NORTHWESTERN UNIVERSITY

# Essays on Macroeconomic Aspects of Economic Development

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Kristina Manysheva

EVANSTON, ILLINOIS

June 2022

© Copyright by Kristina Manysheva 2022
All Rights Reserved

# Abstract

This dissertation is divided into three chapters. Chapter one examines the effects of weak land property rights and limited access to finance on aggregate productivity and the allocation of resources, as well as the role of their interaction in the context of a developing country – Tanzania. Chapter two studies the evolution of innovation across time and space and its effect on productivity using a panel of historical patent data covering a large range of countries over the past century. Chapter three examines how women's employment was effected by the Covid-19 pandemic in developing countries with a focus on Nigeria, the most populous country in Africa.

# Acknowledgements

This dissertation would not have been possible without the support, both intellectual and emotional, from a long list of people. I am first exceedingly grateful for the advice, insight, encouragement, and infinite support of my advisors Matthias Doepke, Martí Mestieri, and Christopher Udry. I could not have imagined having better advisors and mentors for my PhD study.

I also benefited enormously throughout the years from discussions with other Northwestern faculty members including Benjamin Jones, Guido Lorenzoni, Kiminori Matsuyama, Dimitris Papanikolaou, Giorgio Primiceri, and Matthew Rognlie. Many thanks to all the Northwestern economics faculty members for imparting wisdom during coursework, providing feedback in seminars, and assisting with preparation for the job market.

I made substantial progress on this dissertation during visits to the International Monetary Fund in 2019 and the Federal Reserve Bank in Chicago in 2020. My sincere thanks to everyone in both groups for hosting me, especially Chris Papageorgiou at the IMF and Gadi Barlevy, Marco Bassetto, Francois Gourio, and Martí Mestieri at the Chicago Fed.

It was a pleasure and a wonderful opportunity to learn alongside other students at Northwestern, but especially Bence Bardóczy, Enrico Berkes, Ana Danieli, Samuel Goldberg, Carl Hallmann, Fergal Hanks, Egor Kozlov, Joris Mueller. Laura Murphy, Jane Olmstead-Rumsey, Richard Peck, Divid Stillerman, Emre Enes Yavuz, and Brecia Young.

My Master's degree professor Peter Montiel first recommended I undertake a PhD, and I could not be more grateful for his early investments in me and his continued support throughout all these years. More teachers and mentors than I can count have contributed to my success, curiosity, and knowledge over the years and I'm grateful to each and every one of them.

Last but certainly not least, I'm grateful to my friends and family who loved me, supported me, cheered me on, and patiently went through all my exams and presentations. I want to especially thank my parents Svetlana and Valeri, my sister Natallia, my brother-in-law Patrick, and my nephews Daniel and Michael for being the most supportive family I could ever ask for.

# Preface

The three essays composing this dissertation are unified by their focus on the macroeconomic aspects of economic development. In Chapter one titled "Land Property Rights, Financial Frictions, and Resource Allocation in Developing Countries", I study the effects of weak land property rights and limited access to finance on aggregate productivity and the allocation of resources, as well as the role of their interaction. To do that, I develop a dynamic general equilibrium model and use it to quantify the aggregate and distributional impacts of land and financial market imperfections connected via the collateral channel. I discipline the model with longitudinal micro data from Tanzania and show that substantial frictions in land and financial markets affect resource allocation and economic efficiency in agriculture. In the model, these distortions reduce aggregate productivity by allocating land and capital to less efficient producers, and by preventing households from moving out of agriculture and limiting entrepreneurship. An economy-wide land reform that improves land property rights leads to increases in agricultural and non-agricultural output by 7.4% and 8.2%, respectively, as well as a decline in agricultural employment by 8.6%. A land reform also results in higher financial inclusion, especially among the poorest, as land market frictions amplify the effects of financial markets imperfections. While a financial reform can deliver comparable aggregate effects, land reform is more pro-poor and reduces consumption inequality.

In Chapter two, which is a joint work with Enrico Berkes and Martí Mestieri titled "Global Innovation Spillovers and Productivity: Evidence from 100 Years of World Patent Data", we use a panel of historical patent data covering a large range of countries over the past century to study the evolution of innovation across time and space and its effect on productivity. We document a substantial rise of international knowledge spillovers as measured by patent citations since the 1990s. This rise is mostly accounted for by an increase in citations to US and Japanese patents in fields of knowledge related to computation, information processing, and medicine. We estimate the causal effect of innovation induced by international spillovers on sectoral output per worker and total factor productivity (TFP) growth in a panel of country-sectors from 2000 to 2014, as well as on aggregate income per capita since 1960. To assess causality, we develop a shift-share instrument that leverages pre-existing citation linkages across countries and fields of knowledge, as well as heterogeneous countries' exposure to technology waves. On average, an increase of one standard deviation in log-patenting activity increases sectoral output per worker growth by 1.1 percentage points. We find results of similar magnitude for sectoral TFP growth and long-run aggregate income per capita growth.

In Chapter three, which is a joint effort with Titan Alon, Matthias Doepke, and Michèle Tertilt titled "Gendered Impacts of Covid-19 in Developing Countries" we examine whether the fact that in many high-income economies, the recession caused by the Covid-19 pandemic has resulted in unprecedented declines in women's employment, took place in developing countries. We focus our study on Nigeria, the most populous country in Africa. A force affecting high- and low-income countries alike are increased childcare needs during school closures; in Nigeria, mothers of school-age children experience the largest declines in employment during the pandemic, just as in high-income countries. A key difference is the role of the sectoral distribution of employment: whereas in high-income economies reduced employment in contact-intensive services had a large impact on women, this sector plays a minor role in low-income countries. Another difference is that women's employment rebounded much more quickly in low-income countries. We conjecture that large income losses without offsetting government transfers drive up labor supply in low-income countries during the recovery.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Land Property Rights, Financial Frictions, and Resource Allocation in Developing Countries[1]

