018 from Twitter's API. Twitter data is published directly by people without an editor. Since people write messages to each other that is more closely related to how they speak in real life, it is more likely that tweets capture gender norms than, for example, books and news articles. The data used do not include retweeted tweets. The main dataset consists of tweets which include the words *he* (han) or *she* (hon). I replace the true word indicating gender (*he*, *she*) with a placeholder and the true word becomes a variable in the dataset. It is ambivalent which gender tweets containing both *he* and *she* is about. Such tweets are not included in the dataset as they only constitute 0.7 percent of tweets containing any of the two words. Twitter users are anonymous for all practical purposes; for example, there is no information on if it is a man or a woman who writes a specific tweet.[3]

Table 3: Sets and sample sizes

Set	Function	Time	Count of tweets		
			Original	10-25 words	No #metoo and related
Training	Estimate parameters	Year 1	984 337	676 466	
Validation	Model selection	Year 1	246 085	169 184	
Test	Evaluate model	Year 1	307 606	211 242	203 691
Evaluation	Evaluate model	Year 2	1 396 705	1 017 414	989 028

From the he/she tweets, several subsets are formed. I create two main sets; Year 1 at 1.5 million tweets spanning 2016-05-13 to 2017-04-30 and Year 2 at 1.4 million tweets spanning 2017-05-01 to 2018-04-30. The Metoo event takes place in the middle of Year 2. Table 3 displays the sets, their functions and their sample sizes. The he/she tweets of Year 1 are further subdivided into a training, validation and test set according to a random 64-16-20 percent split. The subdivision to a training, validation and test set is standard in the machine learning paradigm. The training set is used to estimate parameters. The validation set is used for model selection, i.e. to estimate hyperparameters. The test set is reserved for evaluating the chosen model on unseen data. The tweets of Year 2 is also a test set in the machine learning sense, but as it performs a separate function in the evaluation of the Metoo movement, I make it distinct and call it the evaluation set. Because tweets with very different word counts are difficult to handle in the same model, the sets are subsampled only to contain tweets that include 10-25 words. The removal cuts the sample sizes to about 70 percent of the original. An alternative would be to train various submodels, but as the main goal is to capture gendered language, the current solution is deemed sufficient. An additional reason for the sample cut is that tweets with a word count below 10 contain little information to predict on. The models are trained and validated on 845 000 tweets and tested on 211 000 tweets. The tweets of Year 2, the evaluation set, is used to evaluate the Metoo movement by measuring the change in gender norms vis-a-vis the test set of Year 1. The two sets are subsampled to only include tweets which do not

[3] The name chosen by a Twitter user is given in the data and might indicate the gender of the user. However, since there is no easy way of judging the measurement error by approximating the user's gender from the Twitter name I refrain from using the information. In general, see Steinert-Threlkeld (2018, pp. 83-84) for a discussion on common background variables that does not exist on Twitter users.

have a hashtag associated with the metoo hashtag and by removing missing dates.[4] The final sample size for assessing the effect of the Metoo movement is 1.2 million tweets.

A model will learn that names, such as *Erik* and *Anna*, refer to males and females. A definition of gender norms is not nuanced if the model simply learns such names. Thus, in the main version of the model, all first names are censored. The name list is retrieved from Statistics Sweden (Statistics Sweden, 2016a,b). It contains all first names used by any Swedish residence and the number of female and male name holders.

5 Model Specification

An LSTM neural network model is presented in this section as it outperforms other models, as shown in Appendix A. The appendix presents the model specifications and selection in much greater detail with the terminology used in the neural network literature.

5.1 Overview of machine learning concepts

Machine learning techniques have recently found increasing applications in economics (see Mullainathan and Spiess, 2017, pp. 99-103, for an overview). The methods are characterized by the functional form being learned from the data. Although papers within economics use machine learning methods, papers in the finance literature have mostly adopted neural networks. A neural network is not a model, it is an algorithm in the sense that the analyst sets up limits and a good model is searched for within those.[5] Neural networks are universal function approximators (Hornik, 1991). Theoretically, they can approximate any functional form and therefore the algorithm can find any model. The result is based on the network searching in such broad limits that it is computationally infeasible and, in practice, the analyst creates boundaries within which the network chooses the

[4]Due to downloading issues data are partly missing for 22 dates from Year 1 and 44 dates for Year 2 as displayed in Appendix Table B.1.

[5]However, the convention of referring to neural networks as models seems established and consequently I refer to neural networks as models in this paper.

best fit. In general, networks are used for high-dimensional data and designed to contain many other types of machine learning algorithms.

A common task is to predict Y from the data, $D = [Y, X]$. More precisely, the analyst forms an error function (loss, objective, cost function):

$$E = -P(Y_i, F_c(X_i, A, \Omega, B)) \tag{1}$$

where $P()$ is a performance measure of the model as a function of the dependent variable Y and the candidate model $F_c(X_i, A, \Omega, B)$. The error function (1) behind a neural network can include many different specifications of P and $F_c()$, i.e. networks can be differently designed. I introduce the concept of having a candidate model, $F_c()$, to be able to compare machine learning versus more standard statistical approaches. The reason for calling it a candidate model is that the specification also includes regularization, Ω, various features existing to penalize overfitting (overtraining/ model complexity), and adherent hyperparameters, A, that guide model selection. The candidate model is not akin to the model obtained by training the network. The functional form of the obtained model is learned by splitting the original data into two parts: a validation set used for model selection and a training set used for estimating model parameters. By choosing the A hyperparameters that give the lowest error in eq. (1) on the validation set, a model is selected that is not overfitted, i.e. which has good generalization performance. The model parameters, B, are estimated on the training set for each hyperparameter combination. By estimating hyperparameters on a new sample, the validation set, new error terms are drawn, and data not used to estimate its parameters evaluates the candidate model.[6] In the more standard statistical framework the training-validation set split is absent, and the data analyst proposes a candidate model $F_c()$.[7] Residuals from the same data approximate the error terms. On a

[6] In the above framework, the lasso regression seems to be most commonly used in economics and is therefore exemplified. For the general error function in eq. (1) to be equivalent to a lasso regression the analyst selects the performance measure to be the mean square error, $P = \sum_{n=1}^{i}(Y_i - F_c(X_i, A, \Omega, B))^2$. Also, the analyst decides that the candidate model is $F_c(X_i, A, \Omega, B) = X_i^T b + a \sum_{k=1}^{k} \mid b_k \mid$. Thus, the analyst decides to investigate the set of linear functions (first part) and to regularize the model by shrinking large coefficients with the L1 norm (second part).

[7] The major exception is in time series analysis when the forecast error on out-of-sample data decides model selection, which is equivalent to minimizing the performance measure on the validation dataset.

conceptual level, in a more standard statistics framework, the analysts are more involved in the algorithmic process of finding a good model (especially if an analyst tests a model proposed by another on new data) whereas, in machine learning, an automated algorithm performs more of the task.

Verification of the model is achieved by evaluating the learned model on unseen data, a test set, separate from the training and validation set mentioned previously. Machine learning methods generally and neural networks specifically are "black box" approaches. "The black box" exists since networks are designed to handle high-dimensional problems and consequently it is hard for the analyst to interpret the parameters of the learned model. Neural networks tend to be used to solve predictive tasks; they are congruent with inference on the model level.

5.2 Model specification

The LSTM neural network model specifies as follows:

$$P = \sum_{n=1}^{1} Y_i \log F_c(X_i, A, \Omega, B) + (1 - Y_i) \log(1 - F_c(X_i, A, \Omega, B))$$

where Y_i is a binary indicator on if *he* or *she* is written in the blank spot in tweet i. The performance measure P is the cross-entropy error. P is in itself equivalent to optimizing the log-likelihood function behind a Logit model. The candidate model $F_c(X_i, A, \Omega, B)$ is a large parametric model which will search a large scope of possible functional forms to learn:

$$F_c(X_i, A, \Omega, B_i) = f_o(\sum_{m=1}^{M} b_m^o f_h(h_{m,t=T}) + c^h)$$

$$f_o(z) = \sigma(z)$$

$$f_h(z) = \begin{cases} \text{z} & \text{if z>1} \\ 0 & \text{if z<1} \end{cases}$$

Regularization in the designed network enters by the hyperparameter the number of hidden nodes (M), which decides the number of b^o_m coefficients and $h_{m,t=T}$-processes. The number of hidden nodes generally guides how large a neural network model should be and thus, how well the model can fit the data.[8] The candidate model is a panel-data model as each $h_{m,t=T}$-process is specified by:

$$\begin{aligned} h_{m,t} =& \tanh(c_{m,t})\sigma(x_{t,i}B^{u1}_m + h_{m,t-1}B^{u2}_m + c^u_m) \\ c_{m,t} =& c_{m,t-1}\sigma(x_{t,i}B^{f1}_m + h_{m,t-1}B^{f2}_m + c^f_m) + \tanh(x_{t,i}B^{c1}_m + \\ & h_{m,t-1}B^{c2}_m + c^c_m)\sigma(x_{t,i}B^{p1}_m + h_{m,t-1}B^{p2}_m + c^p_m) \end{aligned}$$

The presentation of the specification is brief, the above equation is a short form of an LSTM node which is explained in Appendix A, subsection (3) LSTM. The main takeaway message of the current presentation is that the candidate model is a large parametric model which will be reduced during training.

The inputs $x_{t,i}$, representing the *t*:th word in the *i*:th tweet, to the network are word vectors. The usage of word vectors is today common within natural language processing. The word vectors are parameters to the network which are already specified, $x_{t,i} = d_{t,i}B^r$. Dummy-variable encoded vectors, $d_{t,i}$ get reduced by the B^r matrix to word vectors $x_{t,i}$. Each word vector is a k-dimensional vector representing one word and each word vector element is a parameter "looked up" in the B^r matrix.

5.3 Word vectors

I choose to train my own word vectors by the Skip-Gram method of Mikolov et al. (2013). Available Swedish word vectors are based on Wikipedia text and as the type of language used differs a lot to Twitter, using my own encoded vectors likely enhances the performance of the model.

[8] Regularization enters in three ways but only one is explicitly shown in the specification. The other two hyperparameters are the following: The second hyperparameter is training time because the error function is numerically optimized by updating the candidate model with a number of randomly selected observations at each round (stochastic gradient descent). The third hyperparameter is the percentage of b^o_m:s present for each individual update, which pertains to a computational efficient model averaging technique (dropout).

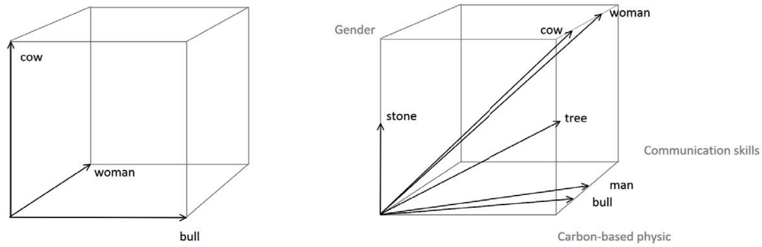


Figure 1: Example of dummy variable representation (left) and word vector representation (right)

An intuitive explanation of the dimensionality reduction performed in the first part of the model, going from dummy-variable encoded vectors to word vectors, follows. Words are usually thought of as dummy variables, but it might not be a good representation. Let us take *cow*, *women*, and *bull* as an example. Formulated as being three dummy variables would mean that they are orthogonal to each other, as visualized to the left in Figure 1. This is clearly not the case as *cow* and *women* are both of female gender and they have some relation that is not captured with a dummy variable encoding. Words can be encoded into many different vector spaces without any inherent truth about whether it provides a good representation or not. One example of such a vector space is the right one in Figure 1, which illustrates the encoding of words to word vectors developed by Mikolov et al. (2013). Such word vectors have good properties, for example, it is possible to perform vector algebra on them and get sensible results, which is illustrated in the figure.[9] When using the pre-trained word vectors, this paper uses a representation, again with no inherent truth of whether or not it is good. If it is a good representation or not is judged by how good the obtained model is at predicting on the test set.

In a related paper, Bolukbasi et al. (2016) classify certain words as gendered in an unwanted way and show that this dimension exists in the word vectors trained on English Google News. They provide an algorithm to remove the unwanted gender dimension in the word vectors for usage in different downstream applications. Bolukbasi et al. (2016) find that one dimension is sufficient to detect unwanted gendered words in the word vector space. If this pattern generalizes to gendered words in general, it suggests that a weighted sum of word vector elements would

[9]Various papers exist evaluating the properties of word vectors. For example, Arora et al. (2016) investigate polysemous words, words where the letters are the same but where the meaning differs and find that they are placed in the center of their respective meanings.

be a good representation for the gendered-words intensity of a sentence. Appendix A shows that performance increases with the non-linear LSTM model, showing that additional predictability over a weighted sum exists for the present case. The main difference between the papers is that this paper detects gendered language in sentences, instead of removing predefined unwanted gendered words from word vectors.

5.4 Masking of words

A model will use words such as *boy - girl* and *Anna - Erik* to predict if a tweet is about a man or woman. One could argue that a good definition of gender norms ought to be more nuanced than capturing such "too obvious" gender-colored language. Thus, this paper also trains the models on masked data where "too obvious" gendered colored words are censored. As an example, in the masked data, the fictive word *boygirl* replaces the original words *boy* and *girl*. Both raw/unmasked and masked versions are trained, but when evaluating the Metoo movement, the masked version is used. The masked version of the main model uses three percent fewer words than the unmasked version.

The words used for masking divides into three groups; pairs of male-female words, family names of famous persons and first names. I have selected words to mask in the following way: Firstly, I selected pairs of male-female words (*boy - girl*) from lists of high predicting words from the Neural Network or Naive Bayes model. I extended the list with the words' respective neighbors according to the pre-trained word vectors (*boy's - girl's, the boy - the girl*). I deemed some high predicting words to be impossible to map, such as *witch* and *tinder dude*, and such words are not masked. Secondly, I selected the family names of famous persons from the same lists. Lastly, I selected all Swedish residents' first names, except of names that also is another word in Swedish (*kalla, du*) or English (*star, honey*), from the name list obtained from Statistic Sweden.[10] For the two latter groups, the fictive word *familyname* and *firstname* replaces all names. See the table "Words used for masking" in the online appendix for the lists.

[10]For additional details please see Appendix A.

6 Model Result

Table 4 evaluates the models on the test set of Year 1. Initial model selection in Appendix A shows that the best-performing model is a one-layer neural network LSTM model. The table presents it alongside a Naive Bayes model, which functions as a baseline. The Naive Bayes model is a well-established model for similar tasks (Hastie et al., 2009, pp. 210-211). The main difference between them is that the neural network model takes into account the sequencing of words, i.e. can learn phrases, whereas the Naive Bayes model does not. The first row of Table 4 presents the chosen performance metric, ROC AUC scores. A ROC AUC score of 0.5 means that the model has no predictive power: the model cannot separate the genders in the text. A score of 1 means that the model has perfect predictive power: the model can completely separate the genders. The first two columns present model results for the Neural Network and Naive Bayes models that are trained and evaluated on the raw/ unmasked data. Those two models use words such as *mum* and *dad* to capture gendered language. Since this might not be in line with the type of gendered language one wants to capture, the last two columns of the table present the masked versions of the models. In the masked data, fictive words such as *mumdad* replace gendered words such as *mum* and *dad*. Inspection of the ROC AUC scores in Table 4 shows that the Neural Network model always is better at predicting than the Naive Bayes model. The Neural Network model is an overarching model type – the Naive Bayes model is a submodel of it. The differences between them are small, 0.039 and 0.026 ROC AUC score points for the unmasked and masked version, respectively. The Neural Network model allows for the use of sequential information, namely, that one word comes before another, and the result indicates this is not very useful in predicting the he/she variable. The Naive Bayes model only uses the relative frequency of words in each class to predict. The result shows that genders to a large extent can be discerned in a text by the words they are bundled with. The table also shows, as expected, that the scores decrease in the masked versions of the models. The masked version of the Neural Network model has a ROC AUC score of 0.76, showing that the genders indeed still can be separated.

Table 4: Model Evaluation on Test Set

	Unmasked		Masked	
	Neural Network	Naive Bayes	Neural Network	Naive Bayes
ROC AUC	0.8496	0.8107	0.7629	0.7372
Unbalanced Sample				
Accuracy	0.8242	0.8058	0.7748	0.7522
Sensitivity	0.4020	0.4015	0.2133	0.3708
Specificity	0.9656	0.9412	0.9629	0.8799
Balanced Sample				
Accuracy	0.7529	0.7224	0.6850	0.6619
Sensitivity	0.7976	0.7308	0.7459	0.6777
Specificity	0.7091	0.7142	0.6255	0.6464

Note: ROC AUC: a threshold-independent performance measure ranging from 0.5 (model classify as good as random) to 1 (model perfectly classify). Accuracy: correctly predicted tweets over total tweets. Sensitivity: correctly predicted she tweets over total she tweets. Specificity: correctly predicted he tweets over total he tweets.

Table 4 further illustrates the main result by measuring accuracy, sensitivity, and specificity. Accuracy is how many tweets the model correctly classifies. Sensitivity and specificity investigate how good the model is in either class. Sensitivity is how many she tweets the model classifies correct over total she tweets. Specificity is how many he tweets the model classifies correct over total he tweets. The unbalanced sample panel presents that the masked version of the Neural Network model has an accuracy of 77 percent as opposed to the Naive Bayes model that has an accuracy of 75 percent. Also, it shows that the sensitivity of the Neural Network model is 21 percent and the specificity is 96 percent. In other words, of the total she tweets, it classifies 21 percent correctly to the she class and, of the total he tweets, it classifies 96 percent correctly to the he class. The fact that the masked neural network model mainly uses the simple rule of classifying to the he class is due to having imbalanced classes while deciding to optimize accuracy. There are around three times more he tweets than she tweets and by optimizing accuracy, the model takes into account that the best guess for any tweet is that it belongs to the he class. The class imbalance is large in magnitude – out of 100 tweets, 74.6 contains *he* and 25.4 contains *she*. The class imbalance does not seem to be specific to Swedish tweets. The *she* class ranges from 23 percent on Wikipedia to 43 percent on popular Swedish blogs.[11] Similar levels of class imbalance exist in the large corpus of scanned English books available at Google's Ngram Viewer (Google's Ngram Viewer). The class imbalance is not a reflection of gender norms according to my definition of it as a set of soft boundaries defining a female and male identity. I interpret the class imbalance as males being talked about to a larger extent than females are. I choose to optimize accuracy because I want to capture the gender norms through a continuous gendered language variable while taking into account that the bulk of tweets is about men. The class imbalance is by itself a result which should not be understated. Later on in the paper, the masked version of the Neural Network model, for which accuracy is optimized, evaluates the Metoo event.

The choice of optimizing accuracy withstanding, it is not a convenient standpoint for illustrating the performance of the model. Table 4, the balanced sample panel, illustrates that the model captures the gendered language, and does not simply predict all tweets to belong to the he class. The balanced sample panel shows results on a balanced sample for the choice of maximizing the predictive power of the model in either class. The balanced sample is created by randomly throwing away excess he tweets from the test set until an equal count of he and she tweets exist. The model is trained with a binary variable representing *he* and *she*, but I consider gendered language be a continuous variable ranging from 0 (male) to 1 (female). The training against the he/she variable is a method to capture the underlying unobserved continuous gendered language variable. The output of the model is a probability distribution since the last stage of the model is specified as a logistic function (just as for a Logit model). To classify into either class, a specific threshold value is specified at the probability distribution. The ROC AUC score is the preferred performance metric since it is independent of any threshold chosen. In the preceding section was the threshold chosen to optimize accuracy, but the threshold can be chosen in many different ways. The balance sample panel presents results from a threshold optimized for making the model being equally good at predicting to the two class. Table 4 presents that the masked version of the Neural Network model now has an accuracy of 68 percent. At the same time, sensitivity is at 75 percent and specificity at 62 percent. In other words, when

[11]For examples of Swedish text sources and the class imbalance found in them see Appendix Table B.2. The class imbalance in the Swedish tweets is not due to females being mentioned in a passive sense; inclusion of the words *him* (honom) and *her* (henne) in the two classes only increase the *she* class by 0.7 percentage points. It is not the case that most Twitter users are males that talk about other males as the likelihood of a random Twitter user being male is 56 - 58 percent (ISS, 2016).

making a fictive reality where tweets are as much about women as men, the model predicts she tweets correctly 75 percent of the time.

Table 5 illustrates which words that the masked Neural Network model uses to predict. The model uses sentences as input and not words in isolation, which is hard to visualize in a table. To easier illustrate what the model does, the median predicted probability ranks the words, termed Word Color (WC) by me since each word "gets colored" by the prediction for the tweets to which it belongs. The higher the WC, the more often the word is used in a tweet classified to the she class. The method is conceptually equivalent to evaluating a Logit model on the sample mean, except that I strive to preserve the ordering of words taken into account by the model by using the median of the predicted probability.[12] The table displays the most frequent words over the predicted probability distribution. For example, the masked word *grandfathergandmother (morfarmormor)* has a word color of 0.77, meaning that the word is found in tweets predicted to belong to the she class. Table 5 displays that words related to sport are found in tweets predicted to be about a male. Words related to personal relations are found in tweets predicted to be about a female, even though the words are masked.[13] [14]

[12]Another way of understanding what a neural network model does is to train a generative network or include an attention mechanism in the network. However, a generative network will still yield chosen examples of what it found, for example, the "mean" sentence of a tweet containing the word *he*. An attention mechanism would only illustrate how the network predicts individual tweets. Thus, at present, I refrain from making such extensions.

[13]The outmost left tail in Table 5, where we find *credit*, *free* and *casino*, is due to very few words having a median predicted probability (WC) close to one. The table also displays the word *the lady* (*tanten*). One could argue that this word should have been masked just as *the ladies* (*tanter*), *lady* (*tant*) and *neighborhood lady* (*granntanter*) are (seen in the table of words used for masking in the online appendix). That only one word on the list in Table 5 denotes a gendered subject shows that most relevant words indeed are masked.

[14]One might want to control which words are relatively more used in either class by evaluating on the balanced sample, as opposed to absolutely used in the unbalanced sample. Appendix Table B.3 replicates the same exercise on a balanced sample and shows that the pattern of sports versus personal relations still holds.

Table 5: Frequent words over the predicted probability distribution for the masked model

WC	English	Swedish
0.01	contract	kontrakt
0.02	united	united
0.03	games	matcher
0.03	player	spelare
0.04	club	klubb
0.05	the season	säsongen
0.06	goal	mål
0.07	the game	matchen
0.08	game	match
0.09	play	spela
0.10	the ball	bollen
0.11	team	lag
0.13	played	spelade
0.14	football	fotboll
0.15	penalty	straff
0.16	field	plan
0.18	ready	klar
0.18	play	spelar
0.18	miss	missar
0.20	em	em
0.21	europe	europa
0.23	the chance	chansen
0.23	worse	sämre
0.25	president	president
0.25	<familyname>	<familyname>
0.27	short	kort
0.27	leave	lämnar
0.27	last	förra
0.28	score	poäng
0.30	bad	dålig
0.32	smallest	minst
0.32	latest	senaste
0.34	manage	lyckas
0.34	still there	kvar
0.35	before	före
0.35	against	mot
0.37	were	varit
0.39	good	bra
0.40	fuck	fan
0.40	in	in
0.43	like	ju

Continued on next page

Table 5 – *Continued from previous page*

WC	English	Swedish
0.43	come	kommer
0.44	in	i
0.44	had	hade
0.44	.	.
0.45	<user>	<user>
0.46	,	,
0.47	___	___
0.48	are	är
0.48	at	att
0.50	<hashtag>	<hashtag>
0.51	and	och
0.52		
0.53	I	jag
0.54	<url>	<url>
0.55	you	dig
0.57	love	älskar
0.58	me	mig
0.59	my	mitt
0.59	<boygirl>	<killetjejkillle>
0.61	my	mina
0.62	*heart eyes*	*heart eyes*
0.63	children	barn
0.63		
0.64	*heart*	*heart*
0.66	<womenmen>	<kvinnormän>
0.66	<thedaughtertheson>	<dotternsonen>
0.66	my	min
0.67	<brothersister>	<brorsyster>
0.69	<sondaugther>	<dotterson>
0.71	friend	<väninnavän>
0.72	<mum'sdad's>	<mammaspappas>
0.72	<grandfathergrandmother>	<farfarfarmor>
0.74	beautiful	vacker
0.75	<mumdad>	<mammapappa>
0.76	fi	fi
0.77	<thewomenthemen>	<kvinnornamännen>
0.77	<grandfathergrandmother>	<morfarmormor>
0.78	pink	rosa
0.80	yearly	åriga
0.81	makeup	smink
0.81	nails	naglar
0.82	hagen	hagen
0.82	the cookie	kakan

Continued on next page

Table 5 – *Continued from previous page*

WC	English	Swedish
0.82	pippi	pippi
0.86	noora	noora
0.87	book	book
0.87	thatcher	thatcher
0.88	raped	våldtagen
0.90	zara	zara
0.90	dress	klänning
0.91	pregnant	gravid
0.92	veil	slöja
0.94	the lady	tanten
0.94	<hishers>	<hanshennes>
1.00	lift	lyft
1.00	credit	credit
1.00	free	free
1.00	casino	casino
1.00	code	code

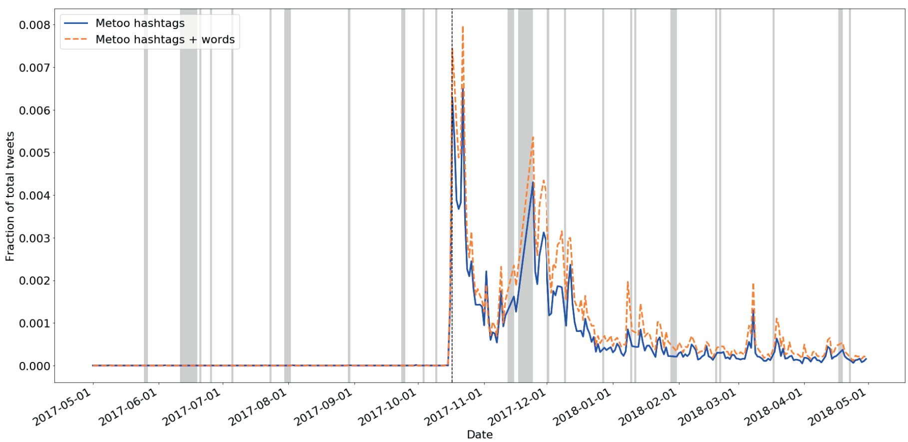
Note: The most “male” word with a WC of close to zero is at the top of the table and the most “female” word with a WC close to one is at the bottom of the table. The table is generated in the following way: The median predicted probability for each word is calculated from the predicted probability of each tweet, named Word Color (WC). The WC of all words is binned into 20 groups and for each quantile is the 5 most frequent words displayed. The table is generated on the test set and evaluated by the masked neural network model. The translation from Swedish is made by the author.

7 The Metoo Movement in Sweden

The Metoo movement started in the US and quickly spread to Sweden. On the morning of the 17th October 2017, people started to share and read the metoo post on social media:

> #MeToo.
>
> If all the women who have been sexually harassed or assaulted wrote ”Me too.” as a status, we might give people a sense of the magnitude of the problem.

During the following hours, people shared more and more metoo posts. Figure 2 displays the Metoo movement on Swedish Twitter. On the 17th of October 2017, 0.63 percent of all Swedish non-retweeted tweets contains the hashtag.[15] The hashtag peaks the 22nd of October at 0.65 percent of all tweets. As displayed in the figure, there is virtually no appearances of the hashtag before the 16th of October 2017. Specifically, there is a total of 10 appearances of the hashtag for the five months before the Metoo event. This paper argues that the Metoo event's exact date of occurrence, the 17th of October, is exogenous in regards to the following of gender norms. The Metoo movement continued in the subsequent weeks after the initial event in Sweden with women in different sectors of the Swedish economy collecting their experiences at the workplace under different hashtags.[16] Figure 2 also shows a leveling out. For the last included month, April 2018, on average, 0.028 percent of all tweets contain the hashtag.



Note: The "metoo hashtag" series includes appearances of the hashtag. The "metoo hashtag + word" series additionally includes appearances of the word *metoo*, not prefixed with #. The shaded regions represent missing dates.

Figure 2: The #Metoo event

[15] I choose to present the Metoo movement with total tweets in the denominator as the event in itself is correlated with a rise in the share of hashtagged tweets, which might be due to the event itself. Those considerations withstanding, the previously quoted figure for the 17th of October is equivalent to 4 percent of all hashtagged tweets containing the metoo hashtag.

[16] To give examples, #tekniskfel, #tystnadtagning and #medvilkenrätt (#technicalerror, #silencetake, #bywhichright).

The Metoo event and the following movement raised general awareness of the sexualization of women. This might lead to people changing their gender norms regarding women and, thus, how they express themselves in a text. The discussions following the event spilled over to gender issues in general. Such debates might lead to persons wishing to categorize people as men and women to a lesser extent. The persons participating in the movement most likely hoped for a less gender-stereotypical environment, but it is possible that it became more gender-stereotypical as well.

There are likely differences in how persons of different ages experienced the Metoo event. The younger generation is more prone to consume social media. Younger persons are more likely impacted by having friends and acquaintances sharing experiences of sexual harassment online. In comparison, the older generation to a greater extent consumes traditional media, such as TV news broadcasts and newspapers. Without a presence on social media, the impact of the Metoo event is more likely the notion of media personalities being accused of sexual harassment, which might not yield an equally strong experience.

This paper captures the norms and the effect on Swedish Twitter users. According to a survey performed by ISS (2016), 18 percent of Internet users use Twitter and 8 percent post tweets in Sweden. The age distribution is skewed: the highest share of Twitter users is among 16 to 25-year-olds at 35 percent and declines with age. Men use Twitter more than females; the likelihood of a random Twitter user being male is 56 to 58 percent. Hence, the gender norms the younger generation holds are defined and investigated, with a slight bias towards men.

8 Research Design

The main dependent variable *Follow Norms* evaluates if gender norms changes in relation to the Metoo movement. The exact model version used for the evaluation is the masked version of the neural network model where accuracy has been optimized, henceforth called Gini for short. *Follow Norms* is a binary indicator of whether Geni predicts a tweet i correct:

$$FollowNorms_i = \begin{cases} 1 & \text{if } Prediction_i = TrueGender_i \\ 0 & \text{if } Prediction_i \neq TrueGender_i \end{cases}$$

An average of *Follow Norms* for a sample of tweets is equivalent to the predictive accuracy of Geni. The variable *Follow Norms* indicates if the norms in Year 1 are followed because Geni is trained on data from this year. In other words, norm changes are benchmarked against a definition of the gender norms obtained in Year 1.

The Metoo event occurred on the 17th of October 2017. To be sure not to capture the effect of the Metoo movement leading to a more intensive debate about gender, tweets containing the metoo hashtag and other related hashtags are removed.[17] Although the event's exact date of occurrence is exogenous to the Swedish society, identification relies on having a counterfactual scenario. Figure 3 illustrates the baseline research design. There might exist time trends in norm following. The tweets of Year 2 (where the Metoo event takes place) allow us to investigate possible time trends five months before the event. There might also exist seasonal variation in the following of gender norms. We can use the tweets of Year 1 (the previous year) as a comparison group to approximate such seasonal variation by calendar day fixed effects. Only the test set of Year 1 functions as a comparison group because Geni is not trained on those tweets. They better approximate the predictive accuracy of Geni, the level of *Follow Norms*, which one expects in Year 2 in the absence of any effect.

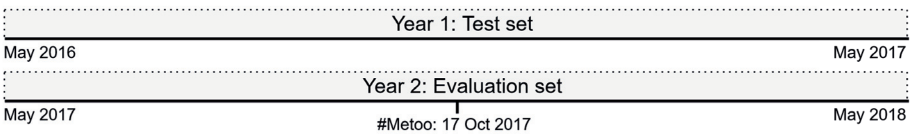


Figure 3: Research design

The baseline specification becomes:

$$FollowNorms_{itu} = b_1 * Year2_t + b_2 * AfterMetoo_t + v_t + u_{itu} \quad (3)$$

where each observation is a tweet i posted at calendar day t by user u. The variable *Year2* is a binary indicator on if the tweet belongs to Year 2. The main independent

[17] This cuts the sample size to 97-98 percent of the original.

variable *AfterMetoo* is a binary indicator of whether the tweet is posted in Year 2 after the Metoo event. The tweets of Year 2 are compared to the tweets from the test set of Year 1 with calendar days time fixed effects v_t to control for seasonal variation. Time is indexed to start at 2017-05-01 for both years, and thus each calendar day has its own intercept (for example, 2016-05-01 and 2017-05-01 have one intercept).

The set-up used in this paper is a difference-in-differences setup, but compared to many other cases when this research design is used, many timepoints exist (large t). It is also similar to regression discontinuity in time (RDIT). As norms are supposed to be slowly changing, the approximately 5-6 month before and after the Metoo event can be seen as the window size used for the RDIT estimate. There is an ongoing discussion on how to cluster standard errors for similar settings as the current one (Bertrand et al., 2004; Abadie et al., 2017). The approach taken in this paper is to present conservative estimates, although they might be unnecessarily harsh. For specifications including calendar day fixed effects or time trends, such as the baseline, standard errors are clustered on calendar days since it provides a framework which easily extends to user fixed effects.

In robustness checks, autocorrelation in the residuals is taken into account as one could argue for calendar day clustered standard errors being problematic since they miss such variation. The data is aggregated per calendar day, $\bar{Y}_{tg} = \sum_{i \in group\ g} Y_{itug}/I$ where g refers to the group (Year 1 or 2). With the transformed data, the baseline specification (3) becomes $(\bar{Y}_{t,g=Year2} - \bar{Y}_{t,g=Year1}) = c + AfterMetoo_t + (u_{t,g=Year2} - u_{t,g=Year1})$ and autocorrelation is estimated by Newey-West standard errors with a lag length of 4 (a bandwidth of 5). Since the data is aggregated, weights – equal to the sum of observations for the two groups – are applied during estimation.

In additional robustness specifications, I extend the baseline specification (3) with user fixed effects. This is feasible since each tweet contains a user id. Standard errors are clustered on calendar days and users with the correction to the variance-covariance matrix estimator presented in Cameron et al. (2011). I choose the baseline specification without including user fixed effects since I want to investigate the change in tweets. If users stop posting tweets due to the event, third-party persons are less affected by the norms no longer followed, and I would like to capture such composition effects. There is no *a priori* reason to expect users to leave Twitter for another online forum, but the user fixed effects control for alike possibilities.

In the robustness analysis, a second control strategy is also employed. Theoretically, Geni should perform worse over time only since users introduce new words and phrases for describing the same thing as the words and phrases she already knows. At first-hand, this is not a great worry because Geni's accuracy on average increases in Year 2 as compared to the test set of Year 1 (not shown). Another model has been trained to predict the words *I* and *we* in the equivalent way that Geni was trained to predict *he* and *she*.[18] The I/We series approximates a togetherness norm, as it captures words and phrases depicting what people do alone or together. The Togetherness series function as a placebo test – it is not expected to change due to the Metoo movement. The series mainly control for the evolution of the Swedish language. In the analysis, a decrease in Geni's accuracy, the *Follow Norms* variable, is interpreted as norms being less followed but it might just arise since the language evolves. The Togetherness series will respond in the same way and is therefore used to filter out possible language evolution effects.

9 Results

Table 6 presents the main result on whether the Metoo event changes gender norms on Swedish Twitter. The first column presents a raw estimate of -0.003 showing that gender norms on average are less followed in tweets six months after the Metoo event compared to five months before. The magnitude is small, attributing the decrease of 0.3 percentage points in *Follow Norms* to the Metoo movement; the estimate indicates the following of gender norms decreasing by 0.4 percent over the sample mean of the five months before. The implicit counterfactual in the comparison is the average level of *Follow Norms* in the five months before. Although the *Follow Norms*- series does not exhibit pre-trends (Appendix Table B.4), the estimate might be biased. There might be seasonal variation in gender norms adherence, which would have changed norm following even in the absence of the Metoo movement.

The second column in Table 6 displays the baseline specification. The test set of Year 1 functions as a comparison group. Seasonal variation is filtered out by allowing the same calendar day to have its own effect and only the variation around the calendar day means contributes to the estimate of *After Metoo*. The coefficient of -0.015 shows gender norms on average being less reflected in tweets after the event.

[18]See Appendix A for details.

Table 6: Metoo results without user fixed effects

	Raw	Baseline	Cutoff independence	Female Focus
	(1)	(2)	(3)	(4)
Dep Var	Follow Norms	Follow Norms	Gendered Language	She
After Metoo	-0.003***	-0.015***	0.013***	0.022***
	(0.001)	(0.005)	(0.004)	(0.007)
Time FE	No	Yes	Yes	Yes
User FE	No	No	No	No
SE:s	Robust	Clustered on day	Clustered on day	Clustered on day
Tweets	989 028	1 192 719	1 192 719	1 192 719
Days (T)	319	365	365	365
Users (N)	95 025	114 040	114 040	114 040
Data	Year2	Year2/1	Year2/1	Year2/1

Note: *** p<0.01, ** p<0.05, * p<0.1.

The magnitude of the estimate is fairly large, 1.9 percent of the sample mean before the event. Extrapolating the difference implies that tweets would not reflect any of the gender norms followed before the event after 18 Metoo movements (based on counting a level of *Follow Norms* of 50 percent as indicating no adherence to past gender norms, since this is the level achieved when Geni has lost all predictive power). Geni does not make any discrepancy between wanted and unwanted gender norms, and the comparison is not normative.

Like any dependent variable, *Follow Norms* include measurement errors. For the present case, additional errors are introduced in the *Follow Norms*- variable because Geni binarizes the probability for a tweet to be about a male or a female (a continuous variable from 0 to 1) by having a specific cutoff for predicting either gender. The specific cutoff yields the predicted value needed to create *Follow Norms*. To verify the main result, Table 6 column (3) presents estimates with the probability as a dependent variable, which includes fewer measurement errors. The variable does not carry the same interpretation as Follow Norms: it measures the degree of genderedness in a tweet and I, therefore, call it *Gendered Language*. The estimate of 0.013 shows that the language in the tweets after the Metoo event is congruent with it being about a female to a higher degree. The magnitude is fairly large, 3.0 percent of the sample mean. If Geni went 45 years into a future, she would think that all tweets are about women without any hesitation. A direction towards "more female" gendered language does not mean that the tweets are about females, as it can equally well describe men.

The *Follow Norms* variable implements a cutoff making Geni good at predicting independent of the choice of writing either he or she since it is not included in the operationalized definition of the gender norms. If there were many more appearances of she tweets after Metoo, the model would still predict those tweets to belong to the she class, conditional on there being the same level of gendered language in the additional she tweets, and the *Follow Norms* variable would indicate adherence to gender norms. One could argue that it would be beneficial for the model to consider how many tweets mention males and females. Instead of including such aspects in the definition of gender norms, I choose to present results on the count of she tweets as a way to measure which gender the tweet is about. Table 6, column (4) shows that there are 2.2 percentage points more she-tweets after the Metoo event than before. This translates to an effect size of 9.2 percent of the sample mean. Before Metoo, 76.3 percent of all tweets are he tweets, and the estimate indicates that it would take 35 Metoo movements to replace all he tweets with she tweets on Twitter. The large volume of talk about men is decreasing.

Table 7 shows the baseline specification for the subsets of she and he tweets separately. Column (1) and (2) show that for both she and he tweets, gendered language increases. Both women and men are more described as women were the previous year. Column (3) shows that for she tweets this, on average, entails following the norms found in Year 1 to a greater extent. Column (4) presents that for the he tweets, the increase in gendered language, on average, is congruent with following the norms found in Year 1 to a smaller extent. I hypothesized that the Metoo movement leads to a norm shift away from stereotyping the genders, implying that women should be described more as men were the previous year, with a resulting decrease in the following of past gender norms for she tweets. Instead, the overall gender norm shifts towards describing both genders more as females were the previous year. After some thought, the finding is quite reasonable – *a priori* there is no reason why a norm shift implies convergence between the genders. The finding is possible because the operationalized definition of gender norms does not include which gender the tweet, in reality, is about. The current table points to males being described less according to past gender stereotypes as contributing to the baseline effect. The estimate of -0.005 only indicates an effect size of 0.5

percent of the sample mean before the event. Even though he tweets contributes more to the average baseline effect since they occur around three times more often, the size of the baseline effect of -0.015 in Table 6, column (2), cannot only be explained by he tweets being less gender stereotypical. The current table also points to she tweets driving the baseline effect, but in the "wrong" direction. The baseline effect might arise since there are more she tweets for which norms are less followed due to the event, but females on average are pictured more according to past gender norms after Metoo.

Table 7: Metoo results estimated separately for tweets about females and males

	She	He	She	He
	(1)	(2)	(3)	(4)
Dep Var	Gendered Language	Gendered Language	Follow Norms	Follow Norms
After Metoo	0.013***	0.007*	0.020***	-0.005***
	(0.003)	(0.004)	(0.005)	(0.001)
Time FE	Yes	Yes	Yes	Yes
User FE	No	No	No	No
SE:s	Clustered on day	Clustered on day	Clustered on day	Clustered on day
Tweets	293 869	898 850	293 869	898 850
Days (T)	365	365	365	365
Data	Year2/1	Year2/1	Year2/1	Year2/1

Note: *** p<0.01, ** p<0.05, * p<0.1.

Compared to previous authors, I measure gender norms in a much more comprehensive way and it might be that some specific dimension of the gender norms has changed, for example, the tweets might sexualize women less. To investigate if any such patterns exist, Table 8 presents the words which experienced the largest change in accuracy, i.e. in *Follow Norms*. The top of the table presents the top of the differences; Geni predicts those specific words more correctly in Year 1 and more incorrectly after the Metoo event in Year 2. The method employed is for illustrative purposes only, no formal tests are conducted on the words. A word is dependent on other words surrounding it; thus, the main method employed in this paper is to model tweets and conduct inference on them. In Table 8, the "Diff" column displays the change in accuracy by calculating the mean predictive accuracy for each word. Due to this reduction, and since Geni might not use a specific word to a high degree, the table is noisy. I can infer what the model uses to predict on and analyze the change by analyzing the underlying tweets.[19] For example, the word *cutie* (sötis) marks a gender stereotype insofar that males are called cutie (with a slightly derogatory tone) when they do specific things, and in Year 2 females are called *cutie* as well. The *Beating Heart- emoji* is used in combination with other words, such as being nice or able, by males and females mostly for describing their partners. In Year 2, the pattern changes: the *Beating Heart- emoji* is used in combination with qualities that the model does not correctly predict, such as males being *very good* (jätteduktig). Relating to the word *those opinions* (åsikterna), *to have opinions* (ha åsikter) is encoded as indicating a male gender stereotype, as it is mostly males that have opinions in Year 1, but in Year 2 women are described as having opinions to a higher degree.[20] The above are selected examples but indicate that there is not one clear-cut dimension of the gender norms that change.

[19]A longer version of the table is available in the online appendix (in Swedish). As far as I understand, I cannot redistribute the tweets according to the terms and services for the Twitter API.

[20]The words *highly pregnant* (höggravid) and *pregnant* (gravid) unsurprisingly describe females to a high degree. In Year 2, the two words are used as an explanation for why a woman experiences an action performed by a man, a pattern the model has not learned. Insofar that pregnant indicates a gender stereotype by just being a state of the world without having any explanatory power, the pattern still indicates a change.

Table 8: Change in accuracy on word- level

Diff	English	Swedish
0.78	highly pregnant	höggravid
0.73	inspiration source	inspirationskälla
0.71	coco	coco
0.68	prisoner	fånge
0.63	cutie	sötis
0.63	main	main
0.63	pregnant	gravida
0.63	handlebars	styret
0.61	N	N
0.60	the universe's	universums
0.60	nuv	nuv
0.58	scared to death	skräckslagen
0.57	paranoid	paranoid
0.57	=(	=(
0.57	put forward	framställa
0.55	cute	<gulligegulliga>
0.55	fringe	lugg
0.55	thai	thai
0.54	remix	remix
0.53	k-g	k-g
0.53	horse	horse
0.52	jemen	jemen
0.50	kreml	kreml
0.48	relate	relatera
0.48	give joy	glädjer
0.47	*Beamed Eighth Notes- emoji*	*Beamed Eighth Notes- emoji*
0.46	quick	hastigt
0.46	necessary	nödvändig
0.45	draw	tecknar
0.45	tattooed	tatuerade
0.45	disagreeing	osams
0.45	misunderstanding	missförstånd
0.44	copy	kopierar
0.43	bodyguard	livvakt
0.43	nato	nato
0.43	*Beating Heart- emoji*	*Beating Heart- emoji*
0.42	mean	taskiga
0.42	svts	svts
0.42	brainwashed	hjärntvättad
0.42	spokesperson	talesperson
0.42	06	06

Continued on next page

Table 8 – *Continued from previous page*

Diff	English	Swedish
0.41	pre-band/dressing of wound	förband
0.37	loaded	laddat
0.37	foe	ovän
0.37	those opinions	åsikterna
0.37	cheat	fuskar
0.36	rinkeby	rinkeby
0.36	melin	melin
0.36	babe	babe
0.36	the dinner	middagen
0.35	claesson	claesson
0.35	finance minister	finansminister
0.34	<brother/sister-in-law>	<svågersvägerska>
0.33	sexist	sexist
0.31	wrestling	brottas

Note: The top of the table shows words included in tweets that were correctly predicted before the metoo-event and became incorrectly predicted afterward. The table is generated in the following way; the data included is five months of the test set, from 2016-10-18 and forward, and five months from the evaluation set, from 2017-10-18 and forward. For each of the periods the mean accuracy of each word is calculated. The difference, named "Diff" in the table, is formed by subtracting the evaluation set period's mean accuracy from the test set's mean accuracy. The difference is discretized to 20 groups of 0.005 step intervals and the 10 most frequent words from the test set are selected for each group, as long as the word is included in at least 5 tweets for both periods. Due to space limitations, the table displays only the top of the difference distribution. The full table and underlying tweets in Swedish are included in the online appendix. The translation from Swedish is made by the author.

9.1 Robustness checks

Depending on what one considers the unit of observation to be, one might wish to have different standard errors. So far we have investigated how tweets changes, but one could also wish to consider how the level of norm following per day changes over time. Therefore, Appendix Table B.5 and B.6 replicate the preceding two tables, namely, 6 and 7, but display estimates for daily aggregated data, where autocorrelation in the residuals are taken into account by Newey-West standard errors. Now, the baseline estimate is only significant at the 10 percent level. The estimate in Appendix Table B.5 is smaller, but the effect size is the same as in the original baseline specification. In general, the standard error increases somewhat but the size of the original standard errors is similar to the size of the Newey-West standard errors, and but all except two coefficients retain their significance.[21]

Table 9: Metoo results with user fixed effects

	Raw: Sample Cut	Baseline: Sample Cut	User FE	Baseline: User FE
	(1)	(2)	(3)	(4)
Dep Var	Follow Norms	Follow Norms	Follow Norms	Follow Norms
After Metoo	-0.003***	-0.016***	-0.005***	-0.012**
	(0.001)	(0.005)	(0.001)	(0.005)
Time FE	No	Yes	No	Yes
User FE	No	No	Yes	Yes
SE:s	Robust	Clustered on day	Clustered on user	Clustered on day & user
Tweets	849 518	1 053 209	849 518	1 053 209
Days (T)	319	365	319	365
Users (N)	35 338	65 165	35 338	65 165
...in year 2	35 338	35 338	35 338	35 338
...in year 1		51 352		51 352
Data	Year2	Year2/1	Year2	Year2/1

Note: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

The selection process behind generating the he/she-sample might be affected by the Metoo event in itself. To address this issue, Table 9 replicates the specifications but with user fixed included.[22] For the Year 2 data, 35 000 users who have tweeted at least once on either side of the Metoo date, are included, which cut the sample of users to 37 percent of the original. An included user typically tweets 8 tweets in Year 2, but some users tweet much more than others, rendering the mean much higher than the presented median of 8. For the baseline specification, the 51 000 users of Year 1 are added regardless of them tweeting at either side of the Metoo

[21]One exception is the raw estimate, which only was included in the original table for illustrating the raw data. The other exception is *Gendered Language* for the subset of he tweets. The p-value for the latter is 10.2 percent. The coefficient and the standard errors are equivalent in the following two tables: Table 7 column (2) and Appendix Table B.6 column (2). Thus, regardless of the estimation method, the coefficient is close to the 10 percent significance threshold.

[22]At first, the worry seems unnecessary. I can calculate the day of the first and the last tweet for any user: the entry and the exit rate. On the full sample of tweets, neither the entry nor the exit rate exhibit any immediate apparent trend break around the event (see Appendix Figure B.2).

event since the tweets of Year 1 are included to allow for time fixed effects. Cutting the sample runs the risk of making it less representative of all users and tweets, but column (1) and (2) replicate the raw and baseline estimate, respectively, on the cut sample and find them to be very similar to their respective counterparts estimated on the full samples. Column (3) extends the raw estimate by including user fixed effects. Each user is allowed its own intercept and only the variation around a user's mean of *Follow Norms* contribute to the estimate of *After Metoo*. The estimate of -0.005 is lower than the baseline because it shows the average over the post-time period, whereas a U-shaped pattern exists: directly after the initial event users tweet more norm-following tweets; right before Christmas, they start to tweet significantly less norm-following; in the middle of March, they are back at the pre-event norm-following level (see Appendix Table B.4, column 3). Column (4) extends the baseline estimate with user fixed effects. The U-shaped pattern disappears when the seasonal variation is taken into account (see Appendix Table B.4, column 4), rendering the average estimate representative. The average individual user decreases his/her level of following norms in the post-period, as compared to how users change their level in Year 1. The estimate of -0.012 is, again, similar to the baseline specification and quite sizable at 1.6 percent of the sample mean before the event. The estimate implies that it would take 22 Metoo movements to change users' tweets not to reflect any gender norms found in Year 1. A comparison of the baseline estimate with and without user fixed effect concludes that most of the average effect stems from users changing their behavior.[23]

[23]The sample size of the Year 1 tweets is 20 percent of the original because I have subsampled it to allow for model training. Thus, the calendar day means will be weighted by 80 percent towards the daily means of Year 2, leading to attenuation bias in estimates. In other words, possible trends do not contribute towards the estimate of *After Metoo* to the same degree it would have done if the sample sizes were equal. Weighting the specifications to take into account the inclusion probability of 20 percent for the tweets of Year 1 does not change any of the estimates (not shown).

Table 10: Metoo results comparing to Togetherness

	Togetherness (Placebo)	Time Trends	Time FE	Time & User FE
	(1)	(2)	(3)	(4)
Dep Var	Follow Norms *100	Follow Norms *100	Follow Norms *100	Follow Norms *100
After Metoo	1.531	11.449**	10.659*	19.153***
	(1.237)	(5.646)	(5.763)	(5.069)
After Metoo*t	-0.010	-0.107**	-0.103**	-0.150***
	(0.010)	(0.048)	(0.049)	(0.039)
After Metoo*t^2	0.000	0.000**	0.000*	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Time FE	No	No	Yes	Yes
User FE	No	No	No	Yes
SEs	Clustered on day	Clustered on day	Clustered on day	Clustered on day & user
Tweets	1 014 356	2 003 384	2 003 384	1 706 588
Days (T)	319	319	319	319
Users (N)	116 376	211 401	211 401	77 863
Data	Togetherness	Year2/ Togetherness	Year2/ Togetherness	Year2/ Togetherness

Note: The coefficients in the table can directly be interpreted as percentage points since the binary dependent variable has been multiplied by one hundred. The specifications interact various indicator variables with time trends. No indicator variables interacted with time is significant more than for *After Metoo*, and hence they are omitted from the table. Specification (1): $Y_{it} = 1(b_0 + b_1 t + b_2 t^2) + AfterMetoo(b_3 + b_4 t + b_5 t^2) + u_{it}$. Only for this placebo regression is After Metoo coded to take the value one also for the Togetherness series. Specification (2): $Y_{it} = 1(b_0 + b_1 t + b_2 t^2) + AfterMetoo(b_3 + b_4 t + b_5 t^2) + HeShe(b_6 + b_7 t + b_8 t^2) + u_{it}$ where the HeShe variable is a binary indicator on using the main series in the paper. Specification (3): $Y_{it} = AfterMetoo(b_0 + b_1 t + b_2 t^2) + HeShe(b_3 + b_4 t + b_5 t^2) + v_t + u_{it}$. Specification (4): $Y_{it} = AfterMetoo(b_0 + b_1 t + b_2 t^2) + v_t + m_i + u_{it}$, where a user which tweets in both series is counted as two separate users. Aggregating the data to days in specification (1) - (3) and estimating Newey-West standard errors with a lag length of 4 does not change the results, except for in specification (2) where the coefficient of the common linear time trend b_1 (not shown) becomes significant at the 5 percent level.

Another source of bias can be that Geni performs worse because users introduce new words and phrases that she does not know, which would lead to a decrease in accuracy, in the level of *Follow Norms*, even in the absence of the Metoo event. So far, we have relied on the test set of Year 1, in which Geni was not trained, for approximating the level. However, the decrease in accuracy due to the language evolving is most likely enhanced by time. Table 10 uses the Togetherness series as a comparison group. We do not expect the main series *Follow Norms* and the Togetherness series to follow the same time trends, and thus such are included in the specifications. Column (1) shows a placebo regression on the Togetherness series and displays that the series is stable over time. The result points to no decreases in accuracy due to the language evolving. The rest of Table 10 re-estimate the main specifications because the predictive accuracy can be worse certain days due to the unrecognition of new words and phrases. Column (2) - (4) present that the previously mentioned U-shaped pattern exists in the main series, as compared to the stable Togetherness series, after the Metoo event. In column (4), user fixed effects are included in addition to the calendar day fixed effects. A user is counted as two separate users even if the user tweets in both series, since the two different norms might be held separately. The users behind the Togetherness series are included on the same basis as the main series, i.e. by tweeting on either side of the event. The specification compares how users on average change their tweeting pattern after the event while controlling for decreases in accuracy due to the language evolving within calendar days. Column (4) exhibits the strongest U-shaped pattern. More specifically, the daily averages of how users tweet from the 8th of December to the 28th of March are significantly lower, at the 5 percent- level, than the average level of *Follow Norms* before the event (not shown). The combined result shows that decreases in accuracy due to users introducing new words and phrases do not drive the main results. Using the tweets of Year 1 as a comparison group is preferred since they allow for controlling for seasonal variation; the baseline estimate in Table 6 column (2) is the preferred one.

10 Conclusion

This paper interprets the exact date for the Metoo event as exogenous and uses it to evaluate if gender norms can change rapidly. Previous authors present evidence of gender norms being constant over generations (Fernández et al., 2004; Alesina et al., 2013; Nollenberger et al., 2016), but this paper finds that gender norms change in relation to the Metoo movement. More and more conversations take place online which generates text data, and this paper utilizes the newer data source efficiently due to the recent developments in machine learning algorithms. In the paper, gender norms are defined in a comprehensive data-driven manner by an LSTM neural network model, allowing for a measure on gender norms experienced by people in their everyday life.

In Swedish tweets six months after the Metoo event, gender norms are on average less adhered to than the five months before. My preferred estimate indicates that 18 Metoo movements would erase adherence to the gender norms that existed the year before. Most tweets are about males, but they start to talk more about females after Metoo. The tweets on average depict both males and females as following female gender roles to a greater extent. The Swedish Twitter sample is not representative of the average gender norm environment in Sweden, and the result is an indication of the possibility for norms to change rapidly.

The quite strong effect found is surprising, and further work could investigate which subpopulations contribute to it. An analysis of which topics contributing tweets cover can be performed to further our understanding of which type of users and tweets that change.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When Should You Adjust Standard Errors for Clustering? NBER Working Paper 24003, National Bureau of Economic Research.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *ArXiv e-prints*.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly journal of economics*, 119(1):249–275.

Bertrand, M., Kamenica, E., and Pan, J. (2015). Gender Identity and Relative Income within Households. *The Quarterly Journal of Economics*, 130(2):571–614.

Bisin, A. and Verdier, T. (2001). The Economics of Cultural Transmission and the Dynamics of Preferences. *Journal of Economic Theory*, 97(2):298–319.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv e-prints*.

Bowles, H. R., Babcock, L., and Lai, L. (2007). Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask. *Organizational Behavior and Human Decision Processes*, 103(1):84–103.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks*, 106:249 – 259.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Fernández, R., Fogli, A., and Olivetti, C. (2004). Mothers and Sons: Preference Formation and Female Labor Force Dynamics. *The Quarterly Journal of Economics*, 119(4):1249–1299.

Fortin, N. M. (2015). Gender role attitudes and women's labor market participation: Opting-out, aids, and the persistent appeal of housewifery. *Annals of Economics and Statistics*, (117/118):379–401. Pre-print accessed.

Fryer, R. G. and Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, 2(2):210–40.

Goldin, C. (2006). The Quiet Revolution that Transformed Women's Employment, Education, and Family. *American Economic Review*, 96(2):1–21.

Google's Ngram Viewer. `https://books.google.com/ngrams`. Accessed: 2018-07-30.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, Gender, and Math. *Science*, 320(5880):1164–1165.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2):251–257.

ISS (2016). Svenskarna och sociala medier 2016: En del av undersökningen Svenskarna och internet 2016. Rapport from ISS, Internetstiftelsen i Sverige. (ISS is a non-profit organization running the .se domain registry).

Leibbrandt, A. and List, J. A. (2014). Do Women Avoid Salary Negotiations? Evidence from a Large-scale Natural Field Experiment. *Management Science*, 61(9):2016–2024.

Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., and Hertel, G. (2015). A Meta-analysis on Gender Differences in Negotiation Outcomes and Their Moderators. *Psychological Bulletin*, 141(1):85.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv e-prints*.

Mueller, G. and Plug, E. (2006). Estimating the Effect of Personality on Male and Female Earnings. *ILR Review*, 60(1):3–22.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nollenberger, N., Rodríguez-Planas, N., and Sevilla, A. (2016). The Math Gender Gap: The Role of Culture. *American Economic Review*, 106(5):257–61.

North, D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

Pope, D. G. and Sydnor, J. R. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives*, 24(2):95–108.

Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2017). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. *ArXiv e-prints*.

Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 616–623.

Shen, L., Lin, Z., and Huang, Q. (2015). Learning Deep Convolutional Neural Networks for Places2 Scene Recognition. *ArXiv e-prints*.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35(1):4–28.

Statistics Sweden (2016a). `http://www.statistikdatabasen.scb.se/sq/54996`. Data on female first names for 2016.

Statistics Sweden (2016b). `http://www.statistikdatabasen.scb.se/sq/54997`. Data on male first names for 2016.

Steinert-Threlkeld, Z. C. (2018). *Twitter as Data*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Williamson, O. E. (2000). The New Institutional Economics: Taking Stock, Looking Ahead. *Journal of Economic Literature*, 38(3):595–613.

Wu, A. H. (2018). Gendered Language on the Economics Job Market Rumors Forum. *AEA Papers and Proceedings*, 108:175–179.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. *ArXiv e-prints*.

Appendix A. Model Specifications and Selection

A.1 Overview

This appendix covers how natural language processing (NPL) type of deep learning models are trained to detect gendered language in an unsupervised fashion. The models learn to predict whether the word *he* and *she* should fill the blank spots (which originally contained the words) in Swedish tweets. One dimension of gendered language is used to infer the rest of the dimensions. This appendix covers specifications and model selection. The final model selected is evaluated on test data in the main article and is additionally used as a measurement tool to evaluate the #Metoo movement on another test data set. A "control" model is also trained to predict *I* and *we* in an equivalent manner. The "control" model is used to generate a comparison series for use in the evaluation of the #Metoo movement.

A.2 Data

I have downloaded 58 million non-retweeted Swedish tweets from 2016-05-01 to 2017-04-30 (called Year 1 in the main article) from Twitter's API. Word embeddings are pre-trained on the full dataset, but for the model training, two main sets are created. The he/she set contains tweets with either the word *he* or *she* appearing and the I/we set contains tweets with either the word *I* or *we* appearing.

Table A.1: Sets and sample sizes

Set	Original		Adjusted	
	Count (millions)	Class imbalance (fraction)	Count (millions)	Count 10-25 words (millions)
He/she	1.5	0.255	1.5	1.1
I/we	8.3	0.240	1.5	1.1

Table A.1 (Original, Count) presents that the total count of the I/we tweets are larger at 8.3 million than the he/she tweets at 1.5 million. Since the I/we tweets yield a "control" model, whose performance is dependent on the size of the input dataset, the I/we tweets are sub-sampled to have the same sample size as the he/she tweets, e.g. 1.5 million (Adjusted, Count). Also, Table A.1 (Original, Class imbalance) presents the class imbalance as the percent of appearances of the infrequent class, *she* or *we*. There are 25.5 percent she tweets in the he/she set and 24 percent we-tweets in the I/we set. The two sets are similarly imbalanced, and, hence, no adjustment is needed for the I/we tweets to yield a good "control" model in this regard.

Table A.1 (Adjusted, Count 10-25 words) displays the sample sizes of 1.1 million for both sets after they are cut to only include tweets with a word count between 10 and 25 words. The lower bound of 10 words is motivated by tweets below containing little information to predict on and reduces the original sample by around 10 percent. The upper bound is motivated by model technical reasons. When the input tweets are of similar length, a better performing model is easier to find. The upper bound reduced the original sample by around 20 percent. An alternative would be to train multiple sub-models and thereby take into account input tweets of various length, but as the main goal is to detect gendered language, a model using 10-25 words is deemed sufficient.[1] Finally, the two sets are further subdivided by a 64-16-20 random split to a training, validation and test set.

One can expect noise in the binary variable indicating if *he* or *she* should be put in place of the placeholder. In Swedish dialects, *he* (han) can be substituted for *him* (honom), e.g. the "grammatically correct" sentence *I gave the ball to him* (Jag gav bollen till honom) can be transformed to *I gave the ball to he* (Jag gav bollen till han). The expectation is due to language-historical reasons. I find that 'grammatically incorrect' Swedish dialects only introduce label noise at around 2 percent in the *he* class.[2] This level of noise is not a concern for the deep type of networks considered in this appendix.

[1]The two sets are still similarly imbalanced at 25.1 and 23.9 percent, respectively.

[2]To investigate the concern of substitution due to "grammatically incorrect" Swedish dialects, I evaluated a randomized balanced sample of 1000 he/she tweets. The substitution occurs in he tweets at a rate of 1.1 to 3.9 percent, with a point estimate of 2.2 percent. (From an exact binomial test where the null hypothesis is that the proportion is equal to zero). There is no language-historical reason to expect the same kind of substitution for the female pronouns *she* (hon) and *her* (henne) and no substitution was found.

A.3 Preprocessing

The tweets are preprocessed by replacing all user names, URLs and hashtags with corresponding placeholders. *NLTK Twitter Tokenizer* is used to divide the tweets into words. It is chosen since it preserves emoticons, such as ☺, that oftentimes are not considered to be words.

A.3.1 Word embeddings

A word vector (word embedding) is a d-dimensional vector striving to represent a word's relationship with other words and it is used as input to the networks. Available word embeddings for Swedish are exclusively trained on Wikipedia text which most likely does not include the same type of language as Twitter, and hence I have trained my own. The word embeddings are trained with the Skip Gram algorithm of Mikolov et al. (2013) since it is a standard training method. They are trained on the full 58 million-tweet data with *Gensim*. Implementation details are presented in Table A.2. There are 575 000 tokens and 2.4 billion tokens to train on after the window length is set to ten. The words with a count lower than ten are replaced with a placeholder to be able to initialize such words during training. The Swedish language does not use compound nouns to the same degree as the English language, i.e. a "full moon" would be written "fullmoon" in Swedish (fullmåne), meaning that the model will have more word vectors but fewer co-references to solve than the standard English case.

Table A.2: Word embedding implementation details

Hyperparameter	Value
Word vector dimension	300
Window Length	10
Negative sample size	5
Min Count	10
Nr of epochs	5

Feature	Counts
Sentences	58 million
Tokens	3.4 billion
Tokens >Window Length	2.4 billion
Unique tokens >Window Length	575 000

A.3.2 Masking of gendered words

It can be argued that the model should capture more nuanced gender-colored language than simply using words such as *boyfriend* and *grandma* to predict. Therefore is in the main model 'too obviously' gender-colored words masked. In the masked data, fictive words are created by collapsing the original ones. For example, the words *mum* and *dad* have been masked to *mumdad*. When using word vectors, the new masked vectors are created by summing the original ones and re-normalizing them. I hand-select suitable words to mask by investigating lists of high predicting words, firstly, from the optimized Naive Bayes model and, later, from an LSTM neural network model trained on already masked data. I deem it suitable to mask the following types of words:

1. Tuples that consist of mappings of female words to a male equivalents, such as sister-brother. The set was expanded with similar words as defined by the trained word vectors. Some high predicting words or neighbors were deemed impossible to map, such as witch (*häxa*) and shallow-and-stupid-female (*bimbo*), and are not masked for this reason.

2. Family names of famous persons.

3. Personal names that consist of all Swedish residents' first names from a list obtained from Statistics Sweden (2016a,b), except for names that are also another word in Swedish (*kalla, du*) or English (*star, honey*).

Table A.3 displays the counts of the mask lists and the effective words found in training.[3] The word count of the training set is reduced by three percent when masking.

[3]See online appendix table "Words used for masking" for complete lists of words used for masking.

Table A.3: Counts of masked words

	Description	Source	Tuple count on list	Word count on list	Word count found in training
1	high- predictors	own	421	932	753
2	famous family names	own	2	815	815
3	first names	Statistics Sweden	2	22751	5531

Note: The tuple count is not equivalent to the word count on the mask lists. For example, the words *wife* and *husband* are in the same tuple as *hubby*. The effective word counts found in the training set is displayed as it differs from the theoretically possible words to mask from the lists.

A.4 Model Specifications

Three main model types are evaluated: a Naive Bayes model, a one-layer LSTM model and a three-layer CNN model. The Naive Bayes model functions as a baseline and does not take into account word order, e.g. no time dimension is modeled. Both the LSTM and CNN models have shown competitive results on NPL tasks[4], and both model a time dimension, although with different flavours.

An input sample is a sentence represented by $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ where $\boldsymbol{x}_t$ represents a word by a one-hot (dummy-variable) encoded vector. $\boldsymbol{X} \in \mathbb{R}^{t \times v}$ where t is the number of words in the sentence and v the number of words in the vocabulary. The curse of dimensionality composed by the general problem formulation is overcome differently dependent on the model type. A count-based model such as the Naive Bayes reduces the one-hot encoded word representation to a scalar of the count of the word, $\boldsymbol{X} \in \mathbb{R}^{v}$. The LSTM and CNN models instead use word vectors. With word vectors, a one-hot encoded word representation is reduced to d-dimensional vector striving to represent the word's relationship to other words, $\boldsymbol{x}_t^{word\ vector} = \boldsymbol{B}^{Reduce} \boldsymbol{x}_t^{one-hot}$ where the $\boldsymbol{B}^{Reduce}$-matrix is defined by the pre-trained word embeddings. $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ where d represents the dimensionality of the word vectors. When re-training of the word vectors are allowed, the parameters in $\boldsymbol{B}^{Reduce}$ are re-estimated and the model operates on the full $\boldsymbol{X} \in \mathbb{R}^{t \times v}$-matrix. The tweets are prepossessed by substituting *he* and *she* by a common placeholder. When word vectors are used, a representation for the common placeholder is initialized by summing up the original ones and re-normalizing them.