## 1.1 Introduction

One of the leading explanations for persistent economic disparities between advanced and developing countries is that low-income countries are less effective in allocating their resources to their most productive use. Widespread market imperfections, including incomplete land and financial markets, are recognized as a potential reason for such misallocation.[2] Many developing countries are characterized not only by a low level of financial development (King and Levine, 1993; Banerjee and Duflo, 2005) but also by limited land markets and insecure land property rights (Adamopoulos and Restuccia, 2014). There are two main reasons for such land market imperfections in low-income countries. First, a large share of land does not have any documentation. Second, the land tenure system in many developing countries is

[1]I thank my advisors Matthias Doepke, Martí Mestieri, and Christopher Udry for a fantastic amount of help and support as part of my dissertation committee. I also thank Bence Bardoczy, Gadi Barlevy, Marco Bassetto, Ana Danieli, Francois Gourio, Egor Kozlov, Jane Olmstead-Rumsey, Chris Papageorgiou, Fernanda Rojas, as well as all attendees of the IMF RESDM Divisional Seminar, Chicago Fed Macro Seminar, STEG Annual Conference, CSAE Conference, BREAD Conference on the economics of Africa, Young Economist Symposium, and NU Macro and Development lunch for helpful comments and discussions. I thank IMF and Chicago Fed for their hospitality during part of this research. All errors are my own.

[2]See Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Restuccia and Rogerson (2013) and Hopenhayn (2014) for the review of the expanding literature on misallocation.

ruled by customary law, particularly in Sub-Saharan Africa (Pande and Udry, 2005). Such customary tenure is a set of rules and norms that govern the allocation, use, access, and transfer of land within communities. These rules include the common "use it or lose it" principle, which means that whoever farms the land can continue using it, but if they do not cultivate the land in a given year, they can lose their rights, and land will be reallocated to someone else.[3]

In this paper, I study the interaction between weak land property rights and limited access to credit and their effect on aggregate productivity and allocation of resources. My paper has two main contributions. First, I develop a heterogeneous-agent dynamic macro model to quantify the aggregate and distributional impact of land and financial market imperfections. The framework incorporates both financial and land market frictions that are connected via the collateral channel. This novel feature of my model enables me to study the interaction of these two markets in a general equilibrium setting. Second, I use household-level data from Tanzania to discipline the model and to show that substantial frictions in both the land and credit markets affect resource allocation and economic efficiency in agriculture in Tanzania. I argue that these imperfections reduce aggregate productivity in the economy by affecting two critical margins: the allocation of factors of productions across households and sectors, and the allocation of households across different occupations.

Empirically, I exploit longitudinal micro data from the Tanzania National Panel Survey (2008-2015), which has a special focus on agricultural production. I use a dynamic panel approach to estimate an agricultural production function. My results imply that agriculture in Tanzania is still mainly labor- and land-intensive and exhibits decreasing returns to scale. I then use these estimates to obtain farmer-level TFP measures. Combining these

[3]Up to 70 percent of land in some low-income countries has no formal or informal documentation (Figure A.1). The percent of communal land in Africa varies from 2 in Rwanda to 97 in Somalia. The statistically significant correlation between land security and level of traditional land suggests that countries with a higher level of communal land feature lower land security (Figure A.3).

productivity measures with the variation in land property rights and access to credit both across households and across time, I test for the efficiency of resource allocation. In this setting, efficient land allocation is proportional to the farmer's productivity. Moreover, efficient allocation requires that the relationship between farm size and farmer's TFP is identical across farmers. However, in the data, I find that such a relationship depends on whether the household's land is under a strong property rights regime and whether the household uses credit for agricultural purposes. Such results suggest that land is not allocated efficiently, and land misallocation is associated with insecure land property rights and limited access to credit. In addition, I find that households that have titled land are more likely to use credit for agricultural purposes and enjoy a larger loan size conditional on being given one. Finally, there is a link between land property rights and occupational choice. Households with titled land are less likely to stay in agriculture and more likely to operate a non-farm enterprise.

I use these empirical findings to discipline a heterogeneous agent incomplete-markets model that incorporates endogenous saving decisions, occupational choice, and communal land evolution. Agents are heterogeneous in their wealth, productivity levels in agriculture and entrepreneurship, and land holdings under either private or communal property rights. Following the main channels of how property rights affect economic activity described in Besley and Ghatak (2010), I incorporate three land market imperfections for communal land: i) it cannot be rented out, ii) it is subject to expropriation risk if it is not used, and iii) it cannot be used as collateral. On the financial side, borrowing is subject to a limit, which is a function of a household's financial wealth, land holdings, and land property rights. The presence of financial market frictions and the inability to use communal land as collateral prevents households without legal land titles that are poor in terms of financial assets from obtaining a loan.

To quantify the effects of potential improvement in land property rights and access to credit, I calibrate the model to Tanzania and perform three sets of counterfactual exercises.

First, I show that an economy-wide land reform that converts communal into private land has a positive effect on agricultural and non-agricultural output as well as total consumption. As a result of the reform, agricultural output increases by 7.4%, driven mainly by higher land utilization and a more efficient land allocation across households. Non-agricultural output increases by 8.2% due to higher access to credit and a more efficient allocation of households across occupations. Land reform leads to changes in labor composition in favor of non-agricultural employment (entrepreneurs and workers), with agricultural employment declining by 8.6%.

I also find that despite substantial welfare gains of land reform for the economy, these gains are not evenly distributed. Welfare gains, measured in consumption equivalent changes, are the highest for those belonging to the communal part of the economy before the reform. These welfare gains are particularly high for those with a low level of financial assets, significant land holdings, and a high level of entrepreneurial skills. Substantial welfare gains are driven by higher financial inclusion as a result of the reform, especially among the poorest households with limited assets but positive land holdings. On the other hand, large private landholders are the main losers of the reform, suggesting that political economy barriers might prevent or slow the progress of land reform in many low-income countries, despite its potential benefits.

In my second counterfactual, I perform a decomposition analysis of the role played by the three communal land market imperfections. To do that, I look at the general equilibrium impact of a policy change that eliminates only one communal land friction at a time.[4] Each channel has a distinct impact on equilibrium prices and average productivity in each sector. I find that the increase in agricultural output is driven mainly by the ability of communal landholders to rent out their unused land. This increase happens as land is reallocated from

[4] Recall, that communal land i) cannot be rented out, ii) is subject to expropriation risk if it is not used iii) cannot be used as collateral.

less to more productive farmers leading to higher agricultural productivity. In addition, the ability to rent out communal land increases land utilization and therefore results in larger land input in agricultural production. By contrast, the increase in non-agricultural production results from eliminating expropriation risk and the ability to use the land as collateral. Such growth is driven by a larger number of entrepreneurs, as well as by the higher labor and capital inputs of these entrepreneurs.