[4]For an overview of recent results see Young et al. (2017).

The output y is a binary indicator for class 1, the *she* class. As standard for a two-class problem, models are trained to minimize cross-entropy loss $-\sum_i y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i})$, where $\hat{y} = P(C_1|\boldsymbol{X}) = \sigma(a) = 1/1 + e^{(-a)}$ with $a = log\left[P(\boldsymbol{X}|C_1)P(C_1)/P(\boldsymbol{X}|C_2)P(C_2)\right]$. Minimizing cross-entropy loss corresponds to maximizing the log-likelihood function behind a logistic model.

(1) NB: A baseline neural network Naive Bayes model is trained, which is equivalent to estimating a linear logistic regression with the count of each word as independent variables:

$$a = log\left[P(C_1)/P(C_2)\right] + \sum_{v=1}^{V} log\left[p_{1,v}/p_{2,v}\right] x_v = c + \sum_{v=1}^{V} b_v x_v \quad \text{(A.1)}$$

where $p_{c,v}$ represents the probability for word v in class c and x_v represents the count of word v in the sentence. The model is naive since it does not take word order into account in the dimensionality reduction, as it assumes that the words in each class are independent. In some standard descriptions of a Naive Bayes model, the baseline count for each word is considered to be a hyperparameter (Rennie et al., 2003), which instead is estimated by the bias node (the baseline log odds ratio $log\left[P(C_1)/P(C_2)\right]$, the constant) in the logistic regression specification.

(2) BoW- WV: A naive 'bag of words' – specification on the word vectors is trained, which is equivalent to a logistic regression where each word vector dimension is an independent variable:

$$a = c + \sum_{t=1}^{T} \sum_{d=1}^{D} b_{d,t} x_{d,t} = f(\boldsymbol{XW} + c) \quad \text{(A.2)}$$

where $x_{t,d}$ represents the d:th dimension in word vector t in the sentence. The right-hand side expressions follow the standard of calling parameters in a network weights and representing them by W. f is a linear activation function, e.g. $f(z) = z$. Bolukbasi et al. (2016) show that unwanted gendered language is linearly separable in the word vectors trained on English Google News. This suggests that a linear specification might be sufficient and the current specification illustrates how performance changes as compared to the non-linear LSTM and CNN specifications. In addition, it illustrates possible performance gains due to the usage of word vectors as compared to the Naive Bayes model.

(3) **LSTM:** The LSTM network implemented in this paper has one layer and a fairly bare-bone structure compared to how LSTM nodes, or other recurrent nodes, are used for NPL tasks in the literature. However, for example, including a bidirectional structure did not aid this specific task when tried during undocumented initial model selection. Thus, instead of trying various design tweaks, I decided to implement a CNN network as a comparison since it is a more flexible design.

A Long Short-Term Memory (LSTM) network is a subclass of Recurrent Neural Networks (RNN). For NPL tasks are input to the network a sequence of word vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$, $\boldsymbol{X} \in \mathbb{R}^{t \times d}$. An LSTM node allows for modeling longer time sequences by "automatically detrending" the time series. Various gates are used to "filter" the input of the LSTM node $\boldsymbol{x}_t$ to the output of the node h_t. More specifically, a forget gate f, input gate i, and an output gate o are used:

$$
\begin{aligned}
i &= \sigma(\boldsymbol{u}^i \boldsymbol{x}_t + w^i h_{t-1} + b_i) \\
f &= \sigma(\boldsymbol{u}^f \boldsymbol{x}_t + w^f h_{t-1} + b_f) \\
o &= \sigma(\boldsymbol{u}^o \boldsymbol{x}_t + w^o h_{t-1} + b_o)
\end{aligned}
\tag{A.3}
$$

where the $\boldsymbol{u}$:s are respective parameter vectors, the w:s are respective weights (parameters) and the b:s respective bias weights (constants). The logistic function $\sigma(z)$ maps the gates to $(0, 1)$. The "filtering" specifies as follows:

$$
\begin{aligned}
h_t &= tanh(c_t)o \\
c_t &= c_{t-1}f + \tilde{c}_t i \\
\tilde{c_t} &= tanh(\boldsymbol{u}^c \boldsymbol{x}_t + w^c h_{t-1} + b_c)
\end{aligned}
\tag{A.4}
$$

where $\boldsymbol{u}^c$ is a parameter vector, w^c a weight and b_c a bias weight. The output of the LSTM node h_t is achieved by filtering the internal memory of the structure c_t through the output gate o. In turn, the internal memory c_t is a combination of its value at the previous timestep c_{t-1} filtered by the forget gate f and a candidate update value $\tilde{c}$ filtered by the input gate i. In the proceeding step, the candidate value $\tilde{c}$ is a combination of the input $\boldsymbol{x}_t$ and the output of the node at a previous timestep h_{t-1}. The $tahn$-functions maps to (-1,1) and thus allow for outputting or updating the sequence between the various timesteps. The full LSTM network consists of many LSTM nodes and specifies as follows:

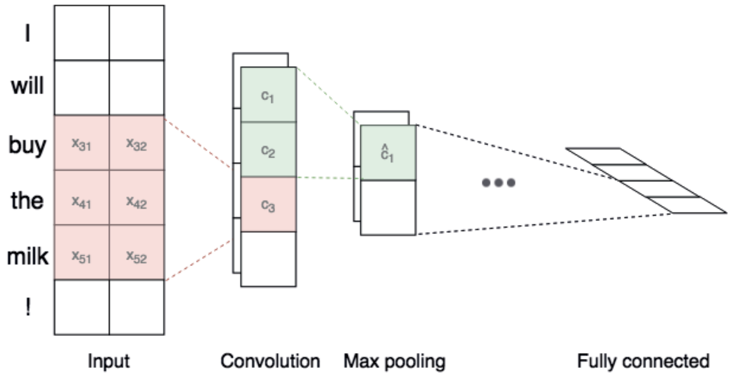
$$
\hat{y} = \sigma(a) = \sigma(\boldsymbol{W}^h f(\boldsymbol{h}_{t=T}) + c^h) \tag{A.5}
$$

where f is a rectified linear activation function (ReLu), $f(z) = max(z, 0)$, applied element-wise. In contrast to the above models, a is now a weighted sum of the output of the various LSTM nodes $\boldsymbol{h}_{t=T}$ taken at the last timestep T. The full equation is equivalent to the last layer of the network. Dropout is applied to the last layer to regularize it. The features $\boldsymbol{h}_{t=T}$ learned in the proceeding hidden LSTM layer are randomly dropped out with probability p during training and the parameters $\boldsymbol{W}^h$ rescaled with the same probability when training is finished. Both the number of nodes in the LSTM layer and the fraction of dropout are examples of hyperparameters that can be tuned.

(4) **CNN:** A three-convolutional layer CNN model is trained which step-wise reduce the number of n-grams from $\{4, 5\}$ in the first convolutional layer, to 3 in the second and 2 in the third. Each convolutional layer consists of 100 kernels. The input layer is zero-padded and the convolutions are narrow/valid. Max pooling is performed with a window size of 2 and, as conventional, the same stride. The last

fully connected layer consists of 100 neurons. The CNN design is similar to the one found in Poria et al. (2017), but overall it follows a standard CNN architecture for an NPL task. Instead of operating on an image, an NPL CNN operates on an input sentence represented by the matrix $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ with 1-d convolutions. Each kernel is applied to a window of h words to produce a new feature.

In a CNN model, a convolutional layer is normally built up in two steps: a convolution and max pooling. The model structure is illustrated in Figure figure A.1. In the convolutional step, a kernel $\boldsymbol{k} \in \mathbb{R}^{hd}$ is applied to a window $\boldsymbol{x}_{t:t+h} \in \mathbb{R}^{hd}$, a concatenation of vectors from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t+h}$, to produce a new feature $c_i = f(\boldsymbol{k}^{\mathrm{T}}\boldsymbol{x}_{t:t+h-1} + b)$. f is a rectified linear activation function (ReLu), $f(z) = max(z, 0)$. One kernel is applied to all possible windows to produce a new feature map $\boldsymbol{c} = \{c_1, \ldots, c_{t-h+1}\}$. The parameters in the kernel are shared across the windows. In the max pooling step, with downsampling and conventional stride settings, another filter is applied, $\hat{\boldsymbol{c}}_j = max\{\boldsymbol{c}_{j:j+h-1}\}$, yielding a final feature vector $\boldsymbol{z}$. In each convolutional layer, multiple kernels and feature maps are normally applied. A CNN model can include many convolutional layers after each other. The last layer in the CNN is a fully connected layer with concatenated input $\boldsymbol{z}$ from the last max-pooling layer, $a = \boldsymbol{W}\boldsymbol{z} + c$ (with the set-up used in this paper). Dropout is again applied to the last layer, but now the features in the last max-pooling layer randomly are dropped out.



Note: The figure illustrates a CNN structure with an example of the input sentence $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\} = \{I, will, \ldots, !\}$ where t=6 and d=2. In the figure is a kernel $\boldsymbol{k} = \{k_1, \ldots, k_6\}$ set to h=3 illustrated sliding over the window $\boldsymbol{x}_{3:5} = \{x_{3,1}, x_{3,2}, x_{4,1}, \ldots, x_{5,2}\}$ to produce the new feature c_3 in the new feature map $\boldsymbol{c}$. The figure also illustrates the max pooling step with a kernel size of h=2 with $\hat{c_1} = max\{\{c_1, c_2\}\}$. That multiple kernels are applied in each set is illustrated by having 2 kernels and feature maps. The possibility of having multiple convolutional layers is in the figure illustrated by the three horizontal dots $(\ldots)$. The figure also illustrates that the last fully connected layer in the CNN is obtained by concatenating the last max pooling layer. (Dropout is not illustrated.)

Figure A.1: CNN illustration

A.5 Hyperparameter Tuning

Hyperparameters are tuned on the masked he/she tweets and, thus, model selection is optimized based on finding a model which takes into account more nuanced gendered-color language. An increase in accuracy of 1 percentage point is considered an improvement.[5]

The main set of hyperparameters tuned is if word vectors should be re-trained and methods to deal with class imbalance. Re-training of the word vectors are only applicable for the LSTM and CNN models. Depending on the quality of the initialized word embeddings for the task at hand, re-training can decrease performance by adjusting the word vectors in the smaller training set only to make

[5] The limit is based on eyeballing what Young et al. (2017) consider to be marked improvements in their overview of the literature on NLP in deep learning.

them less generalizable.[6] I also investigate methods to deal with class imbalanced during training. Undocumented initial model selection showed that overall accuracy decreases without such a method. The first method investigated consists of weightening the loss function to pay more attention to the underrepresented class. There exist around three times more he tweets than she tweets in the training data, and the she class is thus upweighted by a factor close to three. The method is standard, but since binary cross-entropy loss is used, errors will tend towards infinity, and the weightening might be less effective. The second method investigated consists of making each mini batch balanced. The strategy is mentioned by Buda et al. (2018) and used in Shen et al. (2015). The balancing of each mini batch is done by randomly redrawing samples until a sufficient number of the underrepresented class is available and then randomly cutting away samples representing the overrepresented class. Each mini batch approximates the empirical distribution of how the data would be if it was balanced. The method can be thought of as a stochastic gradient decent version of more standard methods of oversampling or undersampling.

After the main set of hyperparameters have been optimized, the following are also tuned: For the LSTM network, the number of nodes and the fraction of dropout applied to the weights are investigated, since they are important hyperparameters guiding generalization performance. More specifically, $\{125, 250, 500\}$ nodes and $\{0, 0.25, 0.5\}$ in dropout fraction was tried. For the CNN network, the fraction of dropout is tuned, again investigating fractions of $\{0, 0.25, 0.5\}$. The numbers of filters are not optimized since 100 filters per layer seem to be fairly standard.

The hyperparameters for the other models used in the main article, the unmasked he/she model and the unmasked I/we model, are set to be equivalent to the ones found for the masked he/she model as they consider a similar problem. In addition, not all hyperparameters are tuned for the masked he/she model since it would cost to much computational resources. Only normalized 300-dimensional word vectors trained by me have been tried. All tweets are prepocessed to the same input length by adding padding to the right of the original input vector, which is applicable only for all model types except the Naive Bayes model. All training is performed on the scale of one-tenth (1/10th) epoch since the sample sizes are large. The networks are trained with a mini batch size of 100 samples for max 5 epochs. Training stops earlier if the validation accuracy does not increase by 0.1 percent-

[6]To clarify, only words found in the training set are included in the model.

age point after 1 epoch. The untuned hyperparameters strive to follow standard settings for natural language processing- types of neural networks.

A.6 Validation Results

Table A.4 presents a summary of the hyperparameter-optimized results for each model type. The NB model presented in the first column is trained with *SciKit's* multinomial Naive Bayes implementation, as training with stochastic gradient decent in the neural network framework showed worse performance (not shown). The NB model's accuracy at 0.754 precisely beats the simple classification rule of predicting any sample to the he-class, which yields an accuracy of 0.749. The second column shows that a naive model with word vectors as input (BoW-WV) do not aid performance. The third column presents the best performing model, an LSTM neural network model which reach a ROC AUC score of 0.764 and an accuracy of 0.773. The fourth column displays that the CNN is a close runner-up with a ROC AUC score of 0.758 and an accuracy of 0.772. Thus, introspection of Table A.4 shows that the major performance gains arise when taking word order into account by adding non-linearities in the LSTM and CNN model types. More precisely, for the LSTM model, the ROC AUC score is 0.03 points higher and accuracy is 0.02 points higher than for the NB model. More nuanced test set results for the Naive Bayes and the LSTM model is found in the main article, but the ROC AUC scores are 0.7372 and 0.7629, respectively.

Table A.4: Validation summary results

Model type	(1) NB	(2) BoW- WV	(3) LSTM	(4) CNN
ROC AUC	0.7315	0.7180	0.7640	0.7588
Accuracy	0.7541	0.7567	0.7733	0.7721
Hyperparameters:				
Class imbalance method		Balanced Batch	Balanced Batch	Weightening
Training of word vectors			No	No
Nr of nodes			125	100
Dropout fraction			0.25	0.5

Note: Hyperparameter optimized results for masked he/she data for each model type evaluated on the unbalanced validation set.

Table A.5 displays result for the main set of hyperparameters iterated over – if word vectors should be re-trained and the method which should be used to deal with the class imbalance. The table shows that regardless of which method is used to deal with the class imbalance, re-training of the word vector hurts performance. One likely reason is that the word vectors are trained on the same type of data and thus re-training only decreases generalization performance. Table A.5 also shows that higher performance is achieved, regardless of re-training of word vectors or not, for the LSTM model type when each mini batch is balanced (Balanced Batch) instead of when the loss function is weighted (Weightening). For the CNN model type is the result less clear, but by favoring ROC AUC as the evaluation metric over accuracy, one reaches the conclusion that weightening is performance enhancing. For the BoW-WV model, the discrepancy between balancing each mini batch and weighting the loss function was also not large (not shown). One possible reason of why the discrepancy between the methods are larger for the LSTM model type can be that LSTM nodes 'detrend' the input data, rendering it more sensitive to imbalanced data.

Table A.5: Main hyperparameter optimization result

LSTM				
ROC AUC	0.7410	0.7540	0.6486	0.7061
Accuracy	0.7672	0.7698	0.7664	0.7651
Training of word vectors	Yes	No	Yes	No
Class imbalance method	Balanced Batch	Balanced Batch	Weightening	Weightening
CNN				
ROC AUC	0.7331	0.7566	0.7325	0.7588
Accuracy	0.7683	0.7725	0.7691	0.7721
Training of word vectors	Yes	No	Yes	No
Class imbalance method	Balanced Batch	Balanced Batch	Weightening	Weightening

Note: Iterations over hyperparameters for masked he/she data for the LSTM and CNN model types evaluated on the original unbalanced validation set. The LSTM model has 250 nodes and a dropout fraction of 0.5. The CNN has 100 filters and a dropout fraction of 0.5.

A.7 Further Suggestions

Some interesting findings of this project are beyond the scope of the current purpose. The idea of capturing gendered language can be used in a more practical application. Then, it would be interesting to investigate how a non-linear neural network model performs when the scalar word count representation is used as input instead of a word vector representation. It is a smaller model to implement and if yielding the same performance, it would possibly allow for a memory- or speed-enhancing experience by an end user. However, to my knowledge, it is not much discussed in the deep NLP literature. Also, it would be interesting to delve deeper into how differently the models predict, beyond using a simple metric such as ROC AUC scores or accuracy. An undocumented finding of this project is that the models predict individual tweets very differently, i.e. the overlap of how they capture gendered language is not huge. If used in a practical application, one would like to favor a model who mark words and phrases as gendered that agree with the thoughts of the end users on the matter.

A.8 Summary

Three model types are tried: a Naive Bayes, an LSTM and a CNN model. The best performing model type is a 1-layer LSTM model with 125 nodes and a dropout fraction of 0.25 which is trained by balancing each mini batch and not allowing for re-training of the word vectors.

Appendix B. Additional Information and Results

B.1 List of survey questions used to capture gender norms

- World Values Surveys

  1. When jobs are scarce, men should have more right to a job than women.
  2. On the whole, men make better political leaders than women do.
  3. Most men are better suited emotionally for politics than are most women.

- General Social Survey (US)

  1. Most men are better suited emotionally for politics than are most women.
  2. A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
  3. A preschool child is likely to suffer if his or her mother works.
  4. It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.

- National Longitudinal Survey of 1972 (US)

  1. A working mother of pre-school children can be just as good a mother as the woman who doesn't work.
  2. High schools counselors should urge young women to train for jobs which are now held mainly by men.
  3. It is more important for a wife to help her husband than to have a career herself.
  4. It is usually better for everyone involved if the man is the acheiver outside the home and the woman takes care of the home and family.
  5. Many qualified women can't get good jobs; men with the same skills have much less trouble.
  6. Men should be given first chance at most jobs because they have the primary responsibility for providing for a family.

7. Most women are happiest when they are making a home and caring for children.
8. Most women are just not interested in having big and important jobs.
9. Schools teach women to want the less important jobs.
10. Young men should be encouraged to take jobs that are usual filled by women (nursing, secretarial, work. etc.).

Table B.1: Missing UTC dates.

Year 1	Year 2	Year 2	Year 2
2016-09-05	2017-05-07	2017-09-24	2018-03-17
2016-09-19	2017-05-25	2017-10-03	2018-04-17
2016-10-15	2017-05-27	2017-10-09	2018-04-18
2016-10-16	2017-06-11	2017-11-12	2018-04-22
2016-10-18	2017-06-12	2017-11-14	
2016-10-20	2017-06-13	2017-11-17	
2016-10-29	2017-06-14	2017-11-18	
2016-11-01	2017-06-15	2017-11-19	
2016-11-02	2017-06-16	2017-11-20	
2016-11-03	2017-06-17	2017-11-21	
2016-11-04	2017-06-18	2017-11-22	
2016-11-05	2017-06-20	2017-11-23	
2016-11-28	2017-06-25	2017-12-01	
2017-02-15	2017-07-05	2017-12-09	
2017-02-26	2017-07-23	2017-12-27	
2017-03-11	2017-07-30	2018-01-09	
2017-03-12	2017-07-31	2018-01-11	
2017-03-13	2017-08-01	2018-01-28	
2017-03-14	2017-08-14	2018-01-29	
2017-03-18	2017-08-29	2018-01-30	
2017-03-28	2017-09-01	2018-02-18	
2017-04-14	2017-09-23	2018-02-20	

Note: UTC dates are presented because it is the level where downloading failures happen. Swedish dates differ to UTC dates with +1 or +2 hours (depending on daylight saving time).

Table B.2: Class imbalance in other Swedish text sources

Corpus	Description	Timespan	She Frac	She & Her Frac	n (thusands)
Bloggmix	Popular blogs	2008-2017	0.429	0.441	2 713
Familjeliv	Disussion forum on family life	2004-2017	0.418	0.424	36 788
Bloggmix	Popular blogs	1998-2007	0.370	0.376	182
SOUs	Governmental reports	1995-2016	0.366	0.367	160
Göteborgsposten	Newspaper	2004-2013	0.336	0.338	1 292
Flashback	Dicussion forum	2000-2017	0.327	0.351	17 029
Webnews	Newspapers	2004-2013	0.278	0.281	1 682
Forskning & Framsteg	Popular science magasine	1992-1996	0.272	0.284	1.8
Wikipedia		2017	0.225	0.231	1 050

Note: She Frac: *she*/(*she*+*he*) where *she* is the count of the word *she* (hon), inclusive of *She* (Hon), and *he* is the count of the word *he* (han), inclusive of *He* (Han).
She & Her Frac: (*she*+*her*) /(*she*+*her*+*he*+*him*) where, in addition to the definitions above, *her* is the count of the word *her* (henne), inclusive of *Her* (Henne) and *him* is the count of the word *him* (honom), inclusive of *Him* (Honom).
n (thousands) is the total count of the words *he*, *He*, *she* and *She* in thousands.
The various text sources are retrieved from Språkbanken/Korp (2018).

Table B.3: Frequent words over the predicted probability distribution evaluated at a balanced sample

WC	English	Swedish
0.01	contract	kontrakt
0.02	united	united
0.03	games	matcher
0.03	player	spelare
0.04	club	klubb
0.05	the season	säsongen
0.06	goal	mål
0.07	the game	matchen
0.08	game	match
0.09	play	spela
0.10	the ball	bollen
0.11	team	lag
0.13	played	spelade
0.14	football	fotboll
0.15	penalty	straff
0.16	field	plan
0.18	ready	klar
0.18	play	spelar
0.18	miss	missar
0.20	em	em
0.21	europe	europa
0.23	the chance	chansen
0.23	worse	sämre
0.25	president	president
0.25	<familyname>	<familyname>
0.27	short	kort
0.27	leave	lämnar
0.27	last	förra
0.28	score	poäng
0.30	bad	dålig
0.32	smallest	minst
0.32	latest	senaste
0.34	manage	lyckas
0.34	still there	kvar
0.35	before	före
0.35	against	mot
0.37	were	varit
0.39	good	bra
0.40	fuck	fan
0.40	in	in
0.43	like	ju

Continued on next page

Table B.3 – *Continued from previous page*

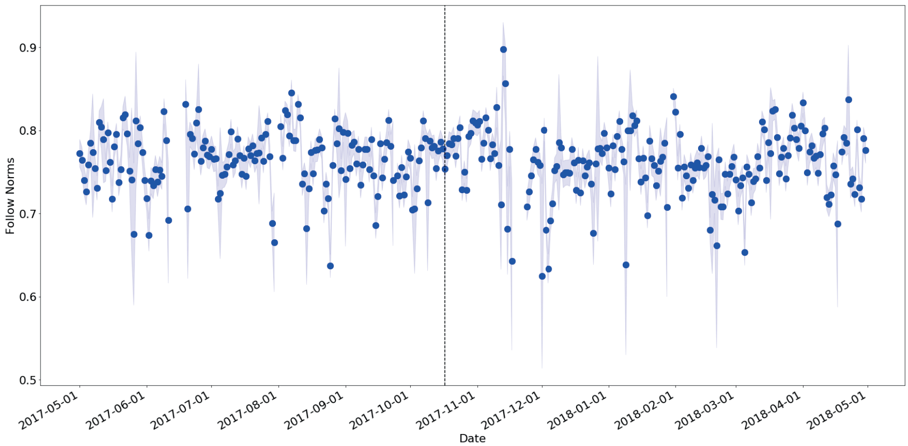
WC	English	Swedish
0.43	come	kommer
0.44	in	i
0.44	had	hade
0.44	.	.
0.45	<user>	<user>
0.46	,	,
0.47	___	___
0.48	are	är
0.48	at	att
0.50	<hashtag>	<hashtag>
0.51	and	och
0.52		
0.53	I	jag
0.54	<url>	<url>
0.55	you	dig
0.57	love	älskar
0.58	me	mig
0.59	my	mitt
0.59	<boygirl>	<killetjejkillle>
0.61	my	mina
0.62	*heart eyes*	*heart eyes*
0.63	children	barn
0.63		
0.64	*heart*	*heart*
0.66	<womenmen>	<kvinnormän>
0.66	<thedaughtertheson>	<dotternsonen>
0.66	my	min
0.67	<brothersister>	<brorsyster>
0.69	<sondaugther>	<dotterson>
0.71	friend	<väninnavän>
0.72	<mum'sdad's>	<mammaspappas>
0.72	<grandfathergrandmother>	<farfarfarmor>
0.74	beautiful	vacker
0.75	<mumdad>	<mammapappa>
0.76	fi	fi
0.77	<thewomenthemen>	<kvinnornamännen>
0.77	<grandfathergrandmother>	<morfarmormor>
0.78	pink	rosa
0.80	yearly	åriga
0.81	makeup	smink
0.81	nails	naglar
0.82	hagen	hagen
0.82	the cookie	kakan

Continued on next page

Table B.3 – *Continued from previous page*

WC	English	Swedish
0.82	pippi	pippi
0.86	noora	noora
0.87	book	book
0.87	thatcher	thatcher
0.88	raped	våldtagen
0.90	zara	zara
0.90	dress	klänning
0.91	pregnant	gravid
0.92	veil	slöja
0.94	the lady	tanten
0.94	<hishers>	<hanshennes>
1.00	lift	lyft
1.00	credit	credit
1.00	free	free
1.00	casino	casino
1.00	code	code

Note: The most "male" word with a WC of close to zero is at the top of the table and the most "female" word with a WC close to one is at the bottom of the table. The table is generated in the following way: The median predicted probability for each word is calculated from the predicted probability of each tweet, named Word Color (WC). The WC of all words is binned into 20 groups and for each quantile is the 5 most frequent words displayed. The table is generated on the test set and evaluated by the masked neural network model. The translation from Swedish is made by the author.



Note: The main dependent variable *Follow Norms*, a binary indicator on if Geni predicts the tweet correct, is displayed on the y-axis. The evaluation set of Year 2 is used. The vertical dotted line represents the Metoo event on the 17th of October 2017 and the lighter shaded regions represent daily confidence intervals.

Figure B.1: Gender norms over time

Table B.4: Metoo results with time trends

Dep Var	Time Trends Either Side (1) Follow Norms *100	Time Trends (2) Follow Norms *100	Time Trends User FE (3) Follow Norms *100	Time Trends User FE + Day FE (4) Follow Norms *100
Constant		76.842***	77.422***	
		(0.738)	(0.243)	
t		0.020	0.006	
		(0.021)	(0.006)	
t^2		-0.000	-0.000	
		(0.000)	(0.000)	
After Metoo	88.290***	11.449**	19.684***	6.354
	(5.598)	(5.646)	(1.632)	(7.770)
After Metoo*t	-0.086**	-0.107**	-0.154***	-0.060
	(0.044)	(0.048)	(0.014)	(0.066)
After Metoo*t^2	0.000*	0.000**	0.000***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Before Metoo	76.842***			-1.730
	(0.738)			(1.234)
Before Metoo*t	0.020			0.017
	(0.021)			(0.033)
Before Metoo*t^2	-0.000			-0.000
	(0.000)			(0.000)
Time FE	No	No	No	Yes
User FE	No	No	Yes	Yes
SEs	Clustered on day	Clustered on day	Clustered on user	Clustered on day & user
Tweets	989 028	989 028	849 518	1 053 209
Days (T)	319	319	319	365
Users (N)	95 025	95 025	35 338	65 165
Data	Year2	Year2	Year2	Year2/1

Note: The coefficients in the table can directly be interpreted as percentage points since the binary dependent variable has been multiplied by one hundred. Aggregating to daily data in specification (1) and (2) and estimating Newey West standard errors with a lag length of 4 do not change the results. *** p<0.01, ** p<0.05, * p<0.1.

Table B.5: Metoo results without user fixed effects, autocorrelation taken into account

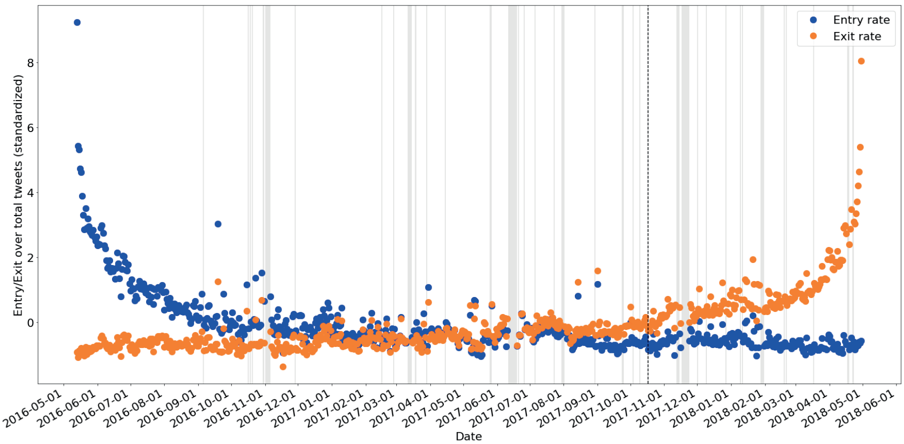
	Raw	Baseline	Cutoff independence	Female Focus
	(1)	(2)	(3)	(4)
Dep Var	Follow Norms	Follow Norms	Gendered Language	She Count
After Metoo	-0.003	-0.013*	0.012***	0.021**
	(0.005)	(0.007)	(0.004)	(0.008)
Time FE	No	Yes	Yes	Yes
User FE	No	No	No	No
SE:s	Newey- West with lag=4	Newey- West with lag=4	Newey- West with lag=4	Newey- West with lag=4
Days (T)	319	287	287	287
Data	Year2	Year2/1	Year2/1	Year2/1

Note: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table B.6: Metoo results estimated separately for tweets about females and males, autocorrelation taken into account

	She	He	She	He
	(1)	(2)	(3)	(4)
Dep Var	Gendered Language	Gendered Language	Follow Norms	Follow Norms
After Metoo	0.014***	0.007	0.023***	-0.004**
	(0.003)	(0.004)	(0.005)	(0.002)
Time FE	Yes	Yes	Yes	Yes
User FE	No	No	No	No
SE:s	Newey- West with lag=4	Newey- West with lag=4	Newey- West with lag=4	Newey- West with lag=4
Days (T)	287	287	287	287
Data	Year2/1	Year2/1	Year2/1	Year2/1

Note: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.



Note: The entry rate is the first observed tweets of users on a daily basis over total tweets. Instead, the exit rate is the last observed tweets. The entry and exit rates are standardized without any loss of pattern detection. The shaded areas represent missing dates. The analysis is limited due to unobserved tweets before and after the sample period, as well as for missing dates. For example, the peak of the exit rate at the end of the sample period is due to those dates being the last observed tweet for many users, but those users are likely going to tweet in future data which is not included. Likewise, before a region of missing dates, the exit rate will be higher since users whose last observed tweet is at a missing date is counted to a date before. An alternative analysis would necessitate a definition of what an active Twitter user is.

Figure B.2: Entry and exit rates

Paper II

Democracy and Economic Growth: Results from a Natural Experiment in Indonesia

Coauthor: Fredrik Sjöholm

Abstract

The form of democracy impacts political and economic conditions and thereby, plausibly, economic growth. However, the existing empirical literature does not show any clear results, partly because of the methodological challenges involved in cross-country studies. This paper examines the causal effect of different democratic systems on economic growth at a local level in Indonesia. Indonesia has changed from a parliamentary system where the local leader was appointed by the local parliament, to a presidential system with direct election of the local leader. The first direct elections of district leaders in Indonesia were performed in a staggered manner and decided such that the year of election is exogenous. Hence, we are able to compare economic growth in districts with a direct election of local leaders to growth in districts with an indirect appointment. Our estimations show no growth effects of the change to direct elections. The lack of growth effect corresponds with very small differences between the two types of districts in a number of governance variables.

Key Words: Asia, Indonesia, Democracy, Elections, Economic Growth

JEL Classification: H11, O10, O43

1 Introduction

One of the major distinctions of democracies is between presidential and parliamentary systems. In a presidential system, the executive (the president) and the legislature (the parliament) are both elected by the people, with a resulting separation of power between the two bodies. In a parliamentary system, the executive is indirectly elected by the legislature. Moreover, different democratic systems have been found to result in different types of political leaders and different types of economic policies, which could have an effect on economic growth. We address the issue of democratic systems and economic growth by relating local economic growth to the staggered and exogenous introduction of a local presidential system in Indonesia.

From a theoretical perspective, different democratic settings can impact economic growth through two mechanisms. The first one relates to the accountability of the political leader. A directly elected president (presidential system) is expected to perform better than an indirectly elected leader (parliamentary system) since the former is more accountable to voters (Persson and Tabellini, 2003). In other words, poorly performing leaders in a presidential system will be punished by the voters and elected out of office.

The second mechanism is the type of policies that different democratic systems foster. Theoretical work shows systematically different economic policies in presidential and parliamentary systems (e.g. Persson and Tabellini, 2003). One reason is that a parliamentary system necessitates the formation of larger coalitions which, in turn, diminish the role for special interest groups. As a result, parliamentary systems are shown to foster economic policies that are relatively better in line with the preferences of the majority of the population (Persson et al., 2000). Moreover, parliamentary systems result in relatively high public expenditures and high taxes, a tendency to redistribute resources towards the majority, less under-provision of public goods, and more rents to the politicians (Persson et al., 2000). Such important differences in economic policies between parliamentary and presidential systems could affect economic growth.