Third, I compare the aggregate and distributional consequences of land reform with the effects of financial reform. To compute the impact of financial reform, I relax the financial constraint so that the loan to collateral value is equal to the level of an advanced economy. I find that the qualitative impact of financial reform on economic outcomes is the same as the impact of the collateral channel of land reform but differs from land reform as a whole. Moreover, distributional consequences are different. In the case of financial reform, marginal entrepreneurs and large asset owners benefit the most. In contrast, those operating communal land do not benefit as much as in the case of land reform. Finally, land reform leads to a lower level of consumption inequality compared to financial reform. This happens as a large share of welfare winners of land reform is among the poorest part of the population before the reform.

I conclude my quantitative analysis by studying the transitional dynamic triggered by a sudden unexpected land reform that removes all land market frictions. I find that most changes happen in the first ten years after the reform, with a substantial initial increase in agricultural and non-agricultural output. Additional adjustment occurs later in transition driven by changes in prices and level of asset accumulation.

**Related Literature.** This paper contributes to two main strands of literature. First, I relate to the literature quantifying the importance of misallocation for aggregate outcomes (e.g. Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009; Bartelsman et al., 2013; Restuccia and Rogerson, 2013; Baqaee and Farhi, 2020), especially in the context of developing countries (e.g. Guner et al., 2008; Banerjee and Moll, 2010; Asker et al., 2011; Oberfield, 2013; Kalemli-Ozcan and Sorensen, 2012; Restuccia and Rogerson, 2017, Bau and Matray (2020)) and with a focus on productivity in the agricultural sector (e.g. Chen, 2017; Adamopoulos et al., 2017; Restuccia and Santaeulalia-Llopis, 2017). Second, it contributes to the literature in macroeconomics using micro data to study macro development issues such as Gollin et al. (2014), Buera et al. (2014), Bick et al. (2016), Santaeulàlia-Llopis and Zheng (2016), Adamopoulos and Restuccia (2020), Buera et al. (2021) among others.

A large share of the misallocation literature focuses on measuring the effect of all sources of misallocation on aggregate output by exploiting cross-sectional dispersion in marginal revenue products without identifying the underlying sources of the distortions. The contribution of my paper is that I not only show the presence of resource misallocation but also link it to specific market distortions. I also measure misallocation under weaker assumptions than some earlier work. Specifically, I estimate the production function instead of assuming that the U.S. parameters can be applied to an African economy. Additionally, I show that my results are robust to alternative production function specifications.

My findings are consistent with the literature that links land property rights to economic outcomes. de Janvry et al. (2015) document that formal land titling enabled a market-based reallocation through sales and rentals to more productive farmers. Beg (2021) provide the evidence that computerized rural land records in Pakistan result in landowning households being more likely to rent out land and to shift into non-agricultural occupations. Consistent with quantitative results of my paper, Chari et al. (2017) find that a land reform in rural China that allowed farmers to lease out their land resulted in a redistribution of land toward more productive farmers and an increase in agricultural output by 8%.[5]

[5] Other work on land property rights and economic outcomes includes Field (2007), Di Tella et al. (2007), Bromley (2010), Macours et al. (2010), and de Brauw and Mueller (2012).

My paper is most closely related to the growing literature that uses micro data and macro models to study the role of different institutions and policies in structural transformation, particularly that focusing on land market institutions. Chen (2017), Adamopoulos et al. (2017), and Restuccia and Santaeulalia-Llopis (2017) use micro data to back out farm-specific TFP and wedges in Ethiopia, China and Malawi, respectively. In all these papers, removing wedges to shift land to more productive farmers brings large gains in aggregate agricultural productivity. Gottlieb and Grobovsek (2019) measure the distortionary impact of land expropriation risk under communal land tenure using dynamic general equilibrium model calibrated to Ethiopia, and find that lifting communal land tenure increases GDP by 9%.[6]

I add to this literature in several ways. First, land market imperfections in my model affect economic outcomes through multiple channels. This allows to perform quantitative analysis of economy-wide land reform that improves property rights and study the implications of different channels of such reform, focusing on each land market friction in isolation. Following the previous literature, I include in the model both the inability to rent out communal land (Chen, 2017) and presence of expropriation risk (Gottlieb and Grobovsek, 2019; Ngai et al., 2019).[7] I also add the third market imperfection – the inability to use land as collateral. Second, this paper includes both financial and land market frictions connected via the collateral channel in a macroeconomic model of growth and development. I show that land market frictions amplify the negative impact of limited access to credit, especially for the poorest part of the population.

At the same time, the presence of financial market imperfections might limit the benefits

[6]Adamopoulos et al. (2017) find that misallocation of land leads to misallocation of workers across different sectors. Adamopoulos and Restuccia (2020) study land reform in the Philippines and find that imposed ceiling on land holdings reduced agricultural productivity by 17 percent.

[7]Chen (2017) build a two-sector general equilibrium model to quantify the impact of untitled land, which cannot be rented in the market. Gottlieb and Grobovsek (2019) use a general equilibrium selection model with communal land that is subject to expropriation and reallocation as a result of such expropriation. Ngai et al. (2019) incorporate reallocation risk of land in a model of migration.

of land market reform. Indeed, there is mixed empirical evidence on the impact of land titling programs on access to formal credit (Deininger and Chamorro, 2004; Galiani and Schargrodsky, 2010; Zegarra et al., 2011; Piza and de Moura, 2016; Agyei-Holmes et al., 2020). Taken together, the findings of these studies suggest that the efficiency of financial markets should be taken into account when the effects of improvements in land property rights are being quantified, as I do in this paper.

My model also allows studying how land property rights affect entrepreneurship. The majority of the entrepreneurship literature on developing countries explores the effect of only financial frictions and does not take land markets into account.[8] I find that improvement in land property rights leads to higher entrepreneurial activity as a lower risk of expropriation makes moving away from agriculture less costly, while the collateral channel provides access to finance to start or expand a business.

In the next section, I describe the data and provide empirical evidence of misallocation in the agricultural sector in Tanzania. In Section 1.3, I introduce a quantitative model of endogenous occupational choice that features incomplete financial and land markets. In Section 1.4, I calibrate the model to the Tanzanian economy and discuss the mechanics of the model. In Section 1.5, I present my main results on the effects of policy interventions. Section 1.6 concludes.

## 1.2 Empirical Evidence: How Do Land and Financial Markets Affect Economic Outcomes?

In this section, I empirically revisit the evidence that insecure land property rights and limited access to finance directly link to resource misallocation, which in turn affects sectoral and aggregate TFP. I start by estimating production functions and farmer-level TFP mea-

[8] See Buera et al. (2015) for the literature survey.

sures for the agricultural sector in an East African country, Tanzania. I then show that land market and credit market imperfections generate resource misallocation across and within sectors. These facts guide subsequent modeling choices and are used to inform the quantitative exercise.

### 1.2.1 Conceptual Framework

To fix ideas, consider an efficient static allocation in a simple model of farm size and input choice. As in Gollin and Udry (2021), there are $n$ heterogenous farmers producing a single homogeneous good according to the production function:

$$Y_i = e_i A L_i^{\alpha_L} \prod_k X_{k,i}^{\alpha_{X_k}}, \quad \text{with} \quad (\alpha_L + \sum_k \alpha_{X_k}) < 1,$$

where $L_i$ is the amount of land used by a farmer $i$ and the $X_{k,i}$ are other inputs like labor and capital used by this farmer. Individual total factor productivity is equal to $e_i A$, with $A$ being common productivity and $e_i$ is individual farming ability.

In this framework, we can characterize efficient static allocation of land across farmers given a fixed land supply. The efficient allocation maximizes aggregate output and solves the following social planner's problem:

$$\max_{\{L_i, X_{k,i}\}} \sum_i e_i A L_i^{\alpha_L} \prod_k X_{k,i}^{\alpha_{X_k}},$$

$$\text{subject to } \sum_i L_i = L, \ \sum_i X_{k,i} = X_k \quad \forall k.$$

The Pareto efficient allocation requires the marginal product of land to be the same across farmers. The efficient land allocation to farmer $i$ is proportional to the farmer's productivity

$e_i$:

$$L_i^* = \frac{e_i^{\frac{1}{1-\alpha_L-\sum \alpha_{X_n}}}}{\sum e_i^{\frac{1}{1-\alpha_L-\sum \alpha_{X_n}}}} L,$$

Hence, $\ln (L_i)^* \propto \ln (e_i)$, implying that farmers with higher farmer ability should operate a farm of larger size. In addition, factor intensity ratios should be identical across farmers. I use this framework to analyze micro data from Tanzania and motivate my empirical exercise that tests the efficiency of resource allocation in the agricultural sector.

### 1.2.2 Data

I use data from the Tanzania National Panel Survey, which represents panel data gathered in waves from the same households. The first wave was surveyed in 2008-09, the second wave in 2010-11, and the last two waves in 2012-13 and 2014-15. The fourth wave uses a new set of households together with a subsample of households from previous waves. The data were collected with support from the World Bank as a part of the LSMS-ISA project. The survey has regionally representative data for all regions on mainland Tanzania and Zanzibar and covers both rural and urban areas (Figure A.4). In addition to demographic and social characteristics of households, the survey includes detailed information on durable goods and financial assets; agricultural production, including land characteristics; and operations of non-farm household enterprises.

I focus on agricultural production at the household level, so the observation unit is a household $i$ in period $t$. One farmer may operate one or several plots of land. I, therefore, aggregate information available at the plot level to the household level. The dataset contains a panel of about 4,000 households and approximately 3,500 households that were added in the last round of the survey. The share of households involved in farming is around 65 percent.

**Output and inputs** In my analysis, I focus on the long rainy season. For each household, I construct a measure of agricultural output in a given year. My baseline measure is real agricultural output aggregated at the household level using actual quantities of each crop harvested by the time of interview and proxies of prices in 2012-13 as weights. The prevalence of intercropping, when several crops are cultivated simultaneously on a given piece of land, makes it impossible to measure output in physical quantities. Moreover, households report harvest in different units even for the same type of crop, which requires making some unit-price conversion to make the data comparable across farmers. To construct proxies of prices, I obtain the median price of different units for each crop at the national level, conditional on the crop being sold to someone outside the household.

There are four inputs for which quantitative data are available: land, labor, capital, and usage of chemicals such as fertilizers and pesticides. All plot areas are reported in acres, and I use farmer estimates for plots that were never measured by GPS.[9] In terms of land input, both the size of available land and the size of the land that was cultivated are available. I am using the latter in my empirical analysis. The measure of labor inputs is the total number of person-days used by the household. The survey distinguishes between work done by household members and by hired workers. The measure of capital input includes both chemical inputs, such as fertilizers and pesticides, as well as farm implements and machinery, such as hand hoe and plough. All types of capital inputs are aggregated at the household level and weighted by the median price of each type of input at the national level in 2012-13. I only use those purchased without a voucher and/or subsidy to compute the median price of chemical inputs. Moreover, some types of chemicals are reported in different units, and in this case, unit-price conversion is used. Capital includes both owned and rented machinery.[10]

[9]As a default, I use GPS measure of a plot. 63% of all plots in the sample were measured with GPS.

[10]I am using the same price weight for both owned and rented machinery, depending on the type of machinery or tool and not on the ownership status.

**Land property rights** Several indicators on land tenure are available in the survey. For each plot that the household owns or uses, the following information is available: i) whether a household has any legal document for this plot, and – if the answer is "yes" – what type of document; ii) whether a household has the right to sell it or to use it as collateral; iii) whether a household feels comfortable leaving this plot fallow without the worry of losing it; iv) whether the plot is used or obtained free of charge. Using this information for each plot, I construct four measures of land property rights at the household level as a share of total land that satisfies the respective criterion. Later, I use those measures of land property rights to assess the role of land market frictions in the allocation of resources.