If the theoretical discussions are correct, that the form of democracies can affect economic growth, the policy implications are of obvious and large importance. However, the existing empirical literature does not show any clear result (see Voigt, 2011 for an overview). One reason is the methodological difficulties involved in such estimations. The existing empirical papers relate cross-country political systems to economic growth rates. One difficulty is that there are many unobserved country characteristics that presumably affect economic growth and which are difficult to control for. A few studies control for unobserved country characteristics by including fixed country effects in the estimations and hence estimate the effect of switching from one system to another. For instance, Persson and Tabellini (2006) find new parliamentary democracies to grow faster than new presidential democracies, whereas Knutsen (2011) find no significant difference in growth rates between the two types of democracy.

Another and perhaps even more serious methodological difficulty is that it is not necessarily a random process that makes some countries change their democratic system, and there could be factors affecting both the change of system and economic growth. Persson (2005) tries to overcome this endogeneity problem by an instrumental variable approach where settler mortality is used as an instrument for the form of democracy (see also Acemoglu et al. 2001). Democratization by adopting a parliamentary system is seen to increase economic growth. However, the used instrument is controversial, as discussed at some length in Albouy (2012).

We estimate the effect of parliamentary versus presidential systems on economic growth by looking at Indonesian districts that switch from having appointed district heads to having directly elected ones. A quasi-experimental research approach can be adopted since direct elections of local leaders in Indonesia were introduced in a staggered manner: the first elections took place in different years in different districts. The year of the first local direct election of a district head depended on when the incumbent district leader's term was due, which was a random event (see the discussions in Section 2.1 and 3.3). As a result, the year of the local election is exogenous. Combined with annual data on Indonesian districts' GDP and other economic variables, this allows for difference-in-differences estimations: we are able to compare the growth rates in districts that have had direct elections of local leaders against growth rates in districts that have not had such elections.

Our results show no effect of elections on economic growth: districts governed by a directly elected leader have about the same growth rate as districts with an appointed leader. Our results are stable to alternative specifications and measurements. For instance, the results are similar in more and less developed regions and are not affected by alternative definitions of the growth variable.

The results suggest that the form of democracy has had a limited impact on governance, which is confirmed when we examine a large number of indicators on governance in districts with and without elections. More precisely, districts, where there have been direct elections of local leaders, do not perform better than other districts when we examine, for instance, the business environment, infrastructure, licensing, and security.

The paper proceeds as follows. We start with a discussion of elections and governance in Indonesia. We continue with a discussion on the empirical approach, followed by the econometric results, a closer look at governance in Indonesia, and a concluding section.

2 The Setting

2.1 Background

The highly centralized authoritarian Indonesian regime of President Suharto came to an end with the Asian economic crisis of 1997/98. Widespread public protests and demonstrations lead to a transition towards a democratic political system, and the first free election was held in 1999.

The democratization has, by and large, been successful and elections are widely perceived as competitive, free and fair (Erb and Sulistiyanto, 2009; The World Bank, 2009, pp. 5-6). Moreover, a multitude of parties have emerged since the fall of the autocratic regime, the civil society is vibrant, and the media is free (Erb and Sulistiyanto, 2009, pp. 7, 15; Praktikno, 2009, p. 62; Freedom House, 2012; The World Bank, 2009; Buehler, 2009, p. 283). The voter turnout in the local elections is 73 percent on average, which is high by international standards (Schiller, 2009, p. 157).

Indonesia has a three-tiered government structure with elections at a national, provincial, and district level. Elections are performed in five-year cycles. There are elected assemblies and also elected leaders at all three levels: president, governor, and district leader.

The appointment of district leader has changed over time. Direct elections of leaders were introduced in 2005.[1] The year of the first direct election was decided by

[1] Law No. 32/2004

when the term of the incumbent leader expired. Some terms expired in 2005 and others in later years. Indirect elections of district heads were in place from 1999. Under indirect elections, a local parliament was elected, which in turn appointed a local district head. The first indirect election was taking place in the year when the incumbent leader's term expired. Some indirect elections took place in 1999 and others in the years before 2005. President Suharto, sometimes on the advice of the district parliaments, personally appointed district heads until 1999. The typical district head was a Javanese member of the Golkar (ruling) party with a background in the military ranks.

Again, the date of when the first indirectly elected leader was appointed was determined by when the term of the leader appointed by President Suharto expired, and the date of the first direct leader was determined by when the term of the indirect leader expired. It is important to note that the appointment of heads during the Suharto era was random as discussed and shown at some detail by Skoufias et al. (2014). More precisely, they examine dates of appointments of district heads between 1994 and 1999, hence during the Suharto regime. The number of yearly appointed heads averaged 42 and varied between 1 and 81. Moreover, they show in econometric estimations that the appointments are exogenous: the time of a new appointment under Suharto cannot be explained by any district characteristics. This exogeneity ensures that also the timing of the first directly elected district leader, from 2005 and onwards, is exogenous, which enables us to estimate a causal effect of direct elections on economic growth.

2.2 Do direct elections change the political setting?

There are indications that the direct elections have had an impact on the changes in leadership. More specifically, using names of the district heads for the period 2001 to 2007 allows us to calculate the share of incumbents that have remained in power. There was a change of district head in 51 percent of the districts that had direct elections in 2005.[2]

Hence, the direct elections of district heads seem to have enabled a change in leadership. Buehler (2010, pp. 273-5) confirms that the local elections have instilled

[2]Unfortunately, we cannot use turnover as a dependent variable since we lack data to identify which district heads that would have stepped down anyways. District heads cannot stay in office for longer than 2 terms.

competition, although mostly among elites. Moreover, Buehler (2009, pp. 117-9) and Mietzner (2011, pp. 133-6) suggest that, although wealth is a necessary requirement for a prospective district head (bribes, campaign advertising, etc.), it is not sufficient: they both highlight that local knowledge and attachment is the determining factor for success in local elections. Buehler (2009, p. 116) states that the parties' demands for money for nomination "act as early blockers or facilitators by default", and that this only "skims off candidates early in the electoral process". Many case studies underline personal characteristics, and not party affiliation, as important for voters' choice of district head (e.g. Sulistiyanto and Erb, 2009, pp. 12,16,20; Praktikno, 2009, p. 70). Sulistiyanto and Erb (2009, p. 20) argue that in the cases where incumbents have been re-elected, it is because they have performed well.

2.3 Can a district leader make a difference?

Indonesia has changed from one of the world's most centralized countries to one of the world's most decentralized (The World Bank, 2009, p. xvi; Buehler, 2010, p. 268). A major administrative decentralization took place in 2001 as a result of the democratization efforts after the fall of President Suharto. The district level became responsible for everything except "security and defense, foreign policy, justice, religious affairs and monetary policy", which is the responsibility of the national government (Mboi, 2009, p. 44). Obligatory responsibilities at the local level include such areas as health, education, public works, agriculture, industry and trade, transport and communications (The World Bank, 2008, p. 113).[3] In the decentralization process, agencies and personnel were transferred to district-level control, and the districts accounted for 69 percent of all civil servants employed in 2004 (Schiller, 2009, p. 148; The World Bank, 2008, pp. 17, 113).

Most power in the district is in the hands of the district leader, who sets the priorities for the budget, including the levels and types of spending, and is responsible for its execution. Moreover, the ultimate power rests with the district leader in cases of disputes between the district leader and the district parliament (Niazi, 2012).

The district level has also gained the financial means needed to perform its new re-

[3]According to Law No. 22/1999, implemented in January 2001, and later replaced by Law No. 32/2004.

sponsibilities. The districts can issue regulations, including taxes and charges (The World Bank, 2008, p. 125; Niazi, 2012, p. 396).[4] Districts are also receiving grants from the central government.[5] The districts have full discretion over their use of those revenues, with the exception of some special purpose grants.[6] The World Bank (2008, pp. 112, xv) concludes that "most regions now have enough resources to make a real difference for the lives of their citizens", and that Indonesia's level of fiscal decentralization is "higher than the OECD average and higher than any other East Asian country except China".

2.4 Does governance differ between districts?

The above discussion shows that districts have a large amount of autonomy and power in shaping their economic policies. An important question is whether this autonomy is reflected in differences in observed policies. There are case studies that indeed show large differences between Indonesian districts in terms of policies and governance. For instance, von Luebke (2009) examines governance in several Indonesian districts and finds substantial differences in the quality of governance. As an illustrative example, a business license that it takes two days to obtain in Yogyakarta takes 20 days in Medan. Niazi (2012, p. 397) argues that around 10-15 percent of regional governments have developed strong and effective leadership with good policies since the decentralization. At the other extreme are a significant number of districts plagued by poor governance, corruption, and money politics. Moreover, KPPOD (2008) surveyed the business climate in 234 Indonesian districts in 2007. The report claimed that there were strikingly large differences in the business climate and in the quality of economic governance.

Other papers find local institutions to differ across Indonesia, and that this difference has economic impacts.[7] For instance, Burgess et al. (2012) find institutional changes at the district level in Indonesia to affect economic behavior. Similarly, Olken (2007) finds local institutions to affect the business climate in Indonesia. Moreover, Skoufias et al. (2014) show that local direct elections affect the compo-

[4] Law No 32/2004.

[5] DAU (Dana Alokasi Umum).

[6] DAK (Dana Alokasi Khusus), which made up 3.2 percent of the districts' revenues in 2005.

[7] On a related note, Martinez-Bravo (2014) examines the effects of local institutions on political outcomes. She finds that appointed village heads are more likely than elected village heads to convince villagers to vote for the district ruler's party.

sition of local public expenditures, and Sjahrir et al. (2013) and Valsecchi (2013) find signs of increased corruption.

Hence, previous studies suggest that changes in governance might be taking place after local elections, but it is less clear whether governance is improving or deteriorating and whether it affects economic growth. We will continue to examine both issues in more detail below.

3 Empirical Approach

3.1 Data

The Ministry of Home Affairs in Indonesia provided us with the dates of district elections. Data on Gross Regional Domestic Product (GRDP), in constant prices and per capita at the district level, between 2003 and 2010, is from Statistics Indonesia (BPS). Growth in 2001 and 2002 and a large number of governance variables at the district level are from McCulloch (2011). Some districts have split into two or more new districts during our period of analysis. We have collapsed these new districts back to the original district. Hence, we use the district division from 1999, to conform to the district definition in some of our data. Thus, an essential data source is a dataset from Statistics Indonesia that enables the conversion of districts between different years. A more detailed specification of the data sources can be found in the Appendix Section A.0.1.

3.2 Difference-in-differences estimations

We will base our analysis on the standard difference-in-differences expression:

$$Y_{its} = \alpha + \lambda * Treat_i + \delta * AfterTreat_{ti} + v_t + u_{its} \tag{1}$$

where Y is the growth rate of constant GRDP. Subindex i denotes district, t denotes years, and s denotes whether the district belongs to the treatment or the control group. The variable $Treat$ takes the value 1 if the district is treated, i.e. held a local election in 2005, and zero otherwise. $AfterTreat$ takes the value 1 if the observation occurs after treatment: if the district held elections in 2005 and the observation is from later years than 2005. The parameter of interest is δ, the Difference-in-Differences estimator, which estimates the effect of direct elections of district leaders on economic growth. v_t refers to year fixed effects. Standard errors are clustered at the district level, to control for possible serially correlated residuals within districts.

We use different measures on economic growth. Firstly, we include growth in regional domestic product (GRDP) and growth in per capita GRDP. Indonesia is an oil-producing country and oil is concentrated in relatively few districts, mainly in East Kalimantan and the province of Riau on Sumatra. Volatility in oil prices will affect measured production and potentially bias our results. We therefore run our estimations also with non-oil GRDP and non-oil GRDP per capita. Finally, the Indonesian statistical bureau has offices collecting data in all districts. However, measuring local-level GDP is not without problems and, moreover, the quality of the staff responsible for doing so might differ between districts. It is therefore possible that GDP at a district level might be measured with errors. This would bias our results if the measurement errors within districts vary over time. We therefore run also estimations with consumption per capita as the dependent variable.[8] Information on consumption is collected in household surveys, which have been conducted for many years in Indonesia with a standardized method.

In the baseline estimations, our pre-treatment period is 2002-2004 and our post-treatment period 2006-2007. Hence, we examine the growth effect two years after local elections are held. Previous cross-country studies on the transition to democracy and economic growth have found a positive effect immediately after the transition.[9] However, one could argue that it should take longer for the economy to fully react to better or worse policies implemented as a result of elections. We, therefore, include additional estimations where we examine growth in the period 2002-2010. The main drawback is that part of the control group itself get treated, i.e. the control districts have local elections in 2008 or later, making a comparison between the treatment and control group confounded by the inclusion of later years. The main advantage is that this specification examines the growth effect up to five years after the election. Moreover, the results will be unbiased under the

[8]Data on consumption come from Statistics Indonesia and are found in McCulloch (2011).

[9]For instance, Papaioannou and Siourounis (2008) find a positive growth effect of around one percentage point already one year after the transition to democracy, and Rodrik and Wacziarg (2005) an annual similar effect in the first five years after the transition.

reasonable assumption that it will take equal time for elections to have a growth effect in the treatment and control districts.

Finally, we will examine the effect of local elections on various local policy variables as an additional way to examine the impact of direct elections of local leaders. Some of these policies could potentially change relatively fast after an election (see Section 5 below).

3.3 Identification and specification of treatment status

The first election of a district leader is the "treatment" of interest, and the effect is the difference in growth rates. Thus, the difference-in-differences estimation compares the differences in growth rates before and after elections in the districts that had local direct elections, to the districts that had no such elections.

The main benefit of our research methodology is the exogenous assignment of elections, which allows us to make a causal interpretation of the relationship between elections and growth. As discussed in section 2.1, elections were implemented in different years in different districts in a staggered manner.[10] To summarize:

- Districts leaders were randomly appointed by President Suharto up until 1999.
- District leaders were indirectly elected between 1999 and 2004. The first indirectly appointed leader was taking office when the term of the leader appointed by President Suharto expired.
- District leaders were directly elected after 2005. The first direct election took place when the term of the indirect leader expired.

Hence, the year in which a district held its first direct election was determined by the end of the term of the incumbent indirectly elected district leader, which in turn was determined by when district heads had been appointed under the previous regime. Consequently, the strict exogeneity assumption underlying the difference-in-differences method is satisfied.

[10] See alsoSkoufias et al. (2014).

Our treatment group includes districts that held local elections in 2005, and the control group includes districts that held local elections in 2008 or later. The districts that held local elections in 2006 and 2007 are not included in the analysis. There were 336 districts in 2002, the first year of our analysis. The 73 districts which held elections in 2006 or 2007 are excluded. Moreover, 44 other districts are also excluded. Some of them because they were split and the two new districts ended up in different groups (treated and control), and others because they belong to the capital Jakarta, which has a special governance system. We end up with a sample of 219 districts, 132 belonging to the treatment group and 87 to the control group.

3.4 Interpretation

Models on democratic systems and economic growth usually divide the effect into a selection effect and an accountability effect. The former arises since citizens taking part in elections have the possibility to choose a high-quality leader, whereas the latter arises since citizens affect the leaders' behavior while in office because of the desire for re-election. As the announcement of direct elections was made in late 2004, the incumbent leaders did not have much time to react. However, the district leaders in our control group had from 2004 until 2008 to prepare for their elections. We are therefore likely to capture mainly the selection effect, the populace choice of quality of leadership when we compare the treatment and control group. The accountability effect might be in operation for both the treatment districts with their elected leaders and the control districts with their unelected leaders.

3.5 Tests of pre-treatment variables

The parallel trend assumption needs to be fulfilled for identification in the difference-in-differences estimation framework: the treatment and the control group must have had parallel trends in the outcome variable before the treatment takes place.

Figure A.1 shows the economic growth rates in our treatment and control groups. The parallel trend assumption seems fulfilled: economic growth is very similar in the two groups in 2002, 2003 and 2004. Moreover, some of the figures suggest that economic growth in 2006 is relatively high in the treated districts.

To verify that the treatment and control districts do not differ in any important respect that could also be correlated with economic growth requires information on district characteristics. The rich data in McCulloch (2011) allows us to perform such a comparison. Table A.1 in the appendix shows the comparison of more than 50 district characteristics in the treatment and control groups. The variables are from 2000-2003, hence before the treatment in 2005. We divide the variables into four broad groups: general characteristics, social characteristics, governance characteristics, and economic characteristics. The comparison shows a large similarity between the treatment- and control groups: almost all variables are balanced. Religious fragmentation is one exception: there are more fractionalized districts in the treatment group. However, religious fragmentation is fairly time-invariant and controlled for in the difference-in-differences estimation.

A more interesting exception is that the two measures of investment, foreign direct investment and domestic investment, are significantly higher in the control group. Higher growth from high levels of investment might bias the results, and we will, therefore, run robustness estimations in which investment is included.

Finally, it would be unfortunate if the years of elections in the districts followed a geographic distribution since geographic factors are also likely to affect economic growth. Appendix Figure 1 shows the geographic distribution of our treatment and control districts: the districts in the groups are evenly distributed across the archipelago.

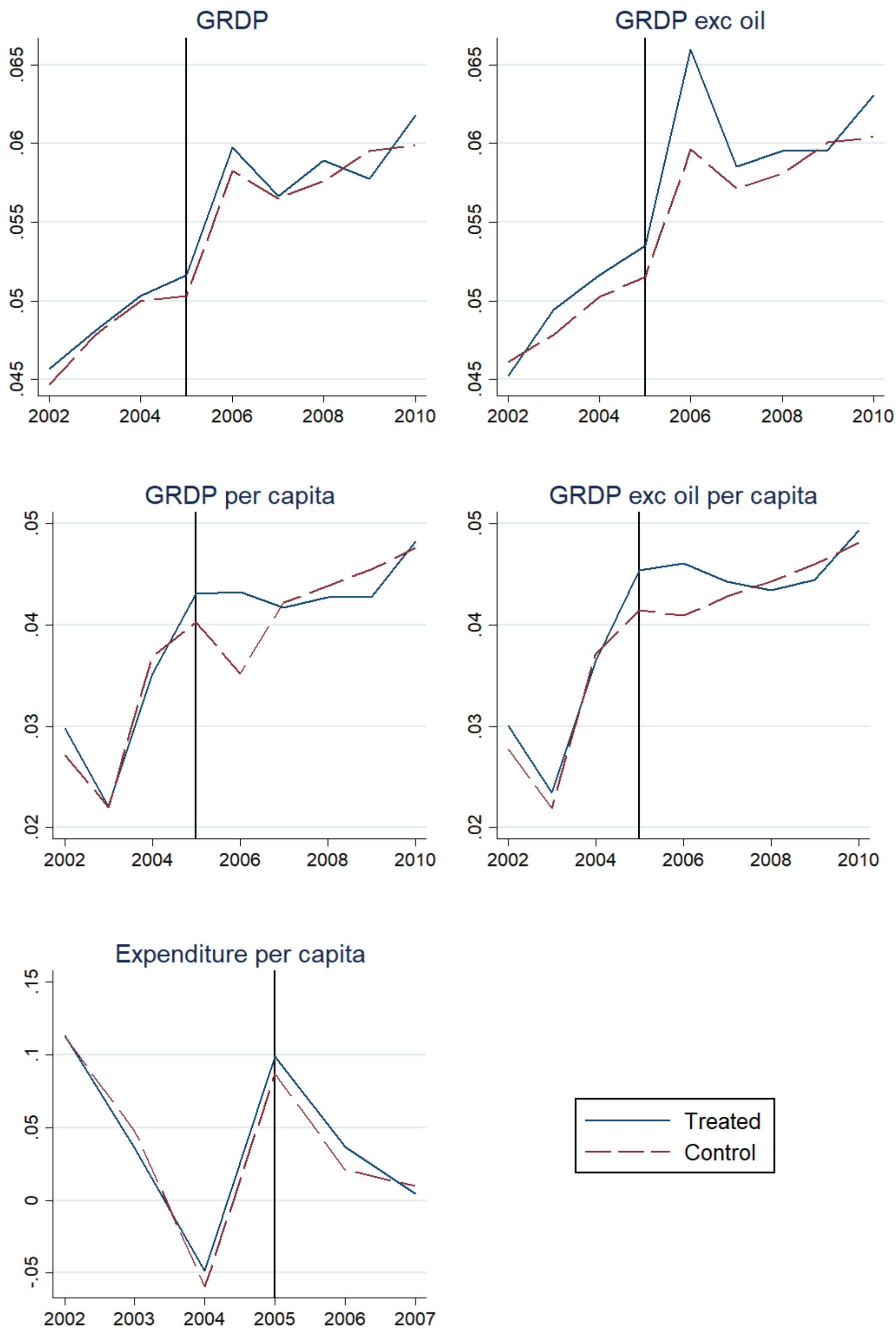


Figure 1: Economic growth in treatment and control districts 2002-2010 (%)

4 Econometric Results

Our base estimations on elections and growth are shown in Table 1. The coefficient is positive for direct elections in all estimations, but the size of the coefficient in the estimation with GRDP as the dependent variable is close to zero. The sizes of the coefficients in the other estimations would suggest that districts with direct elections had a growth rate roughly 0.3 percentage points higher than districts with indirect elections. However, the standard deviations are large, and none of the coefficients are statistically significant. Hence, we find no robust effect of local elections on economic growth.

Table 1: Growth in gross regional domestic product in districts with and without elections (2002-2007)

	(1) GRDP	(2) GRDP per capita	(3) GRDP excl. oil	(4) GRDP per capita excl. oil	(5) Consumption per capita
Did	0.000	0.003	0.003	0.002	0.005
	(0.003)	(0.004)	(0.003)	(0.005)	(0.009)
# obs	1,075	1,076	1,075	1,076	1,064
# clusters	219	219	219	219	219

Note: GRDP is Gross Regional Domestic Product. The growth estimations compare growth in 2006-2007 with growth in 2002-2004. Standard errors in parentheses are clustered at the district level.

Table 2 shows estimations where we expand the post-election period until 2010. Moreover, we show also the individual year effects. For instance, the DiD*2006 coefficient measures the growth effect one year after the election. Expanding the time period decreases the DiD coefficients, and most of the estimates are close to zero. Moreover, the standard deviations are large. Hence, we find no economic or statistically significant effect of elections on economic growth.

Table 2: Growth in gross regional domestic product in districts with and without elections (2002-2010)

	(1) GRDP	(2) GRDP per capita	(3) GRDP excl. oil	(4) GRDP per capita excl. oil
Panel A:				
Did	0.000	0.000	0.002	0.000
	(0.002)	(0.003)	(0.002)	(0.003)
Panel B:				
DiD*2006	0.001	0.008	0.006	0.004
	(0.004)	(0.007)	(0.005)	(0.009)
DiD*2007	-0.000	-0.001	0.001	0.000
	(0.002)	(0.003)	(0.002)	(0.003)
DiD*2008	0.001	-0.001	0.001	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)
DiD*2009	-0.002	-0.003	-0.001	-0.003
	(0.004)	(0.005)	(0.004)	(0.004)
DiD*2010	0.001	0.000	0.002	0.000
	(0.003)	(0.003)	(0.002)	(0.003)
# obs	1 727	1 729	1 728	1 730
# clusters	219	219	219	219

Note: The growth estimations compare growth in 2006-2010 with growth in 2002-2004. Panel A shows the average and Panel B the effect divided by year. GRDP is Gross Regional Domestic Product. Standard errors in parentheses are clustered at the district level.

As previously discussed, some of the districts were split which can impact the date of election and, hence, whether the district belongs to the control or the treated group. Our results will then be biased if a district split is endogenous to economic factors. We use two approaches to control for such bias. Firstly, we estimate a sample of districts from which we exclude the 12 districts that were split. Secondly, we estimate a sample of all districts and include a dummy variable for all districts that have split. The results are seen in columns 1 and 2 in Table 3. Direct elections are not, again, having a statistically significant effect on economic growth.

Table 3: Additional estimations on growth in gross regional domestic product (2002-2007)

	(1) Excluding district that split	(2) Dummy for split	(3) Java and Bali (more developed)	(4) Others Islands (less developed)	(5) Controlling for investmemt	(6) Expanded sample (2004-2007)
Did	0.001	0.000	0.001	0.000	0.001	0.001
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.005)
Dummy for split	No	Yes	No	No	No	No
Investment	No	No	No	No	Yes	No
# obs	1 020	1 075	453	622	587	859
# districts	207	219	92	127	188	322

Note: The growth estimations compare growth in 2006-2010 with growth in 2002-2004. Standard errors in parentheses are clustered at the district level.

Elections could perhaps have more of an effect in more developed parts of the country. Citizens might have better access to information in some districts, for instance, because of a higher literacy rate, and would consequently be better able to hold their local leader responsible when elections are implemented. Other districts might lack those favorable conditions, and be more prone to corruption and local elite capture. We, therefore, follow previous studies on regional development in Indonesia and divide our sample into districts on Java and Bali – the more developed districts – and those outside of Bali and Java – that are less developed (Skoufias et al., 2014; McCulloch and Malesky, 2010). Again, there were no effects of elections on economic growth, as seen in columns 3 and 4 in Table 3.

It was seen above that investment was higher in the control districts, and we continued by controlling for domestic and foreign investments. Thus, by controlling for investments, we essentially examine whether the lack of effect of local elections on growth in the previous estimations was due to lower investments in districts with elections.[11] The coefficient for direct election is once again, as seen in column 5, statistically insignificant.

As previously described, districts that split after 1999 and were the new districts ended up in different groups are excluded from the analysis above. We are able to increase the sample of districts to 323 (202 treated and 121 control districts) if we restrict our pre-treatment period to 2004. However, re-running our estimations on this larger sample of districts have no impact on the results, as seen in column 6.

Another issue is that economic growth rates might be correlated between neighboring provinces. This will not affect our estimates if the correlation is time-invariant. However, we also estimated specifications where the standard errors were clustered at the province level, in order to allow for some spatial correlation between neighboring districts. The results did not show any robust positive effect of direct elections on economic growth (not shown).

[11]It is not obvious that one should control for investments since it can be argued to be endogenous to the political system.

5 Governance

The effect of local elections on growth is supposed to occur via the channel of better governance. We, therefore, continue by examining whether there is any difference between the treatment and control groups in a set of governance variables. McCulloch (2011) and McCulloch and Malesky (2010) surveyed governance indicators in 2007. A fortunate aspect of the survey is that it was designed to measure governance aspects related to economic growth and targeted such areas that are under local government control (McCulloch and Malesky, 2010, pp. 10-11). Moreover, the survey was conducted in early 2007, roughly two years after local elections in the treatment group, but before elections in the control group.

The data allow us to compare means between the treatment and control groups for the governance variables. This is insufficient for us to make any conclusions about the causality between elections and governance, but it can serve as an indication of whether our previous results could be due to a lack of differences in governance.

There are 61 different variables included as shown in Table 4. These variables are also aggregated into nine broader categories covering access to information, business development programs, infrastructure, integrity, interactions with the business community, land issues, licensing, security, and transaction costs.

All variables have been normalized on a scale from 1-100, in such a way that 1 indicates the worst-performing district and 100 indicates the best-performing district. Hence, a higher value indicates better performance on the given policy, even if the variable name may indicate the contrary.[12]

It is worth getting back to the previously discussed issue of how long it will take for a new local head to have an impact on economic policies. Some of the variables included here, improved infrastructure, for example, are likely to take time to implement, and even longer to have an impact on economic growth. Most of the other variables, however, capture policies that are concerned with the general business climate, which should be possible to change in a relatively short period of time.

[12]For more information about the variables please see McCulloch (2011) and McCulloch and Malesky (2010).

Table 4: Comparison of Governance Characteristics in the Treatment and the Control Group

Variable	Mean of Treatment	Mean of Control	P-value
Access to Information Index	47.33	46.99	0.842
Ever tried to access government information	15.05	14.89	0.942
Overall impact of access to information on firm activities	79.60	78.61	0.743
Business Development Programs Index	42.13	41.03	0.577
Average share of firms saying 6 programs exist	30.18	28.95	0.645
Average share of firms participating in 6 programs	34.12	33.34	0.801
Average satisfaction with the programs	66.47	67.41	0.752
Overall impact of business development programs on firm activities	37.73	34.41	0.345
Infrastructure Index	64.06	71.32	0.000
Evaluation of quality of roads	54.99	59.43	0.053
Evaluation of quality of street lighting	67.03	70.64	0.180
Evaluation of quality of local water supply	62.56	67.13	0.147
Evaluation of quality of electricity	65.94	74.18	0.000
Evaluation of quality of telephone	62.05	63.42	0.534
Log time to fix roads	42.05	49.64	0.012
Log time to fix street lighting	66.72	74.07	0.004
Log time to fix local water supply	67.76	75.73	0.009
Log time to fix electricity	84.84	92.38	0.004
Log time to fix telephone	79.87	84.59	0.086
Ownership of a generator	62.72	70.50	0.021
Frequency of blackouts	83.04	91.47	0.002
Overall impact of infrastructure on firm activities	52.57	62.06	0.007
Integrity Index	56.78	56.16	0.736
District head's understanding of business issues	52.96	53.81	0.758
Local officials appointed based on relevant skills	52.78	55.65	0.275
District head takes strong action against corruption	57.97	55.62	0.359
District head (doesn't) take corrupt actions themself	42.73	39.79	0.221
District head is a strong leader	51.03	49.41	0.576
Overall impact of the capacity and integrity of the district head on firm activities	83.20	82.68	0.838
Interaction between Local Government and Businesses Index	55.08	53.56	0.335
Existence of a communication forum	35.58	30.41	0.060
Composite of: does the leader try to solve business problems; do the solutions meet your expectations; do the officials follow up	51.49	51.78	0.912

Continued on next page

Table 4 – *Continued from previous page*

Variable	Mean of Treatment	Mean of Control	P-value
Actions of the local government do not increase business costs	66.68	63.54	0.177
Actions of local government do not increase business uncertainty	55.75	51.12	0.084
Overall impact of issues associated with interaction on firm activities	72.45	72.86	0.885
Land Index	69.67	71.98	0.180
Weeks to get a land certificate	73.95	77.08	0.288
Ease of getting land	43.62	45.74	0.523
Infrequency of eviction in this area	77.64	78.50	0.798
Infrequency of land conflict	79.36	82.29	0.254
Overall constraint of land issues and legal uncertainty on firm activities	73.80	76.28	0.353
Licensing Index	59.40	61.71	0.091
Percentage of firms that have a TDP**	46.59	49.15	0.423
Average of: ease of getting a TDP and mean days to get a TDP. Of which:	73.43	75.60	0.167
–ease of getting TDP	58.80	62.02	0.122
–mean days to get TDP	88.07	89.19	0.482
Average of: cost of TDP and whether cost bothers them. Of which:	79.82	85.14	0.021
–cost of TDP	89.24	94.02	0.011
–whether cost bothers them	71.21	76.27	0.144
Combined score of three measures of efficiency of licensing. Of which:	52.32	52.91	0.838
–business licensing is carried out in an efficient manner	51.64	54.57	0.244
–business licensing is free of illegal collections	56.19	54.80	0.684
–business licensing is free of collusion with officials	49.11	49.36	0.941
Percentage of firms that say there is a complaint mechanism	29.28	30.26	0.786
Overall constraint of licensing on firm activities	74.94	77.21	0.404
Security Index	60.55	59.87	0.721
Composite opinion of how police handle cases	48.03	48.30	0.920
Quality of the police in dealing with worker demonstrations	46.92	48.39	0.515
Overall constraint of security on firm activities	74.92	73.34	0.613
Transaction Costs Index	67.12	66.99	0.952
How much does paying user charges bother the firm	65.06	66.27	0.739

Continued on next page

Table 4 – *Continued from previous page*

Variable	Mean of		P-value
	Treatment	Control	
Existence of user charges on the distribution of goods	66.06	64.46	0.677
Composite of: existence of voluntary donations and how much they bother you. Of which:	63.55	62.67	0.694
–incidence of paying donations	51.81	45.90	0.141
–donation impact of firm performance	75.29	79.43	0.107
Security payments to the police	70.65	70.06	0.861
Overall constraint of transaction costs on firm activities	70.14	71.49	0.658

Note: The P-value is for the t-test of equality of variable means across groups. TDP= business license. Source: McCulloch (2011).

Most indicators are not statistically different between treatment and control districts. However, in the cases where there is a significant difference, it is almost exclusively the case that the control group obtained higher scores than the treatment group. For instance, the control districts have significantly better performance on the variables relating to licensing and infrastructure. The comparison of means for the "licensing index" indicates that the average score of the control districts is 62, whereas the average of the treatment districts is 59. The corresponding figures for infrastructure are 71 for the control group and 64 for the treatment group.

Treated districts perform relatively better than the control districts in none of the aggregated policy variables in Table 4 and in only two disaggregated policy variables. The first is "actions of the local government do not increase business uncertainty", for which the treatment group scores 56 and the control group 51. The second is "existence of a communication forum", which captures one aspect of interaction between the local government and the local business community, and for which the treatment group scores 35 and the control group 30.

In summary, this result is consistent with our earlier finding of an insignificant effect of local direct elections on growth, since the channel of better governance, does not appear to be present in Indonesia.

6 Discussion and Concluding Remarks

How parliamentary and presidential systems affect economic growth has received a great deal of interest. We have approached this issue by examining the growth effect of changing from a parliamentary system to a presidential system in Indonesian districts. There have been large differences in economic growth between Indonesian districts, and there are also several case studies suggesting that governance differs between districts. However, we do not find these differences to be caused by indirect versus direct elections of local leaders. Most of the estimated coefficients are very small suggesting that the economic impact is negligible. In those cases where the coefficients are slightly larger, they remain far from being precisely estimated because of the large standard errors. Our overall conclusion is that we do not find any economic and statistically significant effect of elections. The interpretation is that citizens do not choose higher-quality persons as district heads than those appointed in an indirect way through the local parliament.