**Other variables** The survey asks farmers about their agricultural practices, such as the use of other water sources and additional organic inputs, the number of trees on the plot, and whether specific tools are used at different stages of the agricultural process. The survey also provides information on other soil characteristics, including various soil type attributes and soil quality. In addition, I have information on land improvements and investments made by households in the recent past.

**Household characteristics** The survey data include a detailed description of households and individuals. Data are available on household composition and the age, education, literacy, and health characteristics of each household member; the relationship of each member to the household head; occupational choice of adults within households. In addition, for each household, there are data on different types of assets owned by a household – durable goods; live animals; agricultural tools, and equipment; as well as the outstanding amount of any loans both borrowed and/or lent within one year period from/to any source.

Table A.1 in the Appendix presents summary statistics of the main variables used in the analysis. The statistics show that farmers operate small plots, with an average cultivated area of 1.2 hectares. Also, farmers mostly rely on domestic labor – only half of the households hire any workers, and the average share of household labor is more than 90 percent. Finally, agricultural practices are labor-intensive, with almost no capital used and little chemical inputs (e.g., fertilizers, pesticides).

### 1.2.3 Agricultural Production Function and Measure of Productivity

To obtain a measure of household productivity, I first estimate the agricultural production function. The main challenge in such an estimation is that input choices are not exogenous to productivity, which is unobserved. While an extensive literature addresses this issue in the context of firms, application to agriculture is more limited.[11] Moreover, the literature on firm production function estimation often makes assumptions that are not appropriate to use in an agricultural setting, especially for a low-income country, such as Tanzania. Many approaches require one or several inputs to be monotonic in productivity, which is not a realistic assumption in a developing country due to the presence of numerous frictions and extensive subsidization of inputs such as fertilizers and seeds. Alternatively, imposing a fixed effect on the law of motion for productivity might lead to attenuation bias, especially in the context of small farmers, where most of the labor consists of household members. In this paper, I use the dynamic panel approach as a preferred method to deal with endogeneity issues making assumptions that are more appropriate in the context of small farmers in a developing country.

Consider the Cobb-Douglas production function

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_n n_{it} + \beta_k k_{it} + \omega_{it} + \varepsilon_{it},$$

[11] Firm level production function estimation literature includes Olley and Pakes (1996), Levinsohn and Petrin (2000), De Loecker (2011), Ackerberg et al. (2015), among others

where the unit of observation is household $i$ involved in agricultural activity in period $t$. $l$, $n$, and $k$ stand for (log) land, labor and capital inputs, while $y$ is (log) output. There are two terms, $\omega_{it}$ and $\varepsilon_{it}$, that are unobserved to the econometrician. However, $\omega_{it}$ is known to the farmer when he makes his inputs choices and, therefore, inputs are a function of $\omega_{it}$. Estimating the above equation via OLS leads to biased estimates since more productive farms will use more inputs given that the marginal product of an input is increasing in productivity.

I employ three approaches to production function estimation for comparison purposes. First, I start with simple OLS to estimate the agricultural production function. Second, to account for constant unobserved productivity over time, I add household fixed effects to my OLS regression. In this case, $\omega_{it}$ can be thought of as the agricultural ability of a household. This approach relies on the assumption that productivity is constant over time, i.e.:

$$\omega_{it} = \omega_{i,t-1} = \omega_i.$$

Moreover, in practice, this approach often results in attenuation in inputs like land that does not change much from year to year. To address these concerns, I use a dynamic panel approach as my third and preferred method. This approach relies on the timing of input choices to estimate coefficients.

Assume $\varepsilon_{it}$ is i.i.d. over time and uncorrelated with information set at time $t$, $\mathcal{I}_{it}$, and $\omega_{it}$ is following an AR(1) process:

$$\omega_{it} = \rho\omega_{it-1} + \xi_{it}.$$

Given the law of motion for productivity, we can quasi-difference the production function equation to get the estimating equation:

$$y_{it} - \rho u_{it-1} = (1-\rho)\beta_0 + \beta_l(l_{it} - \rho l_{it-1}) + \beta_n(n_{it} - \rho n_{it-1}) + \beta_k(k_{it} - \rho k_{it-1}) + \xi_{it} + \nu_{it},$$

where $\nu_{it} \equiv \varepsilon_{it} - \rho\varepsilon_{it-1}$. Assuming that $\xi_{it}$ is uncorrelated with $\mathcal{I}_{it-1}$, we can estimate the model using the moment conditions:

$$\mathbb{E}[\xi_{it} + \nu_{it} | \mathcal{I}_{it-1}] = \mathbb{E}\left[(\xi_{it} + \nu_{it}) \cdot \begin{pmatrix} l_{it-1} \\ n_{it-1} \\ k_{it-1} \end{pmatrix}\right] = 0.$$

There are two main issues with the dynamic panel approach. First, the estimation relies on the assumption that changes in land, labor, and capital are correlated with their lagged levels. This assumption fails in a world with perfect markets and without adjustment costs, as inputs are determined by the productivity level irrespectively of their past values. Second, it assumes that farmers have the same information set when they choose each input. Under perfect markets, this implies perfect collinearity between the level of each factor of production. I argue that in a low-income country like Tanzania, various market imperfections allow solving both problems. For example, a limited land market might not allow a farmer to increase land input in case of a positive productivity shock. As a result, the farmer is not able to adjust labor perfectly following his productivity. This implies that the current period labor input will correlate with past labor values and not be perfectly collinear with other inputs. However, such market imperfections rule out a class of structural methods that are often used in the literature in the context of advanced economies.[12]

In addition, unanticipated productivity shocks might change farmers' marginal products

[12]The main assumption of such structural methods is that inputs change monotonically with changes in productivity. Imperfect markets and the inability to freely choose the level of inputs violate this main assumption.

after choosing their factors and make the allocation look inefficient even when markets are perfect. To account for possible misspecification, I include indicators for illness, death in the family, flooding, problems with crop-eating pests, poor rainfall, and low/high prices for agricultural inputs/outputs in the year of farming activity in my estimation of the agricultural production function.