Our results are very robust to changes in the specifications. For instance, local GDP might be plagued by measurement errors but using households' consumption expenditures does not change the results. Moreover, it could be argued that it takes longer time before the full effect is in place and that our treatment period is too short. However, expanding our treatment period does not change the results.

The lack of a growth effect suggests that local governance is not affected by local elections in Indonesia. This is confirmed when we compare many different indicators of governance in districts with and without elections. Governance is rather similar in the two groups of districts and, when there are significant differences, the situation is often better in the districts without direct elections. Districts with direct elections receive better scores than districts without direct elections in only 2 out of the 61 variables capturing various aspects of governance.

There are limitations on the generalizability of our results: the growth effect of national elections might differ from that of local elections. This limitation notwithstanding, we do believe that our paper complements the existing literature in some important respects. Most importantly, our approach has enabled us to overcome the problem of endogeneity, which might have plagued previous studies and thereby allow us to estimate a causal effect of direct elections of district leaders on economic growth.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91(5):1369–1401.

Albouy, D. Y. (2012). The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *American Economic Review*, 102(6):3059–76.

Buehler, M. (2009). The Rising Importance of Personal Networks in Indonesian Local Politics: An Analysis of District Government Head Elections in South Sulawesi in 2005. In Erb, M. and Sulistiyanto, P., editors, *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*, pages 101–124. Institute of Southeast Asian Studies.

Buehler, M. (2010). Decentralization and Local Democracy in Indonesia: The Marginalisation of the Public Sphere. In Aspinall, E. and Mietzner, M., editors, *Problems of Democratisation in Indonesia: Elections, Institutions and Society*, pages 267–285. ISEAS Publishing.

Burgess, R., Hansen, M., Olken, B. A., Potapov, P., and Sieber, S. (2012). The Political Economy of Deforestation in the Tropics. *Quarterly Journal of Economics*, 127(4):1707 – 1754.

Erb, M. and Sulistiyanto, P. (2009). *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*. Institute of Southeast Asian Studies.

Freedom House (2012). Countries at Crossroads: Indonesia. Report.

Knutsen, C. H. (2011). Which Democracies Prosper? Electoral Rules, Form of Government and Economic Growth. *Electoral Studies*, 30(1):83–90.

KPPOD (2008). Local Economic Governance in Indonesia: A Survey in 243 Regencies and Cities in Indonesia. Report, Komite Pemantauan Pelaksanaan Otonomi Daerah (Regional Autonomy Watch).

Martinez-Bravo, M. (2014). The Role of Local Officials in New Democracies: Evidence from Indonesia. *The American Economic Review*, 104(4):1–14.

Mboi, A. B. (2009). Pilkada Langsung: The First Step on the Road to a Dualistic Provincial and District Government. In Erb, M. and Sulistiyanto, P., editors, *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*, pages 38–52. Institute of Southeast Asian Studies.

McCulloch, N. (2011). The Indonesian Sub-National Growth and Governance Dataset. Data.

McCulloch, N. and Malesky, E. (2010). Does Better Local Governance Improve District GrowthPerformance in Indonesia? IDS Working Papers 369, Institute of Development Studies.

Mietzner, M. (2011). Funding Pilkada: Illegal Campaign Financing in Indonesia's Local Elections. In Aspinall, E. and van Klinken, G., editors, *The State and Illegality in Indonesia*, pages 123–138. KITLV Press.

Niazi, T. H. (2012). Decentralization. In Hill, H., Ehsan Khan, M., and Zhuang, J., editors, *Diagnosing the Indonesian Economy: Toward Inclusive and Green Growth*, pages 389–427. Anthem Press and Asian Development Bank.

Olken, B. A. (2007). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy*, 115(2):200–249.

Papaioannou, E. and Siourounis, G. (2008). Democratisation and Growth. *The Economic Journal*, 118(532):1520–1551.

Persson, T. (2005). Forms of Democracy, Policy and Economic Development. NBER Working Papers 11171, National Bureau of Economic Research.

Persson, T., Roland, G., and Tabellini, G. (2000). Comparative Politics and Public Finance. *Journal of Political Economy*, 108(6):1121–1161.

Persson, T. and Tabellini, G. (2006). Democracy and Development: The Devil in the Details. *American Economic Review*, 96(2):319–324.

Persson, T. and Tabellini, G. E. (2003). *The Economic Effects of Constitutions*. MIT Press, Cambridge, Mass., 1. ed. edition.

Praktikno (2009). Political Parties in Pilkada: Some Problems for Democratic Consolidation. In Erb, M. and Sulistiyanto, P., editors, *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*, pages 53–73. Institute of Southeast Asian Studies.

Rodrik, D. and Wacziarg, R. (2005). Do Democratic Transitions Produce Bad Economic Outcomes? *American Economic Review Papers and Proceedings*, 95(2):50–55.

Schiller, J. (2009). Electing District Heads in Indonesia: Democratic Deepening or Elite Entrenchment? In Erb, M. and Sulistiyanto, P., editors, *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*, pages 147–173. Institute of Southeast Asian Studies.

Sjahrir, B., Kis-Katos, K., and Schulze, G. (2013). Political Budget Cycles in Indonesia at the District Level. *Economics Letters*, 120(2):342–345.

Skoufias, E., Narayan, A., Dasgupta, B., and Kaiser, K. (2014). Electoral Accountability and Local Government Spending in Indonesia. Policy Research Working Papers 6782, The World Bank.

Sulistiyanto, P. and Erb, M. (2009). Indonesia and the Quest for Democracy. In Erb, M. and Sulistiyanto, P., editors, *Deepening Democracy in Indonesia? Direct Elections for Local Leaders (Pilkada)*, pages 1–37. Institute of Southeast Asian Studies.

The World Bank (2008). *Spending for Development : Making the Most of Indonesia's New Opportunities*. The World Bank.

The World Bank (2009). Indonesia Development Policy Review: Enhancing Government Effectiveness in a Democratic and Decentralized Indonesia. World Bank Report 53451, The World Bank.

Valsecchi, M. (2013). Local Elections and Corruption during Democratization: Evidence from Indonesia. Mimeo, Department of Economics, University of Gothenburg.

Voigt, S. (2011). Positive Constitutional Economics II— a Survey of Recent Developments. *Public Choice*, 146(1-2):205–256.

von Luebke, C. (2009). The Political Economy of Local Governance: Findings from an Indonesian Field Study. *Bulletin of Indonesian Economic Studies*, 45(2):201 – 230.

Appendix A. Additional Information and Results

A.0.1 Specification of data sources

1. Data on local elections
    - Indonesian Ministry of Home Affairs
2. GRDP data
    - Statistics Indonesia, Badan Pusat Statistik (BPS)
        - Gross Regional Domestic Product of Regencies/Municipalities in Indonesia 2006-2010
        - Gross Regional Domestic Product of Regencies/Municipalities in Indonesia 2003-2007
3. Data on when new districts are formed
    - Indonesian Ministry of Home Affairs
4. Conversion of districts between different years
    - Statistics Indonesia, Badan Pusat Statistik (BPS)
        - Master File Kabupaten 1993-2002
        - Master File Kabupaten 2003-2009
    - Downloaded 2012-04-20 from www.bps.go.id

Table A.1: Comparison of district characteristics in treated and control districts.

	Treatment group		Control group		
Variable	Mean	Sd	Mean	Sd	P-value*
General characteristics					
Dummy for split (own source)	0.298	0.042	0.198	0.043	0.1058
Distance district to province's capital	126	12	101	10	0.1309
Distance district to the capital Jakarta	1015	60	825	65	0.0353
Total district revenue, billion (SIKD)	239	16	231	13	0.7236
Total district revenue per capita, billion	594347	55919	485100	35693	0.1353
Social characteristics					
Population (BPS)	607015	45864	715680	67897	0.1702
Population (Susenas)	597070	45123	709827	67437	0.1502
Poverty headcount (BPS)	21.255	1.017	20.187	1.181	0.4953
Real average annual per capita expenditure (Susenas)	1917723	46532	1917091	59656	0.9932
Urbanization (Susenas)	0.385	0.030	0.464	0.036	0.0934
People in primary school age 7-12 years (Susenas)	74476	5545	90424	9255	0.1196
People in primary school age 7-12 year (share of population)	0.129	0.002	0.127	0.002	0.5591
Share of people ever being in primary school per total population (Susenas)	0.445	0.008	0.449	0.010	0.7521
Unemployment rate (Susenas)	0.039	0.002	0.046	0.003	0.0481
Ethno-linguistic Fragmentation Index, 2000 (Census)	0.417	0.032	0.381	0.035	0.4540
Religion Fragmentation Index, 2000 (Census)	0.204	0.021	0.141	0.018	0.0294
Number of Telephone Subscribers, 2000 (PODES)	7597	1422	12699	2640	0.0710
Telephone access per household, 2000 (PODES)	0.063	0.008	0.083	0.010	0.1168
Road Quality: 1 good - 4 worst, 2000 (PODES)	1.447	0.033	1.435	0.041	0.8219
Number of villages with asphalt roads, 2000 (PODES)	109	6.874	108	9.079	0.9273
Governance characteristics					
Number of corruption cases covered by media, 2004 (ICW)	1.605	0.128	1.618	0.146	0.9492

Continued on next page

Table A.1 – *Continued from previous page*

Variable	Treatment group Mean	Sd	Control group Mean	Sd	P-value*
KPPOD score: Institution, 2002 (KPPOD)	0.052	0.002	0.053	0.002	0.7736
KPPOD score: Social, 2002 (KPPOD)	0.055	0.003	0.054	0.003	0.8224
KPPOD score: Economic, 2002 (KPPOD)	0.034	0.002	0.028	0.002	0.0273
KPPOD score: Labor, 2002 (KPPOD)	0.029	0.002	0.027	0.002	0.5244
KPPOD score: Infrastructure, 2002 (KPPOD)	0.027	0.001	0.028	0.001	0.5750
Economic characteristics					
Real income, GRDP, billion (BPS)	3460	536	3530	438	0.9277
Real income, GRDP, without oil & gas, billion (BPS)	3040	470	3450	433	0.5410
Real income, GRDP, 2003, billion (Own source)	3759	575	3870	480	0.8888
Real income per capita, GRDP, 2003, thousand (Own source)	5784	562	5832	822	0.9609
Sectoral breakdown of GRDP, billion (BPS):					
- Agriculture	714	60	681	73	0.7264
- Mining, Quarrying, Oil & Gas Manufacturing	529	256	154	59	0.2450
- Non Oil & Gas Manufacturing	696	172	1070	266	0.2150
- Electricity, Gas & Water Supply	38	10	53	14	0.3702
- Construction	172	39	142	17	0.5308
- Trade, Restaurant & Hotel	654	144	753	105	0.6091
- Transportation and Communication	216	47	228	38	0.8450
- Financial Services	153	35	145	19	0.8545
- Services	304	38	316	33	0.8163
Sectoral breakdown of GRDP (BPS):					
- Share of agriculture to total GRDP	0.326	0.017	0.300	0.022	0.3231
- Share of mining to total GRDP	0.054	0.013	0.048	0.013	0.7466
- Share of non oil & gas manufacturing to total GRDP	0.142	0.011	0.162	0.018	0.3390
- Share of electricity to total GRDP	0.008	0.001	0.010	0.001	0.0816
- Share of construction to total GRDP	0.054	0.003	0.049	0.003	0.2348
- Share of trade to total GRDP	0.172	0.007	0.199	0.009	0.0171

Continued on next page

Table A.1 – *Continued from previous page*

Variable	Treatment group Mean	Treatment group Sd	Control group Mean	Control group Sd	P-value*
- Share of transportation to total GRDP	0.067	0.005	0.069	0.006	0.8785
- Share of financial service to total GRDP	0.046	0.003	0.047	0.003	0.7782
Value of FDI Realization, 2003, million US$ (Bkpm)	142	19.169	340	46.163	0.0000
- as percentage of current district GDP	0.006	0.001	0.011	0.002	0.0128
Value of domestic direct investment realization, 2003, billion (Bkpm)	436	57.555	894	122.259	0.0003
- as share of current district GDP	0.141	0.019	0.206	0.031	0.0646

Note: * for equality across treatment and control group. A specific variable refers to the value 2001 unless otherwise stated. The number of observation is between 94 and 292. Source: McCulloch (2011). Abbreviations within parentheses after each variable identify other specific sources.

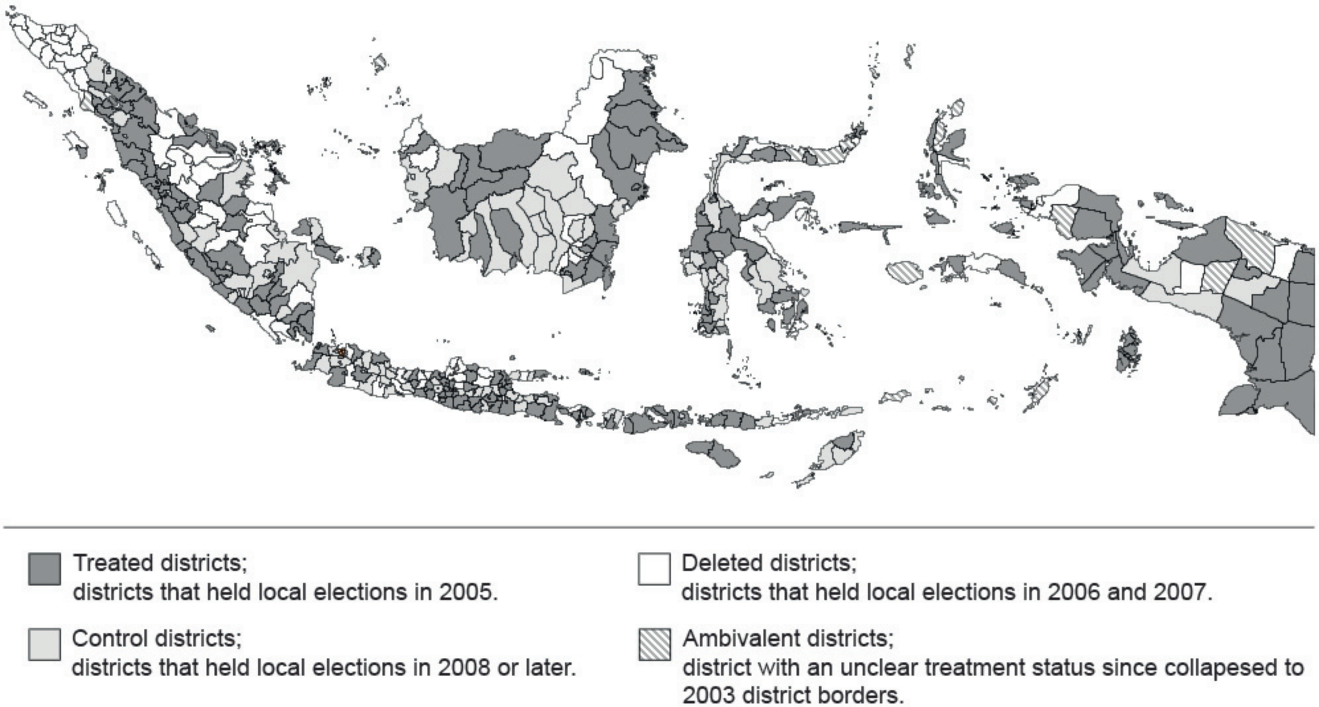


Figure A.1: The distribution of treatment and control districts in Indonesia

Paper III

Explorative Analysis of Municipality-level Inequality in Late Nineteenth Century Sweden

Abstract

This paper presents new data on Sweden's municipality-level inequality for 1871 and 1892. There exists scant information on the evolution of inequality during the 19th century, but thanks to a novel method of obtaining data, this paper shows that the income share of the top one percent in the industrial sector is increasing between the two time points, whereas inequality in the agricultural sector is fairly constant. In addition, the inequality estimates presented here can be used for further studies. The data is not on income but on the vote distribution for municipality elections. The taxation numbers can be backed out since a tax-weighted electoral system was in place.

Key Words: Income Inequality, Economic Development

JEL Classification: N33, D31, O15

1 Introduction

The evolution of inequality over the path of development has gained plenty of interest. Data on inequality is scant from the 19th century, especially data on income inequality. This paper presents new income inequality estimates for 1871 and 1892 for 2300 Swedish municipalities. The major contribution of the paper is to provide a new piece to the puzzle of long-run evolutions in inequality. The paper shows that the top one percent income share in the industrial sector increases between 1871 and 1892. Hence, income inequality increases in the early phase of the Swedish industrialization process. The other contribution of the paper is to provide new inequality estimates for further research.

The high data demands for providing inequality measures constrain research on inequality. Two main data sources are currently in use. Firstly, cross-country survey data exists from the 1970s. The cross-country literature answers the question of the relationship between growth and income inequality as "it depends" (Neves et al., 2016) or "there is a lot of heterogeneity" (Ravallion, 2001). However, survey data generally misses the main driving force behind changes in inequality, namely, changes in top incomes (Atkinson et al., 2011). Top incomes are generally missed because it is rare to select a super-rich person randomly.[1] Secondly, taxation data exists since around the turn of the 20th century for the nowadays developed countries. This data source forms the basis for top income inequality measures. The top income literature documents that income inequality exhibits a u-shaped pattern: at the start of the 20th century to WW1 inequality was high, it decreased heavily from the WW1 to the WW2, continued to decrease to the 1970s and has since started to increase for some countries (for good overviews, see Piketty 2007 and Atkinson et al. 2011). However, the developed countries are a small selected sample, which mostly provides descriptive evidence that is hard to generalize. Regardless of the data source used, there is scant evidence of the evolution of inequality prior to the turn of the 20th century. Such data is valuable as changes in inequality are long-run processes.

Roine and Waldenström (2008) show that Sweden exhibits a similar u-shaped pat-

[1] In addition, authors think that the modern-day survey data on inequality is plagued by measurement errors (e.g. Ravallion 2001 and Easterly 2007). For more complete overviews of data complications, see Cowell and Flachaire (2015) and Cowell and Kerm (2015). Top incomes also seem to be the driving force empirically. Leigh (2007) finds that Gini coefficients from surveys and top income shares correlate to a high degree.

tern as the other nowadays developed countries: at the onset of the 20th century there was a higher level of top income inequality; it dropped until the 1950s and has increased since the 1980s. They also show that top incomes in early 20th century Sweden were at similar levels as for other developed countries. In a study on the 19th century, Modalsli (2018) finds that income inequality in Norway in 1868 was on a similar level to the other few countries where such early estimates exist (Russia, UK, and the US).

Instead of income inequality, wealth inequality is usually measured when going further back in time. One major reason for this approach is that agriculture is the primary sector – agricultural land is an important asset and income from agriculture is usually undocumented. Roine and Waldenström (2009) suggest that Swedish top wealth inequality was fairly stable between data points of the 1870s and 1900s. Then, they show that from 1900, wealth inequality gradually decreased until the 1970s, and from the 1980s there is an upwards trend (but not back to the earlier level). Bengtsson et al. (2018) extend the investigation to 1750-1900. They find that wealth inequality was lower in 1750 than in the US, Great Britain, and France, but Sweden in 1900 was just as unequal as the other countries due to a period of inequality growth since the 1850s. One possible explanation for this pattern could be that Sweden's industrial take-off took place around 1870-1890.

This paper shows that income inequality in the industrial sector increases for municipalities between 1871 and 1892 (municipalities defined as towns are excluded from the sample). The increase in the top one percent income share for the median municipality is at a similar level as the increases for Sweden, the UK, Australia, and Canada from 1980 to 2000, which generally is considered to be substantial. The paper corroborates increases in inequality as found by Bengtsson et al. (2018) by using a different data source and method. In combination with the evidence of income inequality decreasing since 1900 from Roine and Waldenström (2008), this suggests cyclical movements in income inequality in Sweden.

I am able to examine inequality in the 19th century by using a novel method of obtaining income data. Data on the vote distribution for municipality elections form the basis for the estimation of income inequality measures. A tax-weighted electoral system was in place at this time: the votes one person got to cast in municipality elections were dependent on the amount of tax paid by the same person. Thus, the taxation sums can be backed out from the vote data. The data has not been used before to investigate income inequality. To my knowledge, this paper is also the first one that separates income inequality to the agricultural and industrial sector by first classifying people into either sector. The paper provides support for that the two distributions are separate; a person was either involved in agriculture or industry.

The panel dataset used in this paper is a mix between the two common types of data sources. One the one hand, the inequality measure is the top one percent income share, in line with the top income literature. Atkinson et al. (2011) stress that the top one percent and not the top 10 percent drive changes in inequality. On the other hand, a sample of units, the municipalities, are used, in line with the cross-country literature. Both the top income and the cross-country literature struggle with comparability since taxation systems and survey methods change over time and countries. One benefit of the current dataset is that the taxation system is the same, both over units and time.

The municipality-level inequality measures presented in this paper are available online. Piketty (2007) acknowledges that the top income literature's focus on national-level inequality does not generate data which allow for credible identification strategies. Hopefully, the estimates presented in this study are useful for further studies on the effect of inequality on other municipality-level variables.

The outline of the paper is as follows: Section 2 presents the institutional context in which the data was generated to judge if the inequality measures are trustworthy. Sections 4, 5 and 6 describe the data, method, and sample. Sections 7 discuss the empirical results, and section 8 concludes.

2 The Institutional Context

The data used in this paper originate from an old taxation system and, to my knowledge, nobody has explored if the numbers reflect true income or wealth. This section evaluates the context in which the data was generated. The next section shows that we, at large, can approximate people's incomes. Appendix B presents additional information on the institutional context.

The tax system investigated here (*statlig bevillning enligt II artikelen*) represents the seed of what would become the modern Swedish taxation system, and ran in parallel with an older system (*grundskatter*) during the late 19th century. The taxation system under investigation was established in 1862 and should by law reflect people's true income and wealth.[2] In contrast, under the older system, taxes did not reflect underlying true income or wealth during the 19th century, as described by Olsson (2005). For instance, the most important taxes on agricultural property in the old system was to provide for a soldier (*rust- och roteringsbesvär/ indelningsverket*), and taxation lists had not been updated for the last 400 years at this time. The older taxation system was phased out by the turn of the century, as decided in 1892.[3] The year 1903 is usually mentioned as the date of introduction of the Swedish income tax, but the only institutional difference to the tax system under investigation is that income tax returns were introduced.

Table 1: Translation key between votes and the amount of taxes paid (for countryside municipalities, not towns)

Tax paid for agricultural property (1)	Tax paid for others (2, 3)	Votes
1-5 öre	1-10 öre	1
5-10 öre	10-20 öre	2
10-15 öre	20-30 öre	3
...	...	...

Note: 1 SEK=100 öre

I have data on the municipality-level vote distribution and back out the taxation numbers, which is feasible since a taxation-weighted electoral system was in place. For each dollar paid in tax, a person got a specific number of votes. From 1863, the governmental-level tax (*statlig bevillning enligt II artikelen*) gave votes (*fyrkar*) for municipalities in the countryside, according to the translation key in Table 1. The tax constituted three main types: (1) tax on agricultural property, (2) tax on other property and (3) tax on income of labor and capital. The agricultural property tax (1) yielded the double number of votes as the tax paid for other property and other income (2, 3).[4] The taxes are levied on all juridical persons, including firms and other organizations, and consequently, all juridical persons that taxed a certain amount had votes. Women were also juridical persons as long as they were unmarried.[5] Most people did not pay a sufficient amount of the specific taxes needed to receive any vote, and for convenience, I refer to juridical subjects holding one

[2]Nordisk Familjeordbok 1899, Bevillnings-beredning; Nordisk Familjeordbok 1899, Taxeringsvärde; SFS 1861:34, §2

[3]This picture is a simplification. For good overviews of the taxation system as a whole see Olsson (2005) and Gårestad (1987).

[4]SFS 1861:34, §6, §7; SFS 1863:50, §58; Velander 1901, p. 189

[5]SFS 1862:13, §12, §17; SFS 1862:14, §12; Aldén 1901, pp. 227-8

or more votes as vote owners.

A form of "direct taxation-weighted democracy" was implemented; vote owners cast votes directly in the municipality parliament (*kommunalstämma/ allmän rådstuga*) and did not elect representatives.[6] To decide the governmental-level tax, the municipality parliament elected a representative to a committee charged with the task of deciding how much each inhabitant would tax.[7] The committee normally included representatives from 2-4 municipalities and an appointee from the provincial-level government. Instead of income tax returns, inhabitants could give information of his/her property wealth and income to the committee and, besides this, "collected information" is stated to be the primary source in the laws. In turn, the votes functioned as the base for the municipality tax and the municipality parliament decided the municipality tax rate.[8] A vote owner could have incentives to elect the representative both to increase his/her taxation to have more political influence, but he/she could, of course, also want to pay lower taxes. However, checks and balances existed. When the committee had decided the taxation figures for each inhabitant, the municipality parliament translated the taxation lists to electoral lists. Both the tax and electoral lists were publicly available and, by law, the electoral lists "were read aloud from the church pulpit". The inhabitants had the opportunity to complain to a provincial-level instance to rectify the taxation lists. The municipalities of this time were small – the median population of the countryside municipalities was around 1150 inhabitants in 1871 and 1892. It was most likely hard to hide property wealth or income from other inhabitants.

I have found but one qualitative judgment regarding if the governmental-level taxes reflect true income and property wealth. Flodström states, 'It is assumed that agricultural property is often not estimated to its true value. However, the greatest misjudgment in this regard should be regarding Norlandic firm property...' (Authors translation, Flodström, 1906, p. 7).[9] I believe it would have left more

[6]However, the municipality parliament could choose to elect representatives and yield most of their decision powers to the elected parliament (*kommunalfullmäktige/ stadsfullmäktige*), but very few did so. In the elected parliament one counted one person one vote. Town-municipalities with more than 3000 inhabitants had to elect a parliament, but they do not constitute a part of the population of this paper.

[7]SFS 1861:34, §2; Nordisk Familjeordbok 1899, Bevillnings-beredning; Nordisk Familjeordbok 1899, Taxeringsvärde

[8]SFS 1861:13 §59; SFS 1862:13, §63, §65

[9]In Swedish: 'Det antagas visserligen, att jordbruksfastighet ofta icke är uppsakttad till sitt verkliga värde. De största oegentligheterna i detta afseende torde dock gälla norrländsk bolagsegen-

written records, if people perceived the taxes as unfair.

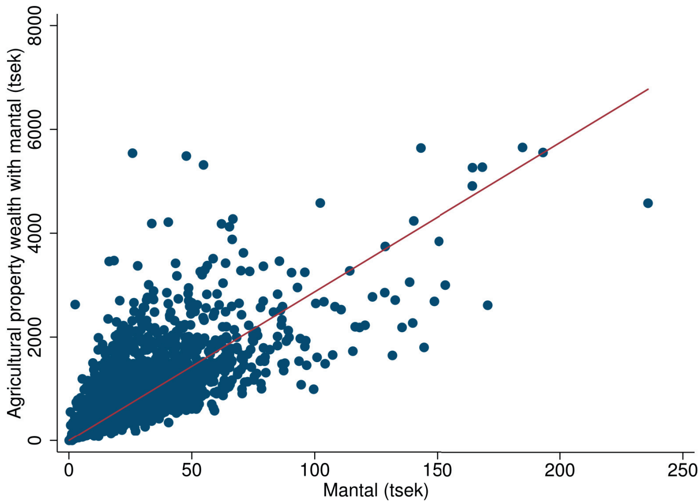


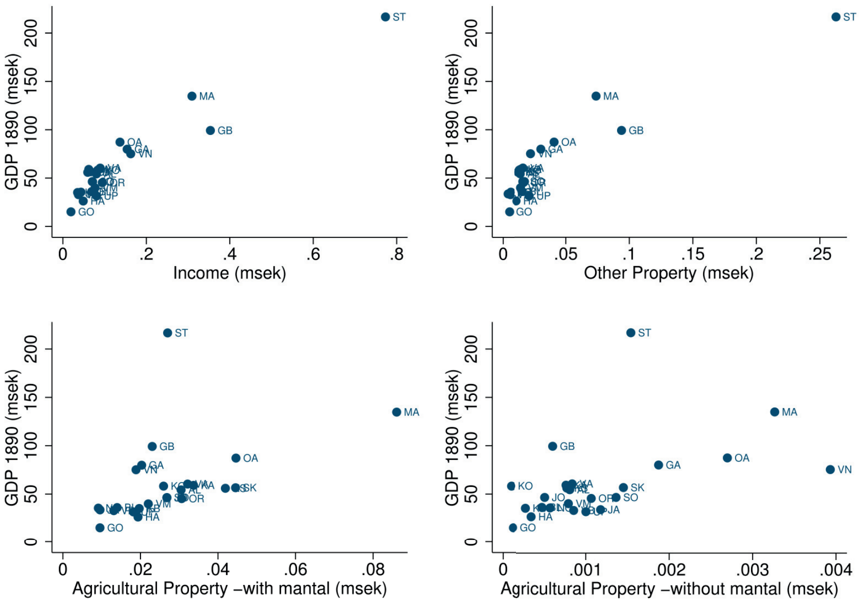
Figure 1: Agricultural property tax arising from farms with mantal versus mantal
Source: BiSOS N and R 1892.

One way to examine if the vote data reflect the true agricultural values produced is to investigate if the taxation values changed compared to the older taxation system. As previously mentioned, the older taxation system did not reflect true agricultural property wealth. Hence, if the taxation were updated from the older to the newer system, we would not see a perfect linear correspondence between the systems. In the older taxation system, agricultural taxation values were given in the unit *mantal*. Figure 1 shows the sum of mantal against the sum of agricultural property wealth for each municipality. Only farms that were with mantal are included in the sum of agricultural property wealth; farms which were not taxed in the old system (without mantal) are not included. Some correlation is expected, as municipalities with good farmland likely score high on both measures, but a "too good" fit would indicate that the newer system simply followed the old one. The figure shows that agricultural taxation was updated from the old to the new system; there is not a perfect linear correspondence.

Another way of verifying if the vote data reflect the true values produced is to com-

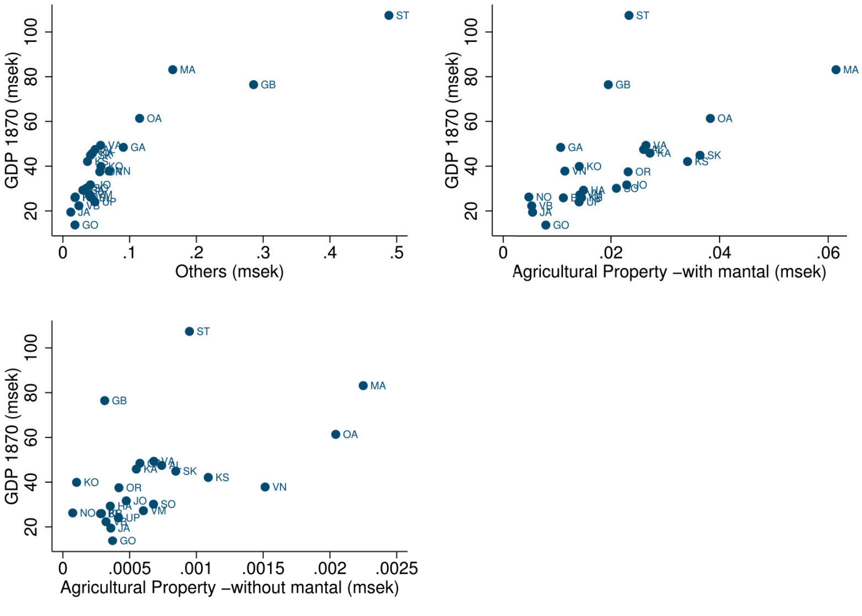
dom...'

pare to GDP. Figure 2 and 3 compare the various taxation bases to GDP estimates at the provincial (*län*) level provided by Enflo et al. (2014). The figures display that the various taxation bases are strongly positively correlated with GDP, strengthening our confidence in that the taxation system under investigation reflects true income and property values.



Note: A linear regression gives the following prediction GDP=14 + 175*Income + 161*Other property + 522*Agricultural property with mantal + 5121*Agricultural property without mantal, with an R2 of 97 percent. The coefficients of the Income and Other property variables should not be interpreted in isolation due to the correlation of 99 percent between them.
Source: BiSOS R 1892, BiSOS U 1892 and Enflo et al. (2014).

Figure 2: GDP versus tax bases 1892



Note: The variable Others refer to Income and Other property combined. A linear regression gives the following prediction GDP=145+ 163*Others + 662*Agricultural property with mantal + 932*Agricultural property without mantal, with an R2 of 95 percent. The coefficients of the Agricultural property variables should not be interpreted in isolation due to the correlation of 77 percent between them.
Source: BiSOS R 1871, BiSOS U 1874 and Enflo et al. (2014).

Figure 3: GDP versus tax bases 1871

3 Interpretation: The Relationship Between Votes and the Modern Day Income Concept

It is notoriously hard to find income data for agricultural societies since agricultural production tends to be non-monetized. This paper uses agricultural property wealth to approximate agricultural incomes. To even have a taxation system based on individuals' property values is rare for earlier periods. Agricultural wealth is a reasonable approximation of agricultural income for two reasons.