Table 1.1 presents estimates of the Cobb-Douglas production function at the household level.[13] I show estimates using simple OLS, OLS with household fixed effects, and dynamic panel estimation. In the latter case, I use a minimal distance procedure to estimate restricted coefficients. In all three specifications, I find decreasing returns to scale. This is plausible as farming in low-income countries is labor-intensive, and a large farm and workforce are harder to manage.

### 1.2.4 Market Distortions and Resource Allocation

Around 70 percent of the land in Tanzania is under customary land rights, and 80 percent of the population in rural areas depends on subsistence farming. One of the weaknesses of customary rights is that they are not formally documented. Only a small share of all land in Tanzania has a title or a certificate, which results in a higher risk of land expropriation and the inability to sell the land and use it as collateral. Moreover, historically the overriding principle in many communities is that the land belonged to the tiller. In other words, the land is subject to the principle "use it or lose it."[14]

Limited land markets result in around 15 percent of all plots not being fully utilized. i.e., part or all of the plot is being left fallow. Although leaving land fallow occasionally is required not to exhaust the soil and keep it fertile, most households are not able to cultivate

[13] Estimates of the production function without shocks are in Table A.6. Results are almost identical to the benchmark specification, suggesting that indeed included shocks were not anticipated. Moreover, the results are statistically identical to the inclusion of district-year fixed effects in all specifications.

[14] More details on the land tenure system in Tanzania can be found in the Appendix A.2.

Table 1.1: Production Function Estimates

| | OLS | OLS FE | DP |
|---|---|---|---|
| | (1) | (2) | (3) |
| log(Land) | 0.343 | 0.264 | 0.299 |
| | (0.015) | (0.026) | (0.071) |
| log(Labor) | 0.404 | 0.366 | 0.368 |
| | (0.017) | (0.025) | (0.161) |
| log(Capital) | 0.111 | 0.051 | 0.035 |
| | (0.006) | (0.009) | (0.025) |
| $\beta_l$ | | | 0.294 |
| $\beta_n$ | | | 0.412 |
| $\beta_k$ | | | 0.050 |
| $\rho$ | | | 0.533 |
| Return to scale | 0.85 | 0.68 | 0.76 |
| Test on common factor restrictions | | | 0.835 |
| # obs. | 8,949 | 6,073 | 3,641 |
| Unexpected shocks | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels. Regressions include year FE, OLS regressions - district-year FE.

the entire plot due to a lack of other inputs rather than soil considerations. If there was a well-functioning land market, those plots would be sold or rented out.

As a proxy of land property rights, I use four different measures that are related to the existence of formal proof of ownership, perception of expropriation risk in case land is unused, ability to sell the land and/or use it as collateral, and whether the land was used/obtained free of charge. Figure 1.1 displays the distribution of each measure in the sample. While all measures are positively correlated, they reflect different aspects of the land tenure system and are complementary in the analysis. I use all of them to test the presence of incomplete markets and the efficiency of resource allocation.[15]

[15]Tables A.2, A.3, A.4, A.5 in the Appendix present summary statistics for plots under different land property rights for each measure. Statistics are computed for plot and land characteristics, as well as for agricultural practices employment by households on a given plot. For most characteristics, there is

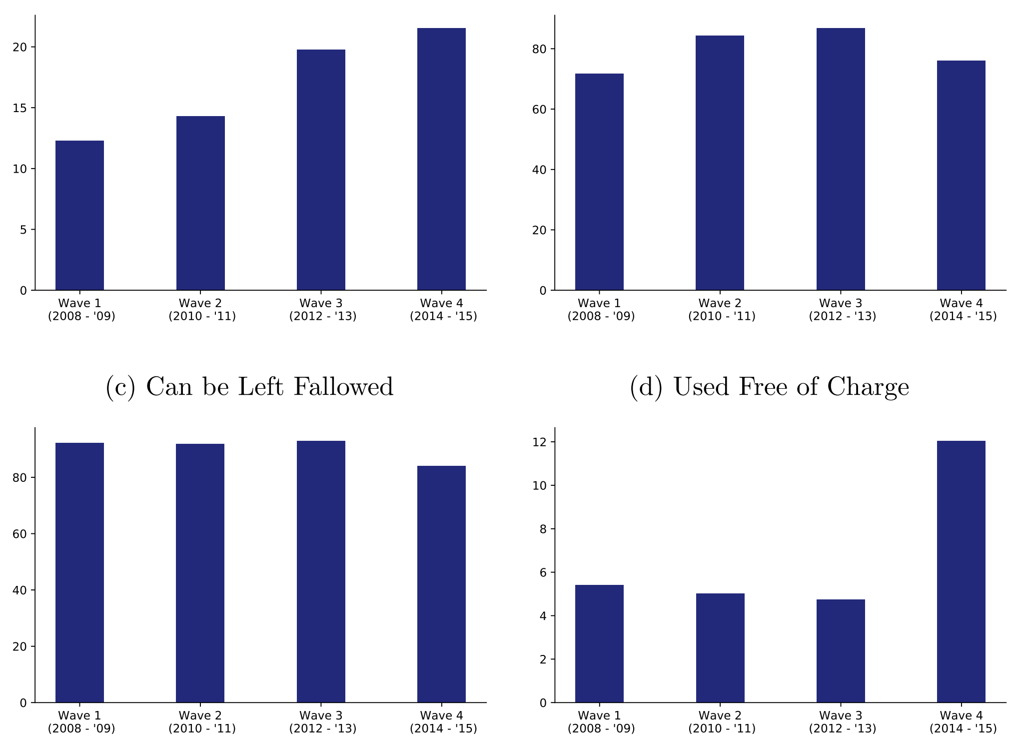
Figure 1.1: Measures of Land Property Rights

(a) Titled Land

(b) Can be Sold or Used as Collateral

(c) Can be Left Fallowed

(d) Used Free of Charge




*Notes:* Each plot depicts the share of land that is owned and/or used by a household and (a) the household has a legal document for this land, (b) the owner of this land has the right to sell it or use as collateral, (c) the household feels comfortable leaving this land fallow without the worry of losing it, (d) this land is used/obtained free of charge.

As discussed in Section 1.2.1, in the efficient static allocation, the amount of land used by the farmer should be positively correlated with the productivity of this farmer. Moreover, the relationship between these two variables should be the same for all farmers in an economy with no frictions. In the case when the land market is limited under the customary tenure system, an additional constraint might be present. For example, if households are unable to

no systematic difference between plots under different property rights regimes that is consistent across all measures. The only exception is plot size and whether the soil type is loam. Plots that are stronger property rights regimes are on average larger and are less likely to have loam soil.

rent land, they face a constraint

$$L_i \leq \bar{L}.$$

In this case, some households will be constrained with $L^* = \bar{L}$, which is independent of productivity. Hence, the relationship between land and productivity would differ for farmers operating under different property rights regimes. It is also straightforward to show that the relationship is not the same for financially constrained and unconstrained households.

To test the presence of resource misallocation that is associated with insecure land property rights and limited access to credit, I use the following baseline regression specification:

$$l_{it} = \phi_0 \ln e_{it} + \phi_1 \left(\ln e_{it} \times Land\_rights_{it}\right) + \phi_2 \left(\ln e_{it} \times Credit_{it}\right) + \delta_{st} + \epsilon_{it},$$

where $l_{it}$ is log of the amount of land used by the farmer $i$ in agricultural production in year $t$, $\ln e_{it}$ is log of farmer's productivity obtained by computing residual using estimated parameters of the production function, $\delta_{st}$ denotes district-year fixed effects to control for things like common weather shocks, and $\epsilon_{it}$ denotes the error term. The interaction terms include a measure of land property rights, $Land\_rights_{it}$, which is computed as a share of land belonging to a specified category (e.g., has a title) to the total amount of household's land in a given period $t$. Additionally, I include an interaction term of productivity and a dummy variable $Credit_{it}$, which is an indicator of whether the household borrowed for agricultural purposes in the past 12 months from any sources.

Table 1.2 displays the results. The main observation is that there is a positive relationship between the size of land used and productivity. However, this relationship is different for farmers depending on whether cultivated land has strong property rights. Similarly, the relationship is different for farmers who borrowed some resources for agricultural purposes compared to those who did not. Moreover, for some land property rights measures, there is a positive and statistically significant relationship between land size and productivity only in the case of strong land property rights.

Table 1.2: Land Misallocation

| | ln(land) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | leave fallow | | right to sell | | title | | obtain free | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| HH productivity | 0.050 | 0.014 | 0.011 | 0.014 | 0.011 | 0.047 | 0.044 | 0.057 | 0.056 |
| | (0.013) | (0.009) | (0.009) | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) |
| HH productivity × land_rights | | 0.044 | 0.044 | 0.056 | 0.056 | 0.023 | 0.023 | -0.060 | -0.059 |
| | | (0.004) | (0.004) | (0.003) | (0.003) | (0.005) | (0.005) | (0.005) | (0.005) |
| HH productivity × credit | | | 0.052 | | 0.050 | | 0.051 | | 0.050 |
| | | | (0.009) | | (0.009) | | (0.010) | | (0.010) |
| # obs. | 8,939 | 8,939 | 8,939 | 8,939 | 8,939 | 8,939 | 8,939 | 8,939 | 8,939 |
| # households | 5,095 | 5,095 | 5,095 | 5,095 | 5,095 | 5,095 | 5,095 | 5,095 | 5,095 |
| Wave#District FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $R^2$ | 0.290 | 0.301 | 0.304 | 0.319 | 0.322 | 0.292 | 0.295 | 0.305 | 0.307 |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels. The second row indicates which measure of land property rights is used in the regression analysis.

In addition, in the case of complete markets, variation across farmers in factor ratios would reflect misallocation.[16] Tables A.7 and A.8 in the Appendix present evidence of different ratios of inputs, first, for households that are subject to different property rights regimes, and, second, for those households that were able and/or willing to obtain a loan for agricultural purposes compared to those that were not. These empirical results suggest that markets are not complete and that market incompleteness is linked to land property rights and access to credit.

### 1.2.5 Robustness and Other Findings

In this section, I test some of the assumptions that could affect the results on resource misallocation. In addition, to provide a micro-foundation for the model, I explore the relationship

[16]This statement generalizes to any homothetic production function.

between land property rights and different household characteristics.

**CES production function** A possible explanation for the observed misallocation could be that the unity substitution elasticity assumption in the Cobb-Douglas production function is invalid. Although the assumption of a Cobb-Douglas production function is standard in the literature on misallocation, I show that using CES production function also leads to the conclusion that there exists market incompleteness associated with land property rights and access to credit.

Suppose

$$Y_i = e_i \left[\alpha L_i^{-\rho} + \beta N_i^{-\rho} + (1-\alpha-\beta) K_i^{-\rho}\right]^{-\frac{\sigma}{\rho}},$$

where $\sigma$ denotes the return to scale and $\epsilon = \frac{1}{1-\rho}$ is the elasticity of substitution between factors. I assume that $e_i$ is the product of household productivity and time and region fixed effects. Table A.9 in the Appendix reports the results of estimating the equation with nonlinear least squares.[17]

In an efficient static allocation, the marginal product of land should be equalized across farmers. I examine whether land property rights and access to credit are sources of variation in MPL across farmers to test whether there exists market incompleteness related to these factors. As evidenced from Table 1.3, the marginal product of land is higher for farmers that are subject to insecure land property rights and lower for those who did not have a loan. The relationship between the marginal product of the land and land property rights can reflect the fact that in the areas with relatively weak property rights, both rental and final markets for land are absent. At the same time, credit for agricultural purposes is used to buy capital and inputs like fertilizers, and, hence, we observe a positive relationship between credit and MPL.

[17]The ideal estimator is the nonlinear equivalent of the dynamic panel, which applies GMM to the first-difference equation using lagged factors as instruments. Unfortunately, this estimator does not converge.

Table 1.3: Marginal product of land and market frictions

| | ln(MPL) | | | |
|---|---|---|---|---|
| | leave fallow | right to sell | title | obtain free |
| land_rights | -0.196 | -0.184 | -0.034 | 0.216 |
| | (0.035) | (0.029) | (0.045) | (0.042) |
| credit | 0.403 | 0.414 | 0.404 | 0.410 |
| | (0.093) | (0.092) | (0.093) | (0.092) |
| # obs. | 8,925 | 8,925 | 8,925 | 8,925 |
| Wave#District FE | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels.