Firstly, persons living during the end of the 19th century likely viewed the agricultural votes as being based on what we today would call income. People that were tenants at an agricultural property did not pay governmental-level tax for the land they used, but they had votes in the municipality parliament after the property value they would have paid tax for if they owned the land. Tenants not only got votes for the property value they used, but they also paid municipality tax based on it.[10] [11] Furthermore, the agricultural property is valued before deduction for costs, such as rents on debts.[12] [13]

Secondly, agricultural wealth is proportional to agricultural incomes under the assumption of perfect markets. When comparing owners to tenants, we are rediscovering the problematic aspect of inferring the value of the imputed rent of owned property. A person's income in period t is equivalent to the value of consumption and the net change in wealth between the beginning and end of the period (Simons, 1938). The owner of a farm incurs specific costs for using the farm (the net change in wealth, the imputed rent) whereas the tenant pays a tenant fee. The net income is equivalent if one assumes perfect markets.[14] For example, if the tenant fee was higher than the imputed rent, a potential tenant could borrow capital and buy a farm. This paper investigates long-run changes in inequality, for time points 20 years apart, and the assumption is likely feasible on this time scale.

In the data on the vote distribution, the taxation on other property and other income (2, 3) are collapsed to one category, which I label industrial throughout the paper. On the aggregate level, the overwhelming majority of the combined industrial category came from the income tax. For 1892, in the median municipality, 86 percent of the industrial tax came from income. For the municipality at the fifth

[10] This is not as strange as it appears at first sight: In my opinion, the concept of ownership underwent a transformation during the 19th century in Sweden. The two tenure categories crown tenants (*kronobönder*) and freeholders (*skattebönder*) paid equivalent amounts of tax in the older in-kind system, but for the first category it was called "tenant fee" and the other one "tax". However, for the noble tenants (*frälsebönder*) the "tax" in the in-kind system was a proper and varying "tenant fee" (Olsson, 2005).

[11] SFS 1862:13, §10, §57; SFS 1861:34, §2; Aldén 1901, p. 227

[12] The taxes (1,2,3) were all "before" deduction for costs (*brutto och inte netto*), in contrast to today (Flodström, 1906, p. 7, 10).

[13] Flodström (1906, p. 7, 10) report that, for 1903, 40 percent of non-incorporated agricultural property is used as collateral for debt.

[14] From today's tax registry data, the imputed rent (f.ex. the income derived from owning an apartment) is often not taken into account when measuring income since it is hard to observe. For our case, the measures include imputed rent.

percentile, 60 percent still came from income.[15]

I assume that the logic behind both the assignment of votes and the presentation of data is that a person either was a farmer that paid agricultural property tax or a person employed in the secondary sector on the countryside that paid taxes on other property and income. In 1871, the median of the percentage of vote owners that only had votes from either source, industry or agriculture, is 94 percent.[16] For both 1871 and 1892, the median of the total individual vote that comes from industrial sources for the richest agricultural vote owners is only 2-3 percent.[17]

4 Data

This paper uses data on the municipality vote distribution in 1871 and 1892 published by Statistics Sweden (Statistics Sweden, 1874, 1895). The inequality estimates presented in this paper and the source data is freely available at https://doi.org/10.5878/cw7b-g897. In addition to the vote data, the paper uses municipality-level taxation data from Statistics Sweden (1894a) and Statistics Sweden (1894b).

Parliamentarism was introduced at the national level in 1866, and vote rights were thoroughly discussed in Sweden in the decades that followed. The 1892 publication was explicitly conducted due to discussions on legislating limits on the maximum number of votes one vote owner could have and not allowing companies etc. to hold votes. The publication of 1892 used the publication of 1871 as a template. The statisticians behind both publications noted that the vote lists the municipalities sent in included errors, such as the division of votes arising due to the various taxation types was incorrectly done or that people were given votes on the wrong basis. The laws regulating taxes and votes included standardized forms that should be followed, so the statistician's complaints were to some degree a reflection of the all too human nature of not filling in forms correctly. The vote lists appear to have been properly checked by the statisticians. The statisticians behind the 1871 publication corrected the material, most likely by cross-checking with the taxation lists. The statisticians behind the 1892 publication instead communicated with the municipalities to correct the vote lists. Thus, the publications to some degree

[15]The information needed for those calculations is only given for 1892.

[16]The information needed for this calculation is only given for 1871.

[17]The figure is based one the municipalities where the category to which the richest vote owner belongs to can be discern (roughly 50 percent of the cases in 1871 and 80 percent in 1892).

reflect how the vote distribution should have been if the laws had been followed, not how it was in reality. Tyrefors et al. (2017) and Andersson and Berger (2018) use the same data to measure political elite concentration. Tyrefors et al. (2017) choose not to use the 1871 publication since they consider the data being of bad quality to measure the realized vote distribution, but the data is sufficiently good for constructing measures on income.

Censored (tabulated) data forms the basis for the new income inequality measures, which is common in the top income literature, especially for older time periods. For each municipality, there are three different sets of interval data as illustrated in Table 2. The first two sets consist of data on the number of predominantly agricultural (industrial) vote owners that have votes above $\{0, 2, 5, 10, 20, 50\}$ percent of the total sum of votes in each municipality (for 1871, the interval limit is 25 percent instead of 20). If a person's majority share of votes comes from agricultural (industrial) taxation, he/she is placed in the agricultural (industrial) interval but votes from the industrial (agricultural) tax base also contribute to the interval placement. This can be regarded as a mild measurement error since, as previously discussed, persons predominantly belong to either the agricultural or industrial sector. The third set of intervals consists of information on the number of vote owners that, for votes arising from both taxation bases, have votes above $\{1, 5, 10, 25, 50, 100, 250, 500, 1000\}$ votes. Since we know the total sum of votes in each municipality, we can recalculate the first two sets of intervals to absolute numbers and overlay it with the third set of intervals. There exists information on the endpoint of the intervals. More specifically, we know the highest number of votes received by one person from both types of taxation and from agricultural property taxation alone. The majority of the population do not pay any tax, do not have any votes and are not placed in any of the interval sets.

Table 2: Example of input data- Giresta 1892

First two sets of intervals:									
Interval limit (percent of total sum of votes):	0	2	5	10	20	50			
Vote owners - predominantly agriculture	21	5	0	2	1	0			
Vote owners - predominantly industry	6	2	1	0	0	0			
Third set of intervals:									
Interval limit (votes):	1	5	10	25	50	100	250	500	1000
Vote owners	4	1	9	8	5	8	0	3	0
Other information:									
Total agricultural votes	3350								
Total industrial votes	1122								
Highest vote in general	995								
Highest vote due to agriculture	724								
Population	603								

Note: The table displays information for the 1892 "mean" municipality (the municipality closes to the center of the input variables shown in the table). Example of interpretation: 21 vote owners with votes predominantly from agriculture have votes between 0 and 2 percent of the total sum of votes; 4 vote owners have between 1 and 5 votes.

The 1892 data allow for dividing the input data to vote owners that are actual persons, limited holding companies, companies, or congregations, but the 1871 data do not. Most vote owners were actual persons: 81 percent of the total agricultural votes and 69 percent of the industrial votes occurred to actual persons. In only 15 percent of the municipalities, the juridical subject holding the highest number of votes were limited holding companies, and in the robustness checks, such municipalities are excluded.

I have matched the municipalities from the various publications by string matching: In general, the municipality name, the name of the sub-region (*härad*), and the province name (*län*) are used to match datasets by finding names that match approximately. After this step, I matched the few remaining municipalities by hand with the help of information provided by the statisticians in the source publications or the municipality name in itself.[18] In the legal formalization of the local governance structure, applied since 1865, each parish (*socken*) was going to constitute one municipality, but some exceptions arose.[19] Until the beginning of the 1950s, the changes and relationships between parishes/municipalities are mostly undocumented. The best source available comes from the Swedish National Archive's internal archival system (NAD). I have made this accessible online in a more suitable format under the name "Kommungränskonverteraren- Beta". Unfortunately, the number of municipalities change from one statistical publication to another also within years, which is not reflected in NAD, and hence I prefer to use string matching in this paper.

5 Method

Generally, the top income literature estimates national-level top income shares from tabulated income data and national income figures. Due to the national level focus, the end goal in other papers has often been to make inequality measures comparable over time and space. In this paper, the tax system under investigation stays the same on both dimensions, allowing us to circumvent comparability problems to a large degree.

[18]The matching of names is straightforward, for example, in one publication, the municipality "Bro and Lossa" appears, and in another, the municipalities "Bro" and "Lossa"appears in the same sub-region.

[19]SFS 1862:13, §1, §2, §3; SFS 1862:14, §1; Aldén 1901, p. 227

My set-up differs to the top income literature in one important aspect: I have data on the sub-national level. I calculate the top one percent income share, both for the agricultural and the industrial sector. Other researchers calculate top income shares based on larger populations and commonly calculated measures such as the 0.1 and 0.01 percent income share, but in my case, where I rely on smaller populations, such measures are conceptually challenging to interpret. The 0.1 percent income share for the median municipality in my sample is equivalent to the income of 1.15 persons. As an alternative measure, I choose to estimate the top one person income share, e.g. the income share of the richest individual. More specifically, I calculate the top one person share only for the agricultural sector due to data limitations.

My set-up also differs in another aspect: various assumptions typically need to be made, for example, on the distribution of individuals within an income interval, and papers generally show that their inequality estimates are robust to alternative assumptions. Instead, I show in Appendix C that the ordinal properties of my top one percent inequality measure are robust to varying such assumptions, but the cardinal properties are not. Since the taxation system is the same over time and space in the current sample, the ability to use the cardinal properties are less needed. The estimates allow for answering the question "Is municipality A more unequal than municipality B?", that is, using the ordinal properties, but not answering "How much more unequal is municipality A compared to B?", that is, using the cardinal properties. The estimates available online is on the ordinal scale.

The main analysis of this paper still uses the cardinal properties but rely on analyzing changes within municipalities. To analyze how inequality changes between 1871 and 1892, I rely on sign tests of the median difference.[20] The difference of a variable x is $x\Delta_i = x_{1892,i} - x_{1871,i}$ for a municipality i. Thus, the analysis relies on comparing within municipalities, but instead of estimating means, I choose to estimate medians. A median is a more transparent summary statistic than a mean in the face of outliers. The results presented are robust to alternative assumptions. For describing patterns in inequality between sectors and the two years, the analysis relies on Spearman rank correlation coefficients, which compare the agreement between an ordinal sorting of variables, since the ordinal properties of the inequality measures are robust. Spearman's rank coefficient is -1 if two variables rank the

[20]Sometimes called paired sample sign test. To be more precise, since the sample size is large, I choose to present the version of the test with normal approximation and adherent continuity correction.

municipalities completely opposite to each other and 1 if two variables rank the municipalities in perfect agreement. The standard Person correlation coefficients rely on the cardinal properties and are provided only for comparability reasons.

5.1 Inequality measures

The top one percent income share is calculated by dividing the income of the top one percent richest with the total income. Formally:

$$Top\ one\ percent\ income\ share_i = \frac{Income\ of\ the\ top\ one\ percent\ richest_i}{Total\ income_i}$$

for the industrial or agricultural sector i in a municipality. The top one person income share in the agricultural sector is achieved merely by exchanging the numerator to the income of the richest person, which is given in the data, and will not be elaborated on further.

Total taxpaying population Information on the total taxpaying population is needed to calculate the top one percent richest and total income. The taxpaying population denotes how many persons that pay tax under the current legal system. It is important in the top-income field when comparing different taxation system across time and space. As baseline total taxpaying population, I use the total population, including married women and children, in each municipality, which is given in the data. For 1892 where data is available, I show in Appendix C that using the actual legal taxpaying population does not affect the ranking. Instead, the baseline total population I use can be criticized for not taking into account how many persons are involved in the industrial and agricultural sector. In Appendix C I show that the total population instead can be weighted by the share of vote owners and the share of taxes in each sector without affecting the ranking.

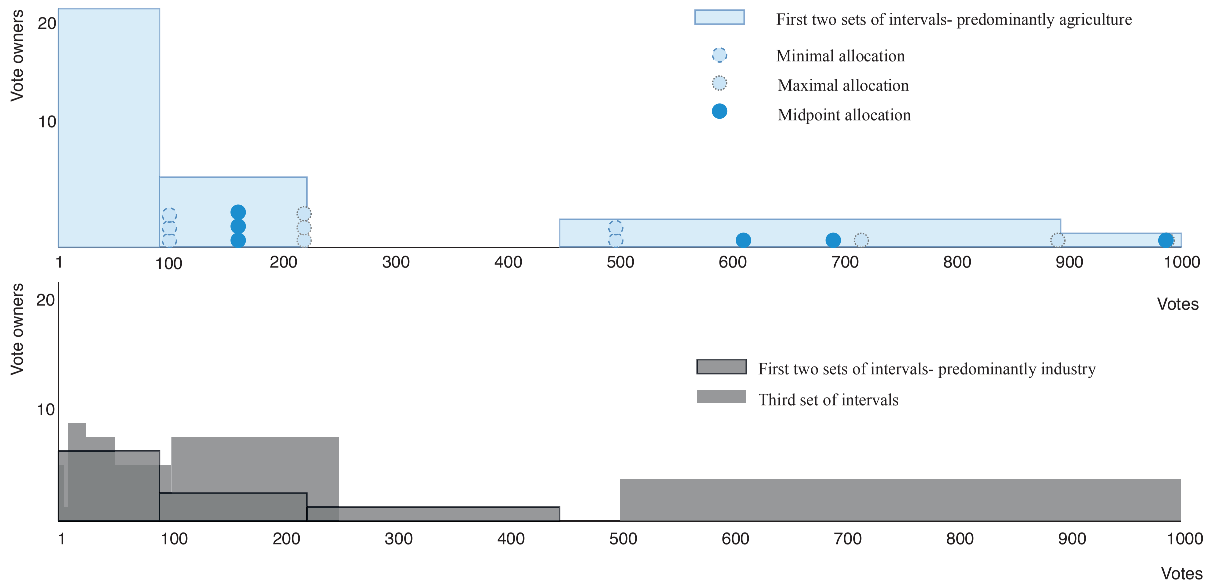
Total income Most people did not pay any taxes and thus we only have data on the top of the distribution. It is unreasonable to expect that the non-taxpaying share of the population did not produce any income. Different methods are employed to estimate the total income in the top income literature (Atkinson, 2007), that is, the sum of the income of the top one percent richest and the rest of the population. In this paper, the total income used is the agricultural (industrial) tax values plus an imputation of the incomes of the people below the taxation limit. In the data, there is information on the total number of votes distributed according to agricultural property taxation and industrial taxation for each municipality, and I obtain the taxation sums from this information. The baseline imputation is that people below the taxation limit are assumed to have 50% of the taxation limit. This translates to it taking two years of saving full labor income for the average person to buy a farm valued at the taxation limit.[21] I show in Appendix C that the imputed bottom incomes can be set to 10% or 75% of the taxation limit without affecting the ranking of municipalities much.

Income of the top one percent richest This paper uses tabulated data, commonly done in the top income literature, instead of data on individual incomes. The Pareto distribution commonly models the upper tail of a wealth or income distribution. My data differs from the usual case in some respects. Firstly, the upper interval limit is known, and no extrapolation of the highest incomes are needed. Secondly, there is no information on the mean income in an interval. However, commonly the interval limits of a tabulation are fixed, whereas my interval limits are endogenous. More specifically, my interval limits are given as a specific percentage of the taxation sum in each municipality for the first two sets of intervals. Moreover, in my data, many intervals have zero or a few individuals in them (see the example in Table 2). Thus, if I would have interpolated following the Pareto distribution, I would have allocated the top-taxed people to the beginning of each interval in a too high degree.

My preferred method is based on minimizing and maximizing the sum of the total number of votes of the top one percent using all three sets of intervals and then taking the midpoint (and it is consequently called *Midpoint* in Appendix C). The method is closely related to allocating persons to the midpoint in each interval, but I get more precision by imposing the restrictions created by all intervals. Figure 4 presents a graphical example. I formally solve one integer linear programming (ILP) problem for each sector in each municipality in each year. It

[21]The comparison implicitly relies on the following information: In 1892, a *dräng* (male agricultural worker with in-house boarding) was paid around 170 sek whereas a *piga* (female equivalent) 80 sek (Statistics Sweden, 1900, p. 45). The equivalent figures for 1871 are 110 sek and 50 sek (Statistics Sweden, 1878, p. 16).

becomes an ILP problem because the sum of vote owners in the first two sets of intervals is equivalent to the number of vote owners in the third set of intervals. The computational burden of the basic ILP problem formulation is too large to be feasible. Instead, I transform the problem to solve it sequentially for each sector by first maximizing the income of the possibly richest person, then the next richest, etc. with the equivalent result. The transformed ILP problems are solved with R's package lpSolve (version 5.6.13). Appendix section A.1 presents the transformed problem formally for finding the maximum allocation for the example in Figure 4. Missing values arise due to the requirement of at least one person representing the top one percent and due to optimization failures. I show in Appendix C that the ranking of municipalities is robust to alternative methods; interpolation by a standard Log-normal or uniform distribution using the information on the two first set of intervals only.



Note: The figure illustrates how the income of the top one percent in the agricultural sector is minimized and maximized to obtain the midpoint sum. Histograms graphically present the information available in the intervals (see Table 2). The blue bars represent the first set of intervals for predominantly agricultural vote owners. The minimal and maximal sum of the top one percent, 6 persons, in this case, is obtained while taking into account the restrictions created by the other gray intervals (and the total number of votes, which is not shown). The light blue dotes (with dotted lines) represent the allocation of the minimum attainable top one income in agriculture while taking the restrictions into account. The light blue dots (with more narrow dotted lines) represent the maximum attainable sum. The clear blue dots represent the midpoint allocation, yielding the midpoint sum of the top one percent in agriculture.

Figure 4: Graphical example of the Midpoint method for top one income in agriculture- Giresta 1892

6 Samples

Table 3: Population and samples

	Industrial		Agricultural	
	1871	1892	1871	1892
(1) According to publication	2354	2386	2354	2386
(2) Matching within same year	2339	2368	2339	2368
(3) Vote owners >1 %	1411	2091	2320	2343
(4) Matching between years	1328	1328	2222	2222

Table 3 displays the construction of the sample. The first row shows the number of municipalities in the source publications, 2354 and 2368 respectively. The second row displays how many municipalities are left after the units are matched within the same year. As an example, in the 1871 publication, sometimes population figures only exist for the aggregate municipality that spans the geographical area of municipality A and B. The publication lists them as two municipalities, but I treat them as one, which reduces the units of observations but not the representativeness. The third row shows how many observations are left in the sample after we have imposed that at least one percent of the population have to be agricultural (industrial) vote owners. The fourth row shows the reduction in the sample due to matching between years (and optimization failures). The fourth, and last, row shows that an analysis of agricultural inequality is representative, as the original sample size is reduced to 2222 municipalities, representing 94 percent of the original sample size. However, the row also shows that an analysis of industrial inequality is not representative as only 1328 municipalities are left in the sample, representing 56 percent of the original. Inspection of the table shows that the reduction happens in the third row when the restriction of the number of industrial vote owners is imposed for 1871. More precisely, going to the third row for industrial measures in 1871, the sample is reduced to 1444 municipalities, representing 59 percent of the original sample size. When using measures on industrial inequality in 1871, the sample will be selected and henceforth I will refer to the 1328 municipalities left after merging between years as the industrial subsample.

Table 4: Correlations, taking into account selection to be able to measure industrial inequality

Variable pair		n	Spearman	Pearson
indOwnershare1871	ind1892	1966	0.39*	0.30*
indOwnershare1871	agr1892	2222	0.30*	0.31*
indOwnershare1871	richestFarm1892	2222	0.18*	0.07*
indOwnershare1871	indOwnershareΔ	2222	0.32*	0.25*

Note: * indicate significance at least at the 5 percent level. Variable description: —ind1892: the top one percent income share in the industrial sector 1892 —agr1892: the top one percent income share in the agricultural sector 1892 — richestFarm1892: the top one person's, e.g. the richest farm owner's, income share in the agricultural sector 1892 — indOwnershareΔ: the difference of the share of industrial vote owners between 1871 and 1892.

Sample selection usually raises red flags, but in this case, it is instead an asset; it arises because we investigate industrial inequality very early in the industrialization process. Table 4 shows that the industrial subsample consists of more unequal municipalities. The sample reduces when I impose the restriction that the share of industrial vote owners in 1871 needs to be above one percent. In other words, the variable that drives the sample selection is the share of industrial vote owners in 1871 (indOwnershare1871). To show how the industrial subsample relates to the full sample of municipalities I cannot use data from 1871; hence, the table correlates indOwnershare1871 with measures in 1892. The table displays that the higher share of industrial vote owners a municipality has in 1871 (indOwnershare1871), the higher is the top income share in 1892, both in the agricultural (arg1892) and industrial sector (ind1892).

7 Results

7.1 Agricultural inequality is unchanged but industrial inequality increases rapidly

Table 5 shows the main results. The first panel shows the change in inequality for the full sample of rural municipalities. The income share of the top one percent richest in the agricultural sector (agr) does not change in the typical municipality between the two time points. More precisely, the point estimate shows that in the median municipality, the top one percent share increases by 0.3 percentage units, a change not significantly different from zero. The income share of the top one person in the agricultural sector, the richest farm owner (richestFarm), increases by 0.1 percentage points. The top one person share (richestFarm) measures a different aspect of inequality, even further in the top, than the top one percent share (agr). The point estimate is significantly different from zero, but since the magnitude is very small, it represents a precisely measured zero estimate. Appendix C shows that the top one person share (richestFarm) is more robustly measured than the top one percent share (agr) and thereby the point estimates confirm that agricultural inequality does not change.

Table 5: Summary Stats of Change 1892-1871

	n	median	p-value*	confidence interval (95%)
Full sample				
agrΔ	2222	0.003	0.112	(0.000, 0.006)
richestFarmΔ	2222	0.001	0.004	(0.000, 0.002)
Industrial subsample				
indΔ	1328	0.048	0.000	(0.043, 0.055)
agrΔ	1300	0.017	0.000	(0.011, 0.025)
richestFarmΔ	1328	0.002	0.004	(0.000, 0.003)

Note: * The p-values refer to a sign test on the median, with the null hypothesis that it is zero. Variable description: —agrΔ: the difference in the top one percent income share in the agricultural sector between 1871 and 1892 — richestFarmΔ: the difference in the top one person's, e.g. the richest farm owner's, income share in the agricultural sector between 1871 and 1892 — indΔ: the difference in the top one percent income share in the industrial sector between 1871 and 1892.

The second panel of Table 5 displays the change in inequality for the industrial subsample. The first row displays that inequality in the industrial sector increases between the two time points. The point estimate of the income share for the top one percent in the industrial sector (ind) is 4.8 percentage points and statistically significant. The magnitude of the change found in this paper is at similar magnitudes as the change in the top one percent income share from 1980 to 2000 for Sweden, the UK, Australia and Canada, which usually is considered to be large (Roine and Waldenström, 2008).

The point estimate of 4.8 percentage points is most likely a lower bound because the industrial sample is selected. We can extrapolate the relationships found in the available data to reason around what the point estimate would have been if no municipalities were excluded. We know that the municipalities excluded from the industrial subsample have lower inequality in 1892 (see section 6 above). At the same time, Table 6 (Ranking: Over time) shows that inequality is persistent: lower inequality in 1892 is related to lower inequality in 1871. Thus, we expect that the excluded municipalities have lower inequality in 1871 as well. Next, Table 6 (Change: Own sector) shows that there is convergence in inequality: a municipality which starts with lower inequality in 1871 experiences larger increases in inequality. Figure 5 displays the relationship graphically for industrial inequality. Thus, we expect that the excluded municipalities would experience larger increases in inequality, rendering the point estimate to be higher than the one presented in Table 5. Table 5 also displays that the change for both the top one percent share (agr) and the top person share (richestFarm) is higher in the industrial subsample than for the full sample, supporting the notion of a lower bound.

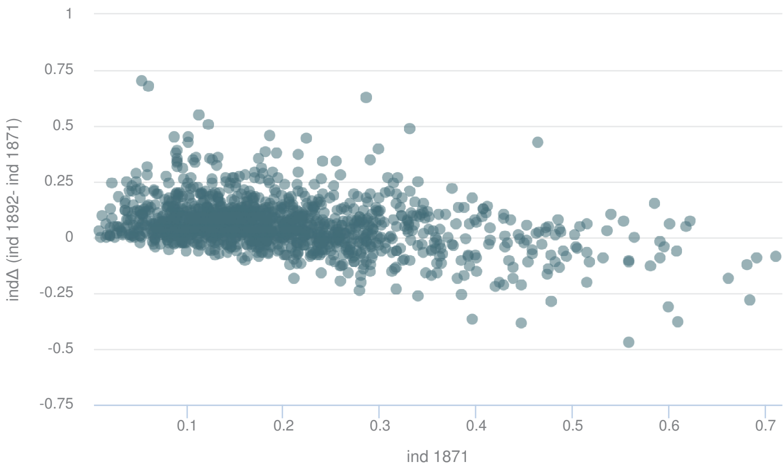


Figure 5: Change versus initial starting point of the top one percent share in industry (ind)

Robustness

It might be that the chosen assumptions affect the conclusion of unchanged agricultural inequality and increasing industrial inequality. Appendix Table C.7 shows summary stats for various ways of interpolating the tabulated data. It is clear that the presented median estimates (agr and ind) indicate the smallest change, rendering the presented estimates to be lower bounds.

Most vote owners are actual persons, but there also exist municipalities dominated by one company. Appendix Table A.1 replicate the main Table 5 but exclude all municipalities from the sample where the richest vote owners is not an actual person in 1892. The median difference is slightly smaller for all measures, but the same pattern is found. Agricultural inequality is constant and industrial inequality increases. The point estimate of the median difference in the top one percent share in the industrial sector (ind) is slightly smaller, 4.3 percentage points.[22]

[22]When using string matching to recover how municipalities change between the two years, municipalities which give up some area to another, without changing the name, are missed. Such border changes lead to measurement error in the variables. Tyrefors et al. (2017) kindly provided their information on border changes. Approximately 200 municipalities had such missed border changes. Relying on analyzing medians instead of means render the summary statistic more robust

Comparison to previous literature

To my knowledge, there are no comparable estimates for the change in income inequality in Sweden during the period, but Roine and Waldenström (2009) and Bengtsson et al. (2018) present evidence on wealth inequality on the national level. Both sets of authors use estate data and thus employ a different method than this paper. Overall, we expect a change in income inequality (the distribution of the flow) to translate to an even larger increase in wealth inequality (the distribution of the stock). Roine and Waldenström (2009) provide estimates on the top one percent wealth share in 1875 and 1907, which recalculated to the same time period as under consideration in this paper, e.g. 21 years, suggests a 3.8 percentage unit increase. Bengtsson et al. (2018) provide evidence on 1850 and 1900, suggesting a 6.7 percentage unit increase in the top one percent wealth share. My finding is more in line with Bengtsson and colleagues' (2018) evidence of rapidly increasing inequality, since, everything else equal, we expect income inequality to translate to even greater wealth inequality.

Bengtsson et al. (2018) show an increase in the wealth Gini coefficient by 3 points for rural areas, whereas for urban areas it increases by one unit only. Recalculating the change to a 21-year period suggests a 1.26 unit increase in the wealth Gini in the rural municipalities. I find a much larger change of 2-4 unit increase in the income Gini coefficint for the agricultural sector and a 10-13 unit increase for the industrial sector for the median rural municipality (see Appendix Table C.7). There are many possible explanations for the diverging result, but it is interesting to note that the current result on growth in industrial sector income inequality at the countryside might be one reason for the growth in rural wealth inequality found by Bengtsson et al. (2018).

7.2 The change in industrial inequality presented geographically

Figure 6 shows the change in industrial inequality in the median municipality in each province. For some provinces, we have smaller original sample sizes and more attrition due to not being able to measure the top one percent industrial inequality, resulting in wide confidence intervals. We can only say with confidence that Malmöhus, Stockholm, and Kristianstad display an increase in industrial inequal-

to measurement error in the variables. The main result in Table 5 does not change when excluding the 200 municipalities where borders might have changed (not shown).

ity. Nevertheless, the point estimates indicate an increase in industrial inequality in all provinces. The same information is provided in a heatmap, Figure 7, which displays the top to the bottom-ranked group of provinces. The general pattern is that the provinces in the South and North exhibit more of an increase.[23]

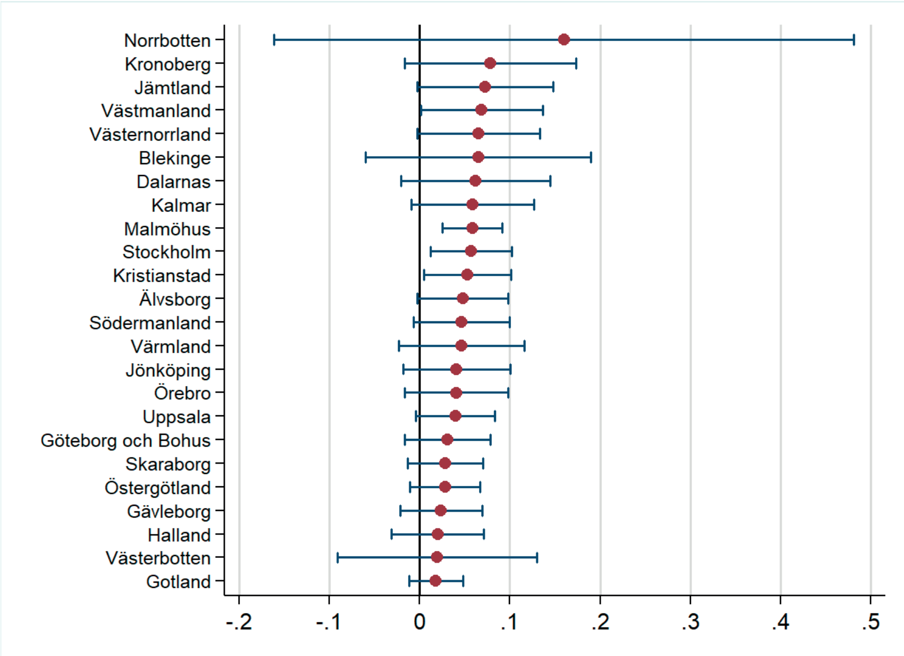


Figure 6: Median municipalities per province for indΔ (ind1892-ind1871)

[23]Regarding the provincial-level analysis (e.g. Figure 6 and 7) the described general pattern is robust to instead using the indLognΔ- variable. However, the ranking of the municipalities is affected when measured as indLognΔ, which is not strange since the difference is fairly small between two provinces. I choose to rely on indΔ as indLognΔ give higher levels and more provinces becomes significantly different from zero in Figure 6. Thus, I choose to present the conservative result.

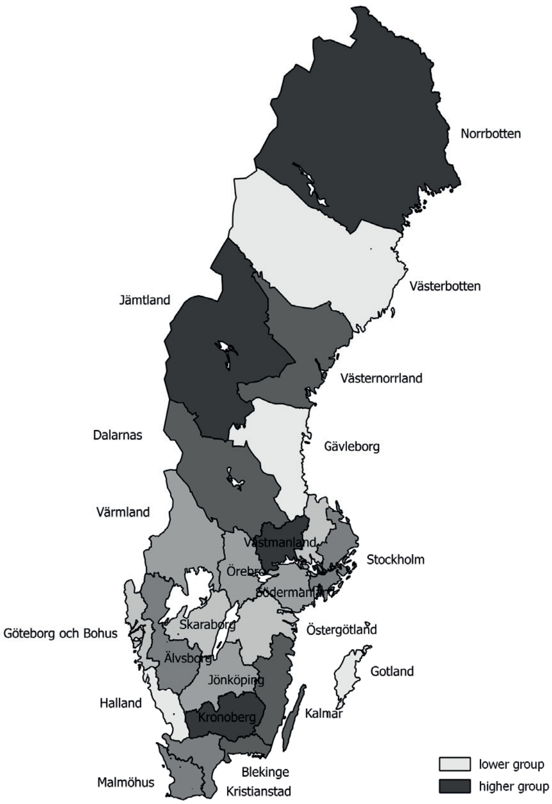


Figure 7: Heatmap of median municipality change per province by ind∆ (ind1892-ind1871)

7.3 Patterns of changes

To further our understanding of how inequality changes, Table 6 shows rank correlations between industrial and agricultural inequality over time and sectors. The table shows that inequality in agriculture and industry go together (Ranking: Over sector). More precisely, the industrial top one percent income share in 1871 (ind1871) is rank correlated at 16 percent to the agricultural top one percent income share in 1871 (agr1871). The strength of the relationship is higher in 1892 than in 1871.

Table 6: Rank correlations

Variable pair		n	Spearman	Pearson
Ranking: Over sector				
ind1871	agr1871	1310	0.16*	0.14*
ind1892	agr1892	1966	0.36*	0.36*
Ranking: Over time				
ind1871	ind1892	1328	0.59*	0.60*
agr1871	agr1892	2222	0.81*	0.77*
Ranking: Cross-sector and time				
ind1871	agr1892	1305	0.33*	0.37*
agr1871	ind1892	1966	0.28*	0.27*
Change: Own sector				
ind1871	indΔ	1328	-0.29*	-0.34*
agr1871	agrΔ	2222	-0.15*	-0.10*
Change: Cross-sector				
ind1871	agrΔ	1300	0.30*	0.37*
agr1871	indΔ	1310	0.11*	0.14*

Note: * indicate significance at least at the 5 percent level. The result is unchanged when using permutations of the measures (see Appendix Table C.8) and when only using the industrial subsample (not shown).

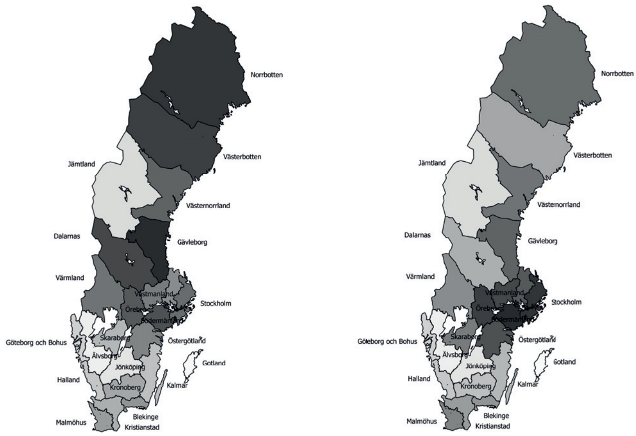
Also, the table shows that inequality is persistent over the two time points (Ranking: Over time). A municipality that scores high in inequality in 1871 also scores high in 1892. Agricultural inequality is more persistent than industrial inequality and the strength of the association is stronger than over sectors. The results are in line with previous literature. For example, Barro (2000, p. 15), from the cross-country literature, shows that Gini coefficients exhibit a correlation of 0.72 across a 20-30-year period and a correlation of 0.85 for a 10-20-year period.

Moreover, the table displays that agricultural inequality in 1871 and industrial inequality in 1892 are positively rank correlated at around 30 percent (Ranking: Cross-sector and time). Initial agricultural inequality is related to later industrial inequality. This is in line with previous literature: Modalsli (2018, p. 76) has data on land inequality for a sample of 62 municipalities from selected Norwegian provinces in 1838 and find a correlation of 0.63 to income inequality in 1868. Likewise, Oyvat (2016, p. 208) using data from the "cross-country" line of the literature find a correlation of 0.47 between land Gini coefficients 1960 and income Gini coefficients 2010.

Table 6 also shows how the change is dependent on the initial starting value in 1871 over sectors and constitutes a way of illustrating convergence or divergence in inequality levels. Before, I showed that within municipalities, agricultural inequality was constant and industrial inequality was increasing. Now, I instead show that this change within municipalities bears a relationship to the initial starting value, that is, inequality in 1871. The table (Change: Own sector) shows that the municipalities where industrial inequality was high in 1871 experienced larger decreases to 1892 than the municipalities which started out with lower industrial inequality. Figure 5 above illustrates this relationship. The result implies convergence in inequality levels over time.

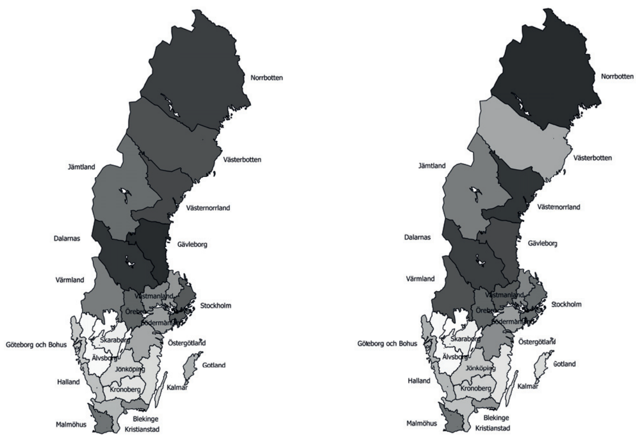
However, the opposite result is, surprisingly, found for the initial value in 1871 compared to changes in the other type of tax base (Change: Cross-sector). High initial industrial inequality is related to a change towards more agricultural inequality, which implies divergence in inequality levels. The table also shows (Change: Own sector) that higher agricultural inequality in 1871 is related to larger decreases in industrial inequality to 1892, but being a fairly small estimate, it is not robust to using other interpolation methods.

Figure 8 and 9 present agricultural and industrial inequality geographically for 1892 only, as inequality is persistent. The maps visualize that inequality in agriculture and industry partly coincide. The ranking of provinces on agricultural inequality in Figure 8 is dependent on the chosen measure and, thus, in addition to the interpolation method used so far (agr and ind), the maps also display measures built on interpolation by the standard normal distribution (indLogn and argLogn). The ranking looks fairly different, but partially this is a visual effect of the largely unpopulated provinces in the North having a large area, as the pattern over the more densely populated provinces in the South is more stable. The provinces of Gotland, Älvsborg, Jönköping, Jämtland, Göteborg och Bohus, and Halland is consistently ranked low on agricultural inequality (in the given order). In Figure 9, we see that industrial inequality partly follows the same geographical pattern as agricultural inequality.



Note: Darker shade = more unequal.

Figure 8: Map of median municipality per province in 1892 for agrLogn1892 (left) and agr1892 (right)



Note: Darker shade = more unequal.

Figure 9: Map of median municipality per province in 1892 for indLogn1892 (left) and ind1892 (right)

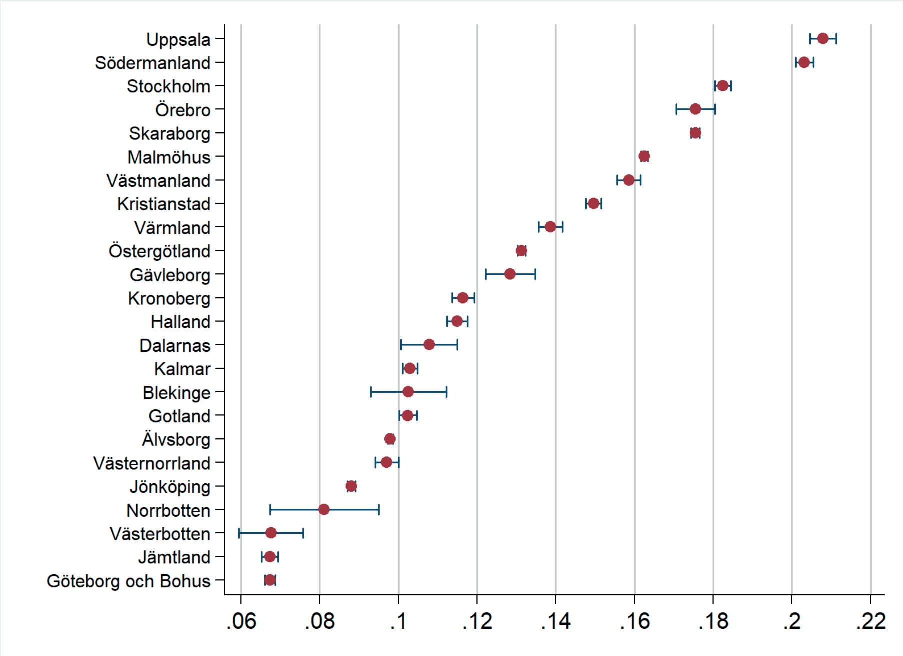


Figure 10: Mean per province for richestFarm 1892

As a robustness check, Figure 10 and 11 presents the geographical pattern of the mean of the income share of the top one person in agriculture (richestFarm). RichestFarm is the most stable measure, the robustness checks in Appendix C show that we can use the cardinal properties of the variable, but it is hard to compare to the previous literature. The richestFarm variable theoretically measures a different aspect of the agricultural income distribution than the top one percent richest, even further in the top. As we expect, also in this data material, the measures coincide but not completely (see appendix section C.2).[24] The map of richestFarm confirms the geographical pattern described in the previous paragraphs, but also shows that we partly measure different aspects as the ranking of the provinces differ. For example, the order of the "bottom ranked" provinces are reversed for richestFarm. Carlsson (1968, p. 25) mentions that half of the country's landowners (godsä-

[24]Appendix Section C.2 shows that the richestFarm measure agree with the agricultural top one percent share (agr) to a much lesser extent than the various interpolated agricultural top one percent share measures agrees with each other. This is not due to the interpolation malfunctioning, as the interpolated income of the richest person in the other measures show a rank correlation of 0.85-0.88 respectively to the richestFarm measure (not shown).

gare) resided in Malmöhus, Kristianstads, Stockholm, Uppsala, Södermanland, Östergötlands, and Skaraborg province in 1855. The results for the richestFarm is striking in that they largely confirm our presumptions: Malmöhus, Kristianstad and the provinces around Stockholm have larger top one person agricultural inequality, which decreases the further one goes from those centers.

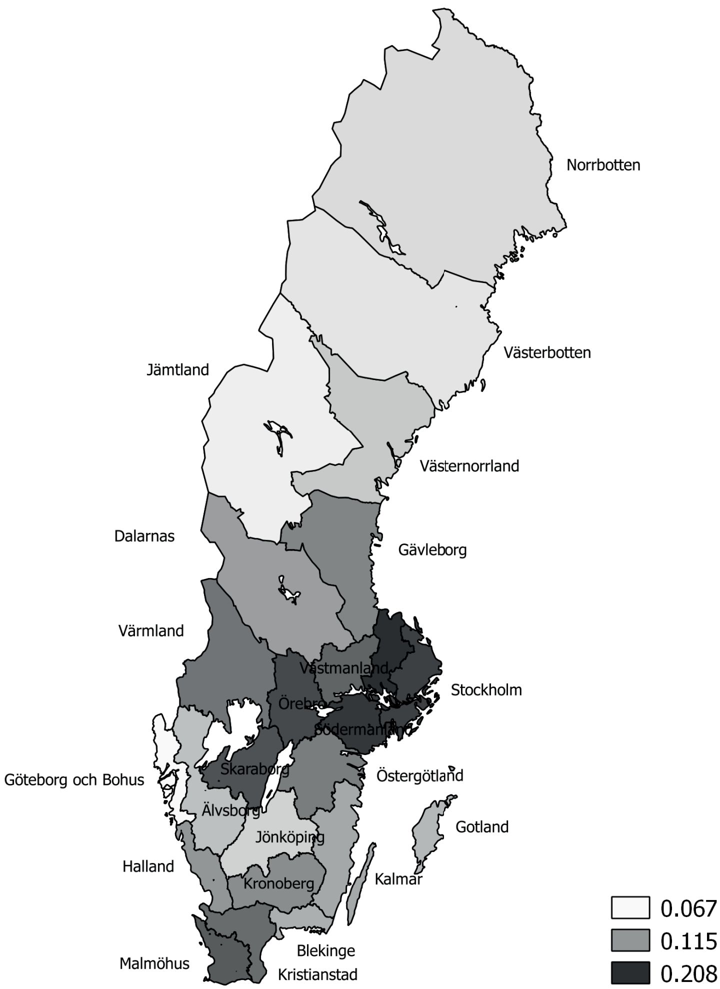


Figure 11: Map of the mean of richestFarm 1892 per province

8 Conclusion

This paper presents new data on municipality-level inequality for Sweden in 1871 and 1892. The paper provides one piece of evidence to further the debate on the relationship between inequality and development. The new municipality-level estimates are possible to construct from data on the vote distribution. In the electoral system at this time, people got votes in relation to their incomes, which renders it possible to back out the incomes of the top one percent. The paper shows that agricultural inequality does not change but that the income share of the top one percent in the industrial sector increases by at least 4.8 percentage points in the median rural municipality. It complements the analysis of the national-level wealth inequality by Roine and Waldenström (2009) and Bengtsson et al. (2018). The current result strengthens the interpretation of rapidly increasing inequality in the second half of the 19th century as found by Bengtsson et al. (2018), but also adds nuance by showing that the growth in inequality takes place in industrial sector incomes. The results presented are robust to alternative assumptions.

The constructed estimates are hopefully useful in further studies on the effect of inequality on other municipality-level variables. For instance, the historical kingdoms of Denmark and Sweden might have had differential propensities to install local elites, and the current data allows for a regression continuity design between municipalities at either side of historical provincial borders.[25]

[25] Göteborg och Bohus, Halland, Kristianstad, Belkinge, and Jämtland have been under Danish rule.

References

Aldén, G. (1901). Kommunalförvaltningen. In Sundbärg, G., editor, *Sveriges land och folk: historisk-statistisk handbok*, pages 226–238. Kungl. boktryckeriet P.A. Nordstedt & söner, Stockholm.

Andersson, J. and Berger, T. (2018). Elites and the Expansion of Education in Nineteenth-century Sweden. *The Economic History Review*, 0(0):1 – 28.

Aronsson, P. (1999). Lokala medborgarskapet- det dolda kulturarvet. In Amnå, E., editor, *Demokrati och medborgarskap (SOU 1999:77)*. Fakta info direkt, Stockholm.

Atkinson, A. (2007). Measuring Top Incomes: Methodological Issues. In Atkinson, A. and Piketty, T., editors, *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*, pages 18–42. Oxford University Press Oxford.

Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature*, 49(1):3–71.

Barro, R. J. (2000). Inequality and Growth in a Panel of Countries. *Journal of Economic Growth*, 5(1):5–32.

Bengtsson, E., Missiaia, A., Olsson, M., and Svensson, P. (2018). Wealth Inequality in Sweden, 1750 - 1900. *The Economic History Review*, 71(3):772–794.

Carlsson, S. (1968). *Yrken och samhällsgrupper: Den sociala omgrupperingen i Sverige efter 1866*. Almqvist & Wiksell, Stockholm.

Cowell, F. A. and Flachaire, E. (2015). Statistical Methods for Distributional Analysis. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2 of *Handbook of Income Distribution*, chapter 6, pages 359 – 465. Elsevier.

Cowell, F. A. and Kerm, P. (2015). Wealth inequality: A survey. *Journal of Economic Surveys*, 29(4):671–710.

Easterly, W. (2007). Inequality Does Cause Underdevelopment: Insights from a New Instrument. *Journal of Development Economics*, 84(2):755–776.

Enflo, K., Henning, M., and Schön, L. (2014). Swedish Regional GDP 1855-2000: Estimations and General Trends in the Swedish Regional System. *Research in Economic History*, 30:47–89.

Flodström, I. (1906). *Redogörelse för statistisk utredning i och för utarbetande af förslag till förändrade bestämmelser dels angående bevillning af fast egendom dels ock om inkomstskatt*. Stockholm.

Gårestad, P. (1987). *Industrialisering och beskattning i Sverige 1861-1914*. Almqvist & Wiksell International, Uppsala.

Leigh, A. (2007). How Closely Do Top Income Shares Track Other Measures of Inequality? *The Economic Journal*, 117(524):F619–F633.

Modalsli, J. (2018). The Regional Dispersion of Income Inequality in Nineteenth-century Norway. *Explorations in Economic History*, 67:62 – 79.

Neves, P. C., Afonso, O., and Silva, S. T. (2016). A Meta-Analytic Reassessment of the Effects of Inequality on Growth. *World Development*, 78:386 – 400.

Nordisk Familjeordbok (1876-1899). 1 edition.

Olsson, M. (2005). *Skatta dig lycklig. Jordränta och jordbruk i Skåne 1660-1900*. Gidlunds förlag, Hedemora.

Oyvat, C. (2016). Agrarian Structures, Urbanization, and Inequality. *World Development*, 83:207–230.

Piketty, T. (2007). Top Incomes over the Twentieth Century: A Summary of Main Findings. In Atkinson, A. and Piketty, T., editors, *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*, pages 1–17. Oxford University Press Oxford.

Ravallion, M. (2001). Growth, Inequality and Poverty: Looking Beyond Averages. *World development*, 29(11):1803–1815.

Roine, J. and Waldenström, D. (2008). The Evolution of Top Incomes in an Egalitarian Society: Sweden, 1903–2004. *Journal of Public Economics*, 92(1):366–387.

Roine, J. and Waldenström, D. (2009). Wealth Concentration over the Path of Development: Sweden, 1873–2006. *The Scandinavian Journal of Economics*, 111(1):151–187.

Rydin, H. L. (1882). PM angående det svenska skatteväsendets utveckling. In *Bihang till Riksdagens protokoll vid lagtima riksdagen i Stockholm år 1883*. Riksdagen, Stockholm.

SFS 1861:34. Kongl. Maj:ts Nådiga Kundgörelse den 17 Maj 1861 angående den av Rikets Ständer vid 1859-1860 års Riksdag åtagna bevillningen (Bevillningsförordningen/Bevillningsstadgan).

SFS 1862:13. Kongl. Maj:ts Nådiga Förordning den 21 Mars 1862 om kommunalstyrelse på landet.

SFS 1862:14. Kongl. Maj:ts Nådiga Förordning den 21 Mars 1862 om kommunalstyrelse i städer.

SFS 1863:50. Kongl. Maj:ts Nådiga Kundgörelse den 15 September 1863 angående förändrad lydelse av vissa §§ i Kongl. Förordningen den 21 Mars 1862 om kommunalstyrelse på landet.

SFS 1868:59. Kongl. Maj:ts Nådiga Kundgörelse den 8 September 1868 angående förändrad lydelse av 12 § 1 mom samt 70 § i kongl. förordningen om kommunalstyrelse i stad den 21 Mars 1862.

SFS 1892:110. Kongl. Maj:ts Nådiga Kundgörelse den 2 December 1892 om viss ändring i förordningen angående bevillning af fast egendom samt af inkomst den 3 Juni 1892 m.m.

SFS 1892:113. Kongl. Maj:ts Nådiga Förordning den 2 December 1892 angående ändradlydelse af §58 i förordningen om kommunalstyrelse på landet.

SFS 1892:44. Kongl. Maj:ts Nådiga Förordning den 3 Juni 1892 angående bevillning av fast egendom samt af inkomst.

SFS 1900:86. Lag den 9 November 1900 angående ändrad lydelse af §§ 11, 21, 59 och 73 i förordningen om kommunalstyrelse på landet den 21 Mars 1862.

Simons, H. C. (1938). *Personal Income Taxation: The Definition of Income as a Problem of Fiscal Policy*. Chicago University, Chicago.

Skatteverket (2003). *Deklarationen 100 år och andra tillbakablickar*. Erlanders Gotab, Stockholm.

Statistics Sweden (1874). Bidrag till Sveriges officiella statistik R Valstatistik II. Data. (BiSOS R 1871).

Statistics Sweden (1878). Bidrag till Sveriges officiella statistik. H, Kungl. Maj:ts befallningshafvandes femårsberättelser, Ny följd IV jemste sammandrag för åren 1871-1875.

Statistics Sweden (1894a). Bidrag till Sveriges officiella statistik N) Jordbruk och boskapsskötsel. XXVIII. Hushållningssällskapens berättelser för år 1892. Data. (BiSOS N 1892).

Statistics Sweden (1894b). Bidrag till Sveriges officiella statistik U) Kommunernas fattigvård och finanser. xix. Data. (BiSOS U 1892).

Statistics Sweden (1895). Bidrag till Sveriges officiella statistik R Valstatistik XV. Data. (BiSOS R 1892).

Statistics Sweden (1900). Bidrag till Sveriges officiella statistik. H, Kungl. Maj:ts befallningshafvandes femårsberättelser, Ny följd VIII jemste sammandrag för åren 1891-1895.

Tyrefors, B., Lindgren, E., and Pettersson-Lidbom, P. (2017). Political Power, Resistance to Technological Change and Economic Development: Evidence from the 19th century Sweden. IFN Working Paper No. 1172, IFN.

Velander, J. (1901). Stadsfinanserna. In Sundbärg, G., editor, *Sveriges land och folk: historisk-statistisk handbok*, pages 187–191. Kungl. boktryckeriet P.A. Nordstedt & söner, Stockholm.

Appendix A. Additional Information and Results

A.1 Example of transformed problem behind the Midpoint method for top one income in agriculture- Giresta 1892

The objective function behind finding the maximum allocation is the following:

max: +C1 +C2 +C3 +C5 +C6 +C7 +C8 +C9 +C12 +C13 +C14 +C15 +C16 +C17 +C18 +C19 +C20 +C21 +C22 +C23 +C24 +C25+C26 +C27 +C28 +C29 +C30 +C31 +C32

where C1 denotes the income of the possibly richest agricultural vote owner, C2 the next possible richest etc. Introspection shows that C4 is not included in the objective function. C4 denotes the possibly richest industrial vote owner, C10 the next possible richest etc. The objective function is maximized under the following restrictions:

C1 +C2 +C3 +C4 +C5 +C6 +C7 +C8 +C9 +C10 +C11 +C12 +C13 +C14 +C15 +C16 +C17 +C18 +C19 +C20 +C21 +C22 +C23 +C24 +C25 +C26 +C27 +C28 +C29 +C30 +C31 +C32 +C33 +C34 +C35 +C36 +C37 +C38 = 4472
C1 = 995
894 <= C1 <= 995
500 <= C2 <= 894
500 <= C3 <= 894
224 <= C4 <= 250
100 <= C5 <= 224
100 <= C6 <= 224
100 <= C7 <= 224
100 <= C8 <= 224
100 <= C9 <= 224
100 <= C10 <= 224
100 <= C11 <= 224
50 <= C12 <= 89
50 <= C13 <= 89
50 <= C14 <= 89
50 <= C15 <= 89
50 <= C16 <= 89
25 <= C17 <= 50

25 <= C18 <= 50
25 <= C19 <= 50
25 <= C20 <= 50
25 <= C21 <= 50
25 <= C22 <= 50
25 <= C23 <= 50
25 <= C24 <= 50
10 <= C25 <= 25
10 <= C26 <= 25
10 <= C27 <= 25
10 <= C28 <= 25
10 <= C29 <= 25
10 <= C30 <= 25
10 <= C31 <= 25
10 <= C32 <= 25
10 <= C33 <= 25
5 <= C34 <= 10
1 <= C35 <= 5
1 <= C36 <= 5
1 <= C37 <= 5
1 <= C38 <= 5

Table A.1: Summary Stats of Change 1892-1871, municipalities where the richest vote owner is not an actual person are excluded

	n	median	p-value*	confidence interval (95%)
Full sample				
agrΔ	1892	-0.002	1.000	(-0.006, -0.002)
richestFarmΔ	1892	0.000	0.945	(-0.001, 0.001)
Industrial subsample				
indΔ	997	0.043	0.000	(0.037, 0.048)
agrΔ	983	0.006	0.048	(0.000, 0.012)
richestFarmΔ	997	0.001	0.081	(0.000, 0.002)

* The p-values refer to a sign test on the median, with the null hypothesis that it is zero. Variable description: —agrΔ: the difference in the top one percent income share in the agricultural sector between 1871 and 1892 — richestFarmΔ: the difference in the top one person's, eg. the richest farm owner's, income share in the agricultural sector between 1871 and 1892 — indΔ: the difference in the top one percent income share in the industrial sector between 1871 and 1892.

Appendix B. Further Institutional Details

B.1 Governmental-level tax rates and deductions

For agricultural property, the tax rate was 0.03 percent of the property value and for other property, it was 0.05 percent of the property value, for each whole 100 SEK.[1] The reason for the differential tax rates might be that agricultural property was also taxed in other terms.[2] For income of labor and capital the tax rate was 1 percent. Incomes lower than 400 SEK were deductible at that amount, and from incomes lower than 1800 SEK, one deducted 300 SEK.[3]

B.2 The process of setting the governmental-level tax base

Figure B.1 illustrates the institutions affecting the tax base decisions. The tax proposal committee (*bevillningsberedning*) proposed the property value and the income of each inhabitant. The tax committee (*taxeringsnämnd/ taxeringskommité*) decided the respective taxation bases.[4] The property values were updated every five years and the income values every year.[5]

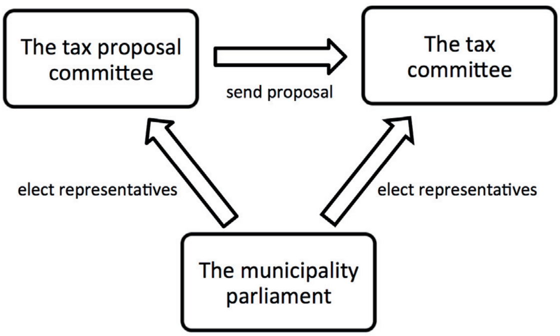


Figure B.1: Overview of institutions affecting tax base decisions.

---

[1] SFS 1861:34, §2; Velander 1901, p. 189

[2] Some of which was non-monetary, for example, the tax to provide for a solider and the "road tax".

[3] SFS 1861:34, §8

[4] SFS 1861:34, §31, §34

[5] Velander 1901, p. 189

The municipality parliament elected representatives to the tax proposal committee.[6] The tax proposal committee consisted of two-three persons from the municipality parliament as well as a president chosen by the provincial-level government (*länstyrelsen*). The tax proposal committee usually spanned a parish, which spanned a municipality, but larger parishes could consist of many tax proposal districts, and smaller parishes could merge into one tax proposal district.[7]

The tax proposal committee's work was to organize the information that had been sent to them and propose the tax base for each inhabitant. By law, the property value should be estimated to its real value with the guidance of sales contract, rental contracts, debt, fire insurance and "other collected information".[8][9] In addition, some of this information, for example, debt information from banks, should by law be sent to the authorities every year.[10] Furthermore, inhabitants could by their own volition give the tax proposal committee information of his/her property value and income.[11] Income tax returns were not introduced until 1905 (decided in 1902).[12]

The proposal was made publicly available. The tax proposal committee sent the proposal to the tax committee, and the taxed persons could attend the tax committee's meetings to rectify it. The tax committee decided the property value and income that each person should tax for.[13]

The municipality parliament elected representatives to the tax committee and at least one of the members in the tax proposal committee needed to be elected to the tax committee.[14] One tax committee regularly spanned 2-4 tax proposal districts (/ parishes/ municipalities).[15] [16] [17] The governmental tax administrator should attend the meetings, as well as the tax proposal committee's president when the

[6]Velander 1901, p. 189; Nordisk Familjeordbok 1899, Bevillnings-beredning;

[7]Nordisk Familjeordbok 1899, Bevillnings-beredning; SFS 1861:34, §34

[8]in Swedish "erhållna uppgifter eller inhämtade upplysningar".

[9]SFS 1861:34, §2; Nordisk Familjeordbok 1899, Bevillnings-beredning; Nordisk Familjeordbok 1899, Taxeringsvärde

[10]SFS 1861:34, §25

[11]SFS 1861:34, §30

[12]Skatteverket 2003, p. 7

[13]SFS 1861:34, §38, §39, §54; Nordisk Familjeordbok 1899, Taxeringsnämnd

[14]SFS 1861:34, §47; Velander 1901, p. 189; Nordisk Familjeordbok 1899, Taxeringsnämnd

[15]But in case of towns the tax committee spanned the town-municipality.

[16]In SFS 1861:34, §43 it is stated that "pastorat" is the basis for the tax committee's area, with the usual exceptions.

[17]Velander 1901, p. 189; Nordisk Familjeordbok 1899, Taxeringsnämnd

matters of his tax proposal district were discussed.[18]

After the tax committee had decided the tax base, it was made publicly available. The inhabitants and the municipality had the right to appeal to another instance at the provincial level (*prövningskomitée*).[19]

B.3 More on votes

The number of votes in the municipality was read aloud from the church pulpit and made publicly available for cross-checking. The inhabitants could complain to the municipality parliament, which decided on changes before the number of votes was decided upon.[20]

The municipalities that were towns did not have agricultural properties and, hence, inhabitants only paid tax for other property and for income and received votes from those sources. In towns, people received one vote for taxes paid up until 1 SEK, and, after that, one vote for every whole SEK paid in tax.[21] From 1869 in the towns, the maximum number of votes a vote owner could hold were 2 percent of the total or 100 votes.[22]

B.4 The municipality tax base

The number of votes arising through the various taxation types decided the municipality-level tax base for each person. Votes arising through agricultural property in *mantalsatt jord* (farmland that historically had been assigned a type of taxation weight) paid twice as much and other agricultural property paid 1/3 more than did the votes arising from taxation of other property or income.[23] In addition to levy taxes on themselves in relation to their governmental-level tax/votes by deciding the municipality tax rate, the vote owners could decide on the rate of the poll tax (tax per capita) for all adult inhabitants.[24]

[18] SFS 1861:34, §42, §49; Nordisk Familjeordbok 1899, Taxeringsnämnd

[19] SFS 1861:34, §56; Nordisk Familjeordbok 1899, Taxeringsnämnd

[20] SFS 1862:13, §65, §66

[21] SFS 1862:14, §12; SFS 1868:59, §12

[22] SFS 1869:59 §12; Aldén 1901, p. 228

[23] SFS 1863:50, §64, Litt.B; Rydin 1882, p. 351

[24] SFS 1861:13 §59, §65.

B.5 Changes in municipality governance structure until universal suffrage was established

The following section presents the changes to the municipality governance structure until universal suffrage was established at the municipality level.

- From 1893 (decided in 1892), the governmental-level tax rate of agricultural property changed to 0.06 percent in conjunction with that the other types of taxes levied on the agricultural property in the older system were abolished.[25] [26] At the same time, from 1893, both categories of taxes received the same number of votes, that is, agricultural property received votes as did the industrial taxation types. The two changes in combination imply that the influence of vote owners holding agricultural property stayed constant.
- From 1893 (decided in 1892), income tax was deductible accordingly: 500 SEK was tax-free; for 500-1200 SEK, a deduction of 450 SEK applied; for 1200-1800 SEK, a deduction of 300 SEK applied. The deductions increased with a maximum of 200 SEK due to high living cost at the place of residence and/or poverty.[27] The taxation of the property was included in the calculation of the deductions: by law, 6 percent of the property value of agricultural property and 5 percent of the property value of other property was thought of as income.[28]
- From 1901 (decided in 1900), in the countryside municipalities, the limitations became that nobody could own more than 10 percent of the municipality's votes and not more than 5000 votes.[29]
- From 1902, the income tax rate was increased to two percent instead of one.

[25] This relates to the other types of taxes on agricultural property (*grundskatter*). In 1885, those taxes were reduced by 30 percent — either by that the monetary tax was abolished or, for example, in the case of the tax to provide for a soldier, in the form of the government paying the farmer for the provision with 30 percent of its estimated value. In 1892 and 1898, it was decided that the same types of taxes were going to be reduced in the same manner during the 12 years, which started in 1893, so the taxes ceased to exist in 1903.

[26] SFS 1892:110, §1; Velander 1901, p. 189; Olsson 2005, pp. 33, 80

[27] SFS 1892:44, §11; Aldén 1901, pp. 227-8; Velander 1901, p. 189

[28] SFS 1892:110, §17

[29] SFS 1900:86, §11; Aldén 1901, p. 228

- From 1909, for both countryside municipalities and towns, the system was that 100 SEK in income tax gave 1 vote for incomes under 2000 SEK and that 500 SEK in income tax gave 1 vote for incomes above 2000 SEK, with a maximum of 40 votes. At the same time, companies and alike juridical persons that were not persons in the standard sense lost their vote rights.[30]

- In 1918, men and women gained vote rights. Certain limitations still existed, such as not taking poverty relief or not having unfulfilled tax payments for more than two of the past three years.[31]

[30]Aronsson 1999, p. 270
[31]Aronsson 1999, p. 270

Appendix C. Sensitivity and Robustness

Permutations of assumptions The baseline methods for interpolating the interval data is *Midpoint*, as described and used in the main article. I also test *Uni* – that the interpolation is done by assuming a uniform distribution within each interval, and *Logn,* – that a standard lognormal distribution is assumed instead. For the later two measures, only information from the first two sets of intervals[1] is used. The baseline imputation is that people below the taxation limit have an income set to 50 percent of the taxation limit. I test the various measures with 10 and 75 percent of the taxation limit. The baseline total population is the total number of inhabitants in each municipality, including women and children. In addition, I test the various measures with the following:

- The baseline total population weighted by the percentage of agricultural (industrial) vote owners (*OwnerShare*). The reason behind this check is that the total relevant population theoretically should be only people involved in agriculture (industry), and it is reasonable to expect that it is proportional to the amount of agricultural (industrial) vote owners.
- The baseline total population weighted by the percentage of agricultural votes (*TaxShare*). The logic is equivalent as for the check above.
- The baseline total population deduced by the fraction of persons that are under 21 years old or married females (*Taxpaying)*. The reason for this check is that you could only get votes when you were above 21 years old, and if you were a married female, you were not a juridical subject. This check corresponds to verifying against the total theoretical taxpaying population. This check is only available for 1892 with the usage of data from the 1890 census.

Also, Gini coefficients are calculated for *Uni* and *Logn* to be able to compare to the previous literature. The Gini coefficients are based on interpolating the non-taxed part of the population between the assumed income below the taxation limit and the taxation limit.

[1] More precisely, information on the number of predominantly agricultural (industrial) vote owners that have votes above $\{0, 2, 5, 10, 20, 50\}$ percent of the total sum of votes in each municipality.

Naming convention An example of a variable name is *agrLogn1871*. Firstly, the type of inequality measured is indicated in the variable names, *agr* stands for agricultural and *ind* for industrial. Then, the type of intrapolation is indicated, such as *Logn* in this example. In the main article is the part indicating the method, *Midpoint*, omitted for readability. Lastly the time dimension used follows, such as *1871* in the example. Note that Δ or *Diff* indicates the variable value 1892 with the variable value 1871 subtracted from it. Note that the *Logn* and the *Uni* variables, without suffixes, refer to measures of the top one percent and that the *LognGini*and the *UniGini* variables, with suffixes, refer to the Gini coefficients.

Rank correlation The analysis partly relies on Spearman rank correlation coefficients. Spearman's rank coefficient is -1 if the two variables list the municipalities completely opposite to each other and 1 if the two variables list the municipalities in perfect agreement.

C.1 Analysis of alternative assumptions

Table C.1 displays the lowest rank correlation between the various versions of each measure, arising from the permutations of the assumptions on total population and total income. The table shows that assumptions on total income or total population do not affect the ranking of municipalities much. Instead, Table C.2 shows that the ranking of the various measures with the baseline assumption on total population and total income. The tables show that the various measures largely agree on the ranking (the next section, appendix section C.2, analyses the table further). In combination, the tables show that the various assumptions do not affect the ranking of the measures much.

Tables C.3, C.4, C.5, and C.6 instead show how varying the assumptions affect the median and the mean. The conclusion is that we cannot rely on the level of the measurements as the differences are too high. The exception to the above statements is *richestFarm* which is insensitive to the assumptions.

C.2 Correlations: Does the measure of inequality matter?

A large debate on different measures of inequality exists in the economic literature since evaluating inequality necessitates some kind of value judgment. Examples also exist when the different measures rank countries differently (Cowell and Kerm, 2015, pp. 682-683). In addition to the above concerns, we are here dealing with interpolated data. Table C.2 shows the ranking of the various measures. Some points are worth noting:

Firstly, the measures largely agree on the ranking. That is, regardless whether we choose to allocate all people in an interval to its middle as by *Midpoint*[2], or evenly over an interval as by *Uni*, or assume that people in an interval follow a standard Log-normal distribution as by *Logn*, the ranking is the same. This strengthens our confidence in that our conclusions will not be dependent upon assumptions imposed on the data.

Secondly, the Gini coefficients and the measures of the top one percent by the same method do not coincide to a great extent for the agricultural measures. With the measures of the top one percent, we only care about how much the one percent richest own of the total, whereas with the Gini coefficients, the next 99 percent influence the measure. For the industrial measures, the Gini coefficients coincide with the measures of the top one percent.

[2] This is not strictly what I do, but very similar in spirit.

Table C.1: Varying Assumptions: Worse Rank Correlation

Measure	Version pair				Spearman
	Version 1		Version 2		
	Total Income	Total Population	Total Income	Total Population	
Agriculture 1871					
agrLogn1871	10% of tax limit	OwnerShare	75% of tax limit	TaxShare	0.96
agrUni1871	10% of tax limit	OwnerShare	75% of tax limit	TaxShare	0.94
agrMidpoint1871	10% of tax limit	Basline	10% of tax limit	TaxShare	0.94
richestFarm1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.99
agrLognGini1871	75% of tax limit	Basline	10% of tax limit	TaxShare	0.63
agrUniGini1871	75% of tax limit	Basline	10% of tax limit	TaxShare	0.59
Agriculture 1892					
agrLogn1892	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.90
agrUni1892	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.87
agrMidpoint1892	10% of tax limit	Basline	10% of tax limit	OwnerShare	0.87
richestFarm1892	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.99
agrLognGini1892	75% of tax limit	Basline	10% of tax limit	TaxShare	0.59
agrUniGini1892	75% of tax limit	Basline	10% of tax limit	TaxShare	0.63
Industry 1871					
indLogn1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.76
indUni1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.76
indMidpoint1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.73
indLognGini1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.71
indUniGini1871	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.63
Industry 1892					
indLogn1892	75% of tax limit	Basline	10% of tax limit	TaxShare	0.74
indUni1892	75% of tax limit	Basline	10% of tax limit	TaxShare	0.68
indMidpoint1892	75% of tax limit	Basline	10% of tax limit	TaxShare	0.70
indLognGini1892	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.51
indUniGini1892	75% of tax limit	Basline	10% of tax limit	OwnerShare	0.43

Source: BiSOS R 1892, Census 1890, BiSOS R 1871.

Table C.2: Correlations, baseline assumptions

Variable pair		n	Spearman	Pearson
	Same measures			
agrLogn1871	agrUni1871	2268	0.97*	0.96*
agrLogn1871	agrMidpoint1871	2268	0.86*	0.66*
agrUni1871	agrMidpoint1871	2268	0.79*	0.55*
agrLogn1892	agrUni1892	2328	0.97*	0.96*
agrLogn1892	agrMidpoint1892	2328	0.87*	0.67*
agrUni1892	agrMidpoint1892	2328	0.81*	0.63*
indLogn1871	indUni1871	1377	0.95*	0.95*
indLogn1871	indMidpoint1871	1377	0.95*	0.92*
indUni1871	indMidpoint1871	1377	0.93*	0.83*
indLogn1892	indUni1892	2077	0.95*	0.96*
indLogn1892	indMidpoint1892	2077	0.94*	0.84*
indUni1892	indMidpoint1892	2077	0.90*	0.77*
agrLognGini1871	agrUniGini1871	2268	0.91*	0.93*
agrLognGini1892	agrUniGini1892	2328	0.90*	0.92*
indLognGini1871	indUniGini1871	1377	0.97*	0.96*
indLognGini1892	indUniGini1892	2077	0.97*	0.96*
	Different measures			
agrLogn1871	agrLognGini1871	2268	0.67*	0.53*
agrLogn1871	agrUniGini1871	2268	0.75*	0.55*
agrUni1871	agrLognGini1871	2268	0.60*	0.43*
agrUni1871	agrUniGini1871	2268	0.75*	0.51*
agrMidpoint1871	agrLognGini1871	2268	0.61*	0.56*
agrMidpoint1871	agrUniGini1871	2268	0.62*	0.56*
agrLogn1892	agrLognGini1892	2328	0.68*	0.42*
agrLogn1892	agrUniGini1892	2328	0.78*	0.48*
agrUni1892	agrLognGini1892	2328	0.59*	0.32*
agrUni1892	agrUniGini1892	2328	0.76*	0.40*
agrMidpoint1892	agrLognGini1892	2328	0.63*	0.51*
agrMidpoint1892	agrUniGini1892	2328	0.70*	0.58*

Continued on next page

Table C.2 – *Continued from previous page*

Variable pair		n	Spearman	Pearson
indLogn1871	indLognGini1871	1377	0.97*	0.85*
indLogn1871	indUniGini1871	1377	0.93*	0.78*
indUni1871	indLognGini1871	1377	0.92*	0.80*
indUni1871	indUniGini1871	1377	0.94*	0.79*
indMidpoint1871	indLognGini1871	1377	0.91*	0.84*
indMidpoint1871	indUniGini1871	1377	0.87*	0.79*
indLogn1892	indLognGini1892	2077	0.97*	0.76*
indLogn1892	indUniGini1892	2077	0.92*	0.67*
indUni1892	indLognGini1892	2077	0.90*	0.72*
indUni1892	indUniGini1892	2077	0.92*	0.67*
indMidpoint1892	indLognGini1892	2077	0.89*	0.79*
indMidpoint1892	indUniGini1892	2077	0.84*	0.71*
agrLogn1871	richestFarm1871	2268	0.46*	0.26*
agrUni1871	richestFarm1871	2268	0.38*	0.15*
agrMidpoint1871	richestFarm1871	2268	0.68*	0.65*
agrLogn1892	richestFarm1892	2328	0.42*	0.14*
agrUni1892	richestFarm1892	2328	0.35*	0.06*
agrMidpoint1892	richestFarm1892	2328	0.65*	0.54*
richestFarm1871	agrLognGini1871	2268	0.40*	0.31*
richestFarm1871	agrUniGini1871	2268	0.30*	0.23*
richestFarm1892	agrLognGini1892	2328	0.40*	0.30*
richestFarm1892	agrUniGini1892	2328	0.34*	0.27*

Note * indicate significance at least at the 5 percent level.

Table C.3: Varying assumptions: Worse Summary Stats Agriculture 1892

Varying Assumptions: Worse result for each measurement: Agriculture 1892				
	Baseline	Highest Differences		
	Value	Value	Total Population	Total Income
richestFarm1892				
Median	0.099	0.107	TaxShare	10% of tax limit
Mean	0.135	0.146	OwnerShare	10% of tax limit
agrMidpoint1892				
Median	0.426	0.336	Taxpaying	75% of tax limit
Mean	0.466	0.374	OwnerShare	75% of tax limit
agrLogn1892				
Median	0.379	0.278	TaxShare	75% of tax limit
Mean	0.456	0.333	TaxShare	75% of tax limit
agrUni1892				
Median	0.595	0.401	TaxShare	75% of tax limit
Mean	0.797	0.522	TaxShare	75% of tax limit
agrLognGini1892				
Median	0.826	0.913	Baseline	10% of tax limit
Mean	0.805	0.904	Baseline	10% of tax limit
agrUniGini1892				
Median	0.871	0.899	Baseline	10% of tax limit
Mean	0.857	0.891	Baseline	10% of tax limit

Source: BiSOS R 1892, Census 1890, BiSOS R 1871.

Table C.4: Varying assumptions: Worse Summary Stats Agriculture 1871

Varying Assumptions: Worse result for each measurement: Agriculture 1871				
	Baseline	Highest Differences		
	Value	Value	Total Population	Total Income
richestFarm1871				
Median	0.095	0.103	TaxShare	10% of tax limit
Mean	0.132	0.144	TaxShare	10% of tax limit
agrMidpoint1871				
Median	0.411	0.358	TaxShare	75% of tax limit
Mean	0.439	0.387	TaxShare	75% of tax limit
agrLogn1871				
Median	0.357	0.257	TaxShare	75% of tax limit
Mean	0.401	0.299	TaxShare	75% of tax limit
agrUni1871				
Median	0.568	0.387	TaxShare	75% of tax limit
Mean	0.665	0.453	TaxShare	75% of tax limit
agrLogNGini1871				
Median	0.786	0.900	OwnerShare	10% of tax limit
Mean	0.763	0.892	OwnerShare	10% of tax limit
agrUniGini1871				
Median	0.847	0.884	OwnerShare	10% of tax limit
Mean	0.830	0.876	OwnerShare	10% of tax limit

Source: BiSOS R 1892, Census 1890, BiSOS R 1871.

Table C.5: Varying assumptions: Worse Summary Stats Industry 1892

Varying Assumptions: Worse result for each measurement: Industry 1892				
	Baseline	Highest Differences		
	Value	Value	Total Population	Total Income
indMidpoint1892				
Median	0.202	0.521	Baseline	10% of tax limit
Mean	0.225	0.521	Baseline	10% of tax limit
indLogn1892				
Median	0.154	0.410	Taxpaying	10% of tax limit
Mean	0.194	0.438	Baseline	10% of tax limit
indUni1892				
Median	0.289	0.749	Baseline	10% of tax limit
Mean	0.353	0.825	Baseline	10% of tax limit
indLognGini1892				
Median	0.672	0.908	OwnerShare	10% of tax limit
Mean	0.631	0.897	OwnerShare	10% of tax limit
indUniGini1892				
Median	0.791	0.886	OwnerShare	10% of tax limit
Mean	0.732	0.879	OwnerShare	10% of tax limit

Source: BiSOS R 1892, Census 1890, BiSOS R 1871.

Table C.6: Varying assumptions: Worse Summary Stats Industry 1871

Varying Assumptions: Worse result for each measurement: Industry 1871				
	Baseline	Highest Differences		
	Value	Value	Total Population	Total Income
indMidpoint1871				
Median	0.170	0.503	Baseline	10% of tax limit
Mean	0.193	0.499	Baseline	10% of tax limit
indLogn1871				
Median	0.118	0.387	TaxShare	10% of tax limit
Mean	0.146	0.398	TaxShare	10% of tax limit
indUni1871				
Median	0.228	0.683	Baseline	10% of tax limit
Mean	0.277	0.730	Baseline	10% of tax limit
indLognGini1871				
Median	0.479	0.916	OwnerShare	10% of tax limit
Mean	0.475	0.891	OwnerShare	10% of tax limit
indUniGini1871				
Median	0.613	0.894	OwnerShare	10% of tax limit
Mean	0.589	0.880	OwnerShare	10% of tax limit

Source: BiSOS R 1892 and BiSOS R 1871.

Table C.7: Robustness Check for Summary Stats of Change 1892-1871

	n	mean	sd	median	p-value*
Full sample					
agrMidpointΔ	2222	0.02	0.15	0.003	0.112
agrLognΔ	2222	0.05	0.27	0.012	0.000
agrUniΔ	2222	0.12	0.66	0.017	0.000
agrLognGiniΔ	2222	0.04	0.05	0.037	0.000
agrUniGiniΔ	2222	0.03	0.03	0.023	0.000
Industrial subsample					
indMidpointΔ	1328	0.06	0.11	0.048	0.000
indLognΔ	1328	0.08	0.12	0.062	0.000
indUniΔ	1328	0.12	0.18	0.097	0.000
indLognGiniΔ	1328	0.14	0.16	0.129	0.000
indUniGiniΔ	1328	0.11	0.12	0.097	0.000

* The p-values refer to a sign test on the median, with the null hypothesis that it is zero.

Table C.8: Correlations, permutations of measures used

Variable pair		n	Spearman	Pearson
Agriculture vs Industry				
indLogn1871	agrMidpoint1871	1349	0.10*	0.13*
indMidpoint1871	agrLogn1871	1349	0.26*	0.28*
indLogn1892	agrMidpoint1892	2020	0.33*	0.44*
indMidpoint1892	agrLogn1892	2020	0.40*	0.37*
Agriculture 1871 vs 1892				
agrLogn1871	agrMidpoint1892	2274	0.75*	0.61*
agrMidpoint1871	agrLogn1892	2236	0.66*	0.41*
Industry 1871 vs 1892				
indLogn1871	indMidpoint1892	1361	0.55*	0.54*
indMidpoint1871	indLogn1892	1338	0.58*	0.59*
Agriculture 1871 vs Industry 1892				
agrLogn1871	indMidpoint1892	2018	0.36*	0.32*
agrMidpoint1871	indLogn1892	1989	0.24*	0.26*
Industry 1871 vs Agriculture 1892				
indLogn1871	agrMidpoint1892	1365	0.29*	0.39*
indMidpoint1871	agrLogn1892	1343	0.37*	0.36*
Agriculture 1871 vs Change				
agrLogn1871	agrMidpointΔ	2222	-0.04*	0.06*
agrMidpoint1871	agrLognΔ	2236	-0.12*	-0.01
Industry 1871 vs Change				
indLogn1871	indMidpointΔ	1328	-0.28*	-0.29*
indMidpoint1871	indLognΔ	1338	-0.09*	-0.07*
Agriculture 1871 vs Change in Industry				
agrLogn1871	indMidpointΔ	1311	0.09*	0.07*
agrMidpoint1871	indLognΔ	1321	0.18*	0.20*
Industry 1871 vs Change in Agriculture				
indLogn1871	agrMidpointΔ	1329	0.32*	0.38*
indMidpoint1871	agrLognΔ	1339	0.26*	0.28*

Note * indicate significance at least at the 5 percent level.

Lund Economic Studies

1. Guy Arvidsson	Bidrag till teorin för verkningarna av räntevariationer, 1962
2. Björn Thalberg	A Trade Cycle Analysis. Extensions of the Goodwin Model, 1966
3. Bengt Höglund	Modell och observationer. En studie av empirisk anknytning och aggregation för en linjär produktionsmodell, 1968
4. Alf Carling	Industrins struktur och konkurrensförhållanden, 1968
5. Tony Hagström	Kreditmarknadens struktur och funktionssätt, 1968
6. Göran Skogh	Straffrätt och samhällsekonomi, 1973
7. Ulf Jakobsson och Göran Norman	Inkomstbeskattningen i den ekonomiska politiken. En kvantitativ analys av systemet för personlig inkomstbeskattning 1952-71, 1974
8. Eskil Wadensjö	Immigration och samhällsekonomi. Immigrationens ekonomiska orsaker och effekter, 1973
9. Rögnvaldur Hannesson	Economics of Fisheries. Some Problems of Efficiency, 1974
10. Charles Stuart	Search and the Organization of Marketplaces, 1975
11. S Enone Metuge	An Input-Output Study of the Structure and Resource Use in the Cameroon Economy, 1976
12. Bengt Jönsson	Cost-Benefit Analysis in Public Health and Medical Care, 1976
13. Agneta Kruse och Ann-Charlotte Ståhlberg	Effekter av ATP - en samhällsekonomisk studie, 1977
14. Krister Hjalte	Sjörestaureringens ekonomi, 1977
15. Lars-Gunnar Svensson	Social Justice and Fair Distributions, 1977
16. Curt Wells	Optimal Fiscal and Monetary Policy - Experiments with an Econometric Model of Sweden, 1978
17. Karl Lidgren	Dryckesförpackningar och miljöpolitik - En studie av styrmedel, 1978
18. Mats Lundahl	Peasants and Poverty. A Study of Haiti, London, 1979
19. Inga Persson-Tanimura	Studier kring arbetsmarknad och information, 1980
20. Bengt Turner	Hyressättning på bostadsmarknaden - Från hyresreglering till bruksvärdesprövning, Stockholm 1979
21. Ingemar Hansson	Market Adjustment and Investment Determination. A Theoretical Analysis of the Firm and the Industry, Stockholm 1981
22. Daniel Boda Ndlela	Dualism in the Rhodesian Colonial Economy, 1981
23. Tom Alberts	Agrarian Reform and Rural Poverty: A Case Study of Peru, 1981
24. Björn Lindgren	Costs of Illness in Sweden 1964-75, 1981

25. Göte Hansson	Social Clauses and International Trade. An Economic Analysis of Labour Standards in Trade Policy, 1981
26. Noman Kanafani	Oil and Development. A Case Study of Iraq, 1982
27. Jan Ekberg	Inkomsteffekter av invandring, 1983
28. Stefan Hedlund	Crisis in Soviet Agriculture?, 1983
29. Ann-Marie Pålsson	Hushållen och kreditpolitiken. En studie av kreditrestriktioners effekt på hushållens konsumtion, sparande och konsumtionsmönster, 1983
30. Lennart Petersson	Svensk utrikeshandel, 1871-1980. En studie i den intraindustriella handelns framväxt, 1984
31. Bengt Assarsson	Inflation and Relative Prices in an Open Economy, 1984
32. Claudio Vedovato	Politics, Foreign Trade and Economic Development in the Dominican Republic, 1985
33. Knut Ödegaard	Cash Crop versus Food Crop Production in Tanzania: An Assessment of the Major Post-Colonial Trends, 1985
34. Vassilios Vlachos	Temporära lönesubventioner. En studie av ett arbetsmarknadspolitiskt medel, 1985
35. Stig Tegle	Part-Time Employment. An Economic Analysis of Weekly Working Hours in Sweden 1963-1982, 1985
36. Peter Stenkula	Tre studier över resursanvändningen i högskolan, 1985
37. Carl Hampus Lyttkens	Swedish Work Environment Policy. An Economic Analysis, 1985
38. Per-Olof Bjuggren	A Transaction Cost Approach to Vertical Integration: The Case of Swedish Pulp and Paper Industry, 1985
39. Jan Petersson	Erik Lindahl och Stockholmsskolans dynamiska metod, 1987
40. Yves Bourdet	International Integration, Market Structure and Prices. A Case Study of the West-European Passenger Car Industry, 1987
41. Krister Andersson and Erik Norrman	Capital Taxation and Neutrality. A study of tax wedges with special reference to Sweden, 1987
42. Tohmas Karlsson	A Macroeconomic Disequilibrium Model. An Econometric Study of the Swedish Business Sector 1970-84, 1987
43. Rosemary Vargas-Lundius	Peasants in Distress. Poverty and Unemployment in the Dominican Republic, 1989
44. Lena Ekelund Axelson	Structural Changes in the Swedish Marketing of Vegetables, 1991
45. Elias Kazarian	Finance and Economic Development: Islamic Banking in Egypt, 1991
46. Anders Danielson	Public Sector Expansion and Economic Development. The Sources and Consequences of Development Finance in Jamaica 1962-84, 1991

47. Johan Torstensson	Factor Endowments, Product Differentiation, and International Trade, 1992
48. Tarmo Haavisto	Money and Economic Activity in Finland, 1866-1985, 1992
49. Ulf Grönkvist	Economic Methodology. Patterns of Reasoning and the Structure of Theories, 1992
50. Evelyne Hangali Maje	Monetization, Financial Development and the Demand for Money, 1992
51. Michael Bergman	Essays on Economic Fluctuations, 1992
52. Flora Mndeme Musonda	Development Strategy and Manufactured Exports in Tanzania, 1992
53. Håkan J. Holm	Complexity in Economic Theory. An Automata Theoretical Approach, 1993
54. Klas Fregert	Wage Contracts, Policy Regimes and Business Cycles. A Contractual History of Sweden 1908-90, 1994
55. Per Frennberg	Essays on Stock Price Behaviour in Sweden, 1994
56. Lisbeth Hellvin	Trade and Specialization in Asia, 1994
57. Sören Höjgård	Long-term Unemployment in a Full Employment Economy, 1994
58. Karolina Ekholm	Multinational Production and Trade in Technological Knowledge, 1995
59. Fredrik Andersson	Essays in the Economics of Asymmetric Information, 1995
60. Rikard Althin	Essays on the Measurement of Producer Performance, 1995
61. Lars Nordén	Empirical Studies of the Market Microstructure on the Swedish Stock Exchange, 1996
62. Kristian Bolin	An Economic Analysis of Marriage and Divorce, 1996
63. Fredrik Sjöholm	R&D, International Spillovers and Productivity Growth, 1997
64. Hossein Asgharian	Essays on Capital Structure, 1997
65. Hans Falck	Aid and Economic Performance - The Case of Tanzania, 1997
66. Bengt Liljas	The Demand for Health and the Contingent Valuation Method, 1997
67. Lars Pålsson Syll	Utility Theory and Structural Analysis, 1997
68. Richard Henricsson	Time Varying Parameters in Exchange Rate Models, 1997
69. Peter Hördahl	Financial Volatility and Time-Varying Risk Premia, 1997
70. Lars Nilsson	Essays on North-South Trade, 1997
71. Fredrik Berggren	Essays on the Demand for Alcohol in Sweden - Review and Applied Demand Studies, 1998
72. Henrik Braconier	Essays on R&D, Technology and Growth, 1998
73. Jerker Lundbäck	Essays on Trade, Growth and Exchange Rates, 1998
74. Dan Anderberg	Essays on Pensions and Information, 1998

75. P. Göran T. Hägg	An Institutional Analysis of Insurance Regulation – The Case of Sweden, 1998
76. Hans-Peter Bermin	Essays on Lookback and Barrier Options - A Malliavin Calculus Approach, 1998
77. Kristian Nilsson	Essays on Exchange Rates, Exports and Growth in Developing Countries, 1998
78. Peter Jochumzen	Essays on Econometric Theory, 1998
79. Lars Behrenz	Essays on the Employment Service and Employers' Recruitment Behaviour, 1998
80. Paul Nystedt	Economic Aspects of Ageing, 1998
81. Rasha M. Torstensson	Empirical Studies in Trade, Integration and Growth, 1999
82. Mattias Ganslandt	Games and Markets - Essays on Communication, Coordination and Multi-Market Competition, 1999
83. Carl-Johan Belfrage	Essays on Interest Groups and Trade Policy, 1999
84. Dan-Olof Rooth	Refugee Immigrants in Sweden - Educational Investments and Labour Market Integration, 1999
85. Karin Olofsdotter	Market Structure and Integration: Essays on Trade, Specialisation and Foreign Direct Investment, 1999
86. Katarina Steen Carlsson	Equality of Access in Health Care, 1999
87. Peter Martinsson	Stated preference methods and empirical analyses of equity in health, 2000
88. Klas Bergenheim	Essays on Pharmaceutical R&D, 2000
89. Hanna Norberg	Empirical Essays on Regional Specialization and Trade in Sweden, 2000
90. Åsa Hansson	Limits of Tax Policy, 2000
91. Hans Byström	Essays on Financial Markets, 2000
92. Henrik Amilon	Essays on Financial Models, 2000
93. Mattias Lundbäck	Asymmetric Information and The Production of Health, 2000
94. Jesper Hansson	Macroeconometric Studies of Private Consumption, Government Debt and Real Exchange Rates, 2001
95. Jonas Månsson	Essays on: Application of Cross Sectional Efficiency Analysis, 2001
96. Mattias Persson	Portfolio Selection and the Analysis of Risk and Time Diversification, 2001
97. Pontus Hansson	Economic Growth and Fiscal Policy, 2002
98. Joakim Gullstrand	Splitting and Measuring Intra-Industry Trade, 2002
99. Birger Nilsson	International Asset Pricing, Diversification and Links between National Stock Markets, 2002
100. Andreas Graflund	Financial Applications of Markov Chain Monte Carlo Methods, 2002

101. Therése Hindman Persson	Economic Analyses of Drinking Water and Sanitation in Developing Countries, 2002
102. Göran Hjelm	Macroeconomic Studies on Fiscal Policy and Real Exchange Rates, 2002
103. Klas Rikner	Sickness Insurance: Design and Behavior, 2002
104. Thomas Ericson	Essays on the Acquisition of Skills in Teams, 2002
105. Thomas Elger	Empirical Studies on the Demand for Monetary Services in the UK, 2002
106. Helena Johansson	International Competition, Productivity and Regional Spillovers, 2003
107. Fredrik Gallo	Explorations in the New Economic Geography, 2003
108. Susanna Thede	Essays on Endogenous Trade Policies, 2003
109. Fredrik CA Andersson	Interest Groups and Government Policy, A Political Economy Analysis, 2003
110. Petter Lundborg	Risky Health Behaviour among Adolescents, 2003
111. Martin W Johansson	Essays on Empirical Macroeconomics, 2003
112. Joakim Ekstrand	Currency Markets - Equilibrium and Expectations, 2003
113. Ingemar Bengtsson	Central bank power: a matter of coordination rather than money supply, 2003
114. Lars Pira	Staples, Institutions and Growth: Competitiveness of Guatemalan Exports 1524-1945, 2003
115. Andreas Bergh	Distributive Justice and the Welfare State, 2003
116. Staffan Waldo	Efficiency in Education - A Multilevel Analysis, 2003
117. Mikael Stenkula	Essays on Network Effects and Money, 2004
118. Catharina Hjortsberg	Health care utilisation in a developing country -the case of Zambia, 2004
119. Henrik Degrér	Empirical Essays on Financial Economics, 2004
120. Mårten Wallette	Temporary Jobs in Sweden: Incidence, Exit, and On-the-Job Training, 2004
121. Tommy Andersson	Essays on Nonlinear Pricing and Welfare, 2004
122. Kristian Sundström	Moral Hazard and Insurance: Optimality, Risk and Preferences, 2004
123. Pär Torstensson	Essays on Bargaining and Social Choice, 2004
124. Frederik Lundtofte	Essays on Incomplete Information in Financial Markets, 2005
125. Kristian Jönsson	Essays on Fiscal Policy, Private Consumption and Non-Stationary Panel Data, 2005
126. Henrik Andersson	Willingness to Pay for a Reduction in Road Mortality Risk: Evidence from Sweden, 2005

127. Björn Ekman	Essays on International Health Economics: The Role of Health Insurance in Health Care Financing in Low- and Middle-Income Countries, 2005
128. Ulf G Erlandsson	Markov Regime Switching in Economic Time Series, 2005
129. Joakim Westerlund	Essays on Panel Cointegration, 2005
130. Lena Hiselius	External costs of transports imposed on neighbours and fellow road users, 2005
131. Ludvig Söderling	Essays on African Growth, Productivity, and Trade, 2005
132. Åsa Eriksson	Testing and Applying Cointegration Analysis in Macroeconomics, 2005
133. Fredrik Hansen	Explorations in Behavioral Economics: Realism, Ontology and Experiments, 2006
134. Fadi Zaher	Evaluating Asset-Pricing Models in International Financial Markets, 2006
135. Christoffer Bengtsson	Applications of Bayesian Econometrics to Financial Economics, 2006
136. Alfredo Schclarek Curutchet	Essays on Fiscal Policy, Public Debt and Financial Development, 2006
137. Fredrik Wilhelmsson	Trade, Competition and Productivity, 2006
138. Ola Jönsson	Option Pricing and Bayesian Learning, 2007
139. Ola Larsson	Essays on Risk in International Financial Markets, 2007
140. Anna Meyer	Studies on the Swedish Parental Insurance, 2007
141. Martin Nordin	Studies in Human Capital, Ability and Migration, 2007
142. Bolor Naranhuu	Studies on Poverty in Mongolia, 2007
143. Margareta Ekbladh	Essays on Sickness Insurance, Absence Certification and Social Norms, 2007
144. Erik Wengström	Communication in Games and Decision Making under Risk, 2007
145. Robin Rander	Essays on Auctions, 2008
146. Ola Andersson	Bargaining and Communication in Games, 2008
147. Marcus Larson	Essays on Realized Volatility and Jumps, 2008
148. Per Hjertstrand	Testing for Rationality, Separability and Efficiency, 2008
149. Fredrik NG Andersson	Wavelet Analysis of Economic Time Series, 2008
150. Sonnie Karlsson	Empirical studies of financial asset returns, 2009
151. Maria Persson	From Trade Preferences to Trade Facilitation, 2009
152. Eric Rehn	Social Insurance, Organization and Hospital Care, 2009
153. Peter Karpestam	Economics of Migration, 2009
154. Marcus Nossman	Essays on Stochastic Volatility, 2009
155. Erik Jonasson	Labor Markets in Transformation: Case Studies of Latin America, 2009

156. Karl Larsson	Analytical Approximation of Contingent Claims, 2009
157. Therese Nilsson	Inequality, Globalization and Health, 2009
158. Rikard Green	Essays on Financial Risks and Derivatives with Applications to Electricity Markets and Credit Markets, 2009
159. Christian Jörgensen	Deepening Integration in the Food Industry – Prices, Productivity and Export, 2010
160. Wolfgang Hess	The Analysis of Duration and Panel Data in Economics, 2010
161. Pernilla Johansson	From debt crisis to debt relief: A study of debt determinants, aid composition and debt relief effectiveness, 2010
162. Nils Janlöv	Measuring Efficiency in the Swedish Health Care Sector, 2010
163. Ai Jun Hou	Essays on Financial Markets Volatility, 2011
164. Alexander Reffgen	Essays on Strategy-proof Social Choice, 2011
165. Johan Blomquist	Testing homogeneity and unit root restrictions in panels, 2012
166. Karin Bergman	The Organization of R&D - Sourcing Strategy, Financing and Relation to Trade, 2012
167. Lu Liu	Essays on Financial Market Interdependence, 2012
168. Bujar Huskaj	Essays on VIX Futures and Options, 2012
169. Åsa Ljungvall	Economic perspectives on the obesity epidemic, 2012
170. Emma Svensson	Experimenting with Focal Points and Monetary Policy, 2012
171. Jens Dietrichson	Designing Public Organizations and Institutions: Essays on Coordination and Incentives, 2013
172. Thomas Eriksson	Empirical Essays on Health and Human Capital, 2013
173. Lina Maria Ellegård	Political Conflicts over Public Policy in Local Governments, 2013
174. Andreas Hatzigeorgiou	Information, Networks and Trust in the Global Economy - Essays on International Trade and Migration, 2013
175. Gustav Kjellsson	Inequality, Health, and Smoking, 2014
176. Richard Desjardins	Rewards to skill supply, skill demand and skill match-mismatch: Studies using the Adult Literacy and Lifeskills survey, 2014
177. Viroj Jienwatcharamongkhol	What Drives Exports? Empirical Evidence at the Firm Level, 2014
178. Anton Nilsson	Health, Skills and Labor Market Success, 2014
179. Albin Erlanson	Essays on Mechanism Design, 2014
180. Daniel Ekeblom	Essays in Empirical Expectations, 2014

181. Sofie Gustafsson	Essays on Human Capital Investments: Pharmaceuticals and Education, 2014
182. Katarzyna Burzynska	Essays on Finance, Networks and Institutions, 2015
183. Mingfa Ding	Corporate Ownership and Liquidity in China's Stock Markets, 2015
184. Anna Andersson	Vertical Trade, 2015
185. Cecilia Hammarlund	Fish and Trips in the Baltic Sea - Prices, Management and Labor Supply, 2015
186. Hilda Ralsmark	Family, Friend, or Foe? Essays in Empirical Microeconomics, 2015
187. Jens Gudmundsson	Making Pairs, 2015
188. Emanuel Alfranseder	Essays on Financial Risks and the Subprime Crisis, 2015
189. Ida Lovén	Education, Health, and Earnings – Type 1 Diabetes in Children and Young Adults, 2015
190. Caren Yinxia Nielsen	Essays on Credit Risk, 2015
191. Usman Khalid	Essays on Institutions and Institutional change, 2016
192. Ross Wilson	Essays in Empirical Institutional Analysis, 2016
193. Milda Norkute	A Factor Analytical Approach to Dynamic Panel Data Models, 2016
194. Valeriia Dzhamalova	Essays on Firms' Financing and Investment Decisions, 2016
195. Claes Ek	Behavioral Spillovers across Prosocial Alternatives, 2016
196. Graeme Cokayne	Networks, Information and Economic Volatility, 2016
197. Björn Thor Arnarson	Exports and Externalities, 2016
198. Veronika Lunina	Multivariate Modelling of Energy Markets, 2017
199. Patrik Karlsson	Essays in Quantitative Finance, 2017
200. Hassan Sabzevari	Essays on systemic risk in European banking, 2017
201. Margaret Samahita	Self-Image and Economic Behavior, 2017
202. Aron Berg	Essays on informational asymmetries in mergers and acquisitions, 2017
203. Simon Reese	Estimation and Testing in Panel Data with Cross-Section Dependence, 2017
204. Karl McShane	Essays on Social Norms and Economic Change, 2017
205. Elvira Andersson	From Cradle to Grave: Empirical Essays on Health and Economic Outcomes, 2017
206. Yana Pryymachenko	Heavy Metal Exposure in Early Life - Health and Labour Market Perspectives, 2017
207. Alemu Tulu Chala	Essays on Banking and Corporate Finance, 2017
208. Jim Ingebretsen Carlson	Essays on economic behavior, focusing and auctions, 2018
209. Jörgen Kratz	Essays on Matching, 2018

210. Anna Welander Tärneberg	Essays on Health in Developing countries, 2018
211. Osmis Areda Habte	Essays on competition and consumer choice, 2018
212. Thomas Hofmarcher	Essays in Empirical Labor Economics, 2019
213. Hjördis Hardardottir	Time and inequality – A study of individual preferences, 2019
214. Erik Grenestam	Essays in Applied Microeconomics, 2019
215. Sara Moricz	Institutions, Inequality and Societal Transformations, 2019