**Variation across time** In my baseline analysis, I explore the efficiency of resource allocation using a variation of land property rights both across time and space. By adding household fixed effects to my baseline specification, I exploit whether a positive relationship between land and productivity is present in the data for the transitory part of productivity. In other words, I test whether households adjust the amount of land used in agricultural production when they experience transitory productivity shock and whether there exists any difference in this adjustment depending on the strength of land property rights and access to credit.

Table A.10 in the Appendix displays the results. I find a positive relationship between productivity and land usage only for those households who operate more secure land in terms of property rights. These results are consistent with the prediction that inability to rent out or sell the land that is not formally registered or subject to expropriation risk prevents households from making adjustments in the amount of land inputs when they experience productivity shocks.

**Land property rights and other household characteristics** To further motivate my modeling choice in the next section, I examine whether there exists any association between land property rights and different household characteristics. Table 1.4 reports the results of this exercise. I find that households with titled land are more likely to rent out the land, potentially due to lower expropriation risk.

Households that have an official document for their land are not only more likely to obtain credit in the last 12 months but also enjoy a larger loan size conditional on being given one. Since in every regression I include household fixed effects, such a relationship can be explained by the collateral channel. Suggestive evidence that supports this theory is the fact that in the year 2014/2015 around 49.2 billion shillings had been issued as loans by various financial institutions, using Certificates of Customary Rights as collateral (URT (2016)). Finally, there is a link between land property rights and occupational choice. Households with titled land are less likely to stay in agriculture (as an occupation of the head and as a share of household labor) and more likely to operate a business.

Table 1.4: Land property rights and other household characteristics

| | Dependent variable | | | | |
|---|---|---|---|---|---|
| | rent out land | head of HH in agriculture | obtained credit | size of a loan | operate a business |
| land_rights | 0.015 | -0.037 | 0.028 | 0.574 | 0.023 |
| | (0.006) | (0.014) | (0.013) | (0.199) | (0.015) |
| # obs. | 7,874 | 11,752 | 11,752 | 448 | 11,752 |
| Household FE | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the household and district levels. Regressions with dependent variable on occupation or presence of business also include dummy variable indicating whether HH owns a plot.

## 1.3 A Model with Incomplete Land and Financial Markets

In this section, I suggest a model that links access to finance, occupational choice, and land ownership. It is a standard occupational choice model with financial frictions but enriched with an additional feature – land ownership, either private or communal.

Time is discrete in the economy. The economy is populated by a continuum of infinitely lived households of measure one indexed by $i \in [0, 1]$. In each time period, a household's state consist of five elements: i) productive skill in the agricultural sector, $z_a > 0$; ii) productive skill in entrepreneurship, $z_e > 0$; iii) endowment of land, $l \geq 0$; iv) property rights regime, $pr = c, p$, either communal or private; v) level of assets, $a \geq 0$. Skills are exogenous and the evolution process is known to a household. Assets evolve endogenously by forward-looking saving behavior.

The total land endowment of land in the economy is $L$, with a fraction $\lambda_l \in [0, 1]$ being communal (weak land property rights), while the rest is private (strong land property rights). The total and individual levels of private land are fixed and can be used for agricultural production and also can be used as collateral. The total amount of communal land is fixed. However, individual communal land holdings evolve endogenously due to the presence of expropriation risk, and communal land is neither allowed to be rented out nor used as collateral.

### 1.3.1 Setup

**Preferences** Individual preferences are described by the following expected utility function over sequences of consumption, $c_t$:

$$U(c) = \mathbb{E}_t \left[ \sum_{t=0}^{\infty} \beta^t u(c_t) \right], \text{ where } u(c_t) = \frac{c_t^{1-\sigma}}{1-\sigma},$$

where $\beta$ is the discount factor, and $\sigma$ is the coefficient of relative risk aversion.

**Occupational Choice** At the beginning of each period, a household chooses whether to operate his own business, become a worker, or cultivate a farm. Firms and the agricultural sector produce a single final good. Each firm is run by one entrepreneur, who produces the good using as inputs his entrepreneurial ability, labor, and capital. Each farm is run by one farmer, who produces the good using his productivity in the agricultural sector, land, and capital as inputs.[18] All occupational choices are mutually exclusive within a period $t$. There is no cost of switching between occupational choices across periods.[19]

**Land and Financial Markets** Agents have access to a perfectly competitive financial intermediary who receives deposits from households and makes loans to farmers and entrepreneurs. The deposit rate $r_t$ is determined endogenously by the capital market clearing condition at period $t$. Households use loans to finance capital. Competitive financial intermediation implies that loan contracts are paid at the gross interest rate, $r_t^k = r_t + \delta$, where $\delta$ denotes the depreciation rate of capital. Also, there is a competitive intermediary that

[18] I abstract from hired labor input and assume that labor input is embedded in agricultural household productivity, $z_a$. This is not a strong assumption, given that household members supply the majority of agricultural hours in Tanzania as is shown in Table A.1 in the Appendix.

[19] This assumption allows to avoid carrying additional state variable and is common in literature on entrepreneurship and development (For a summary see Buera et al., 2015).

collects all leased land and then rents it out at rate $r_t^l$.[20]

Financial markets are incomplete in several dimensions. First, no state-contingent bonds can be purchased. Hence, there is no opportunity to insure against productivity risk. Second, I do not allow borrowing for consumption smoothing across periods by imposing $a_t \geq 0$, therefore entrepreneurs and farmers only can borrow within periods to finance production. Third, similar to Jermann and Quadrini (2012) and Mendoza (2010), I assume that there is a cash flow mismatch, such that the amount of capital that exceeds the current level of assets owned by the household must be financed in advance of production. Thus, households need to borrow intraperiod to finance capital. However, the total amount of borrowing is limited by a collateral constraint due to the limited enforceability of debt contracts. A novel ingredient in my model is that in addition to assets, land can also be used as part of the collateral.

Consider a household with wealth $a_t$ and land holding $l_t$ that is asking for a loan $x_t$ from a financial intermediary at rate $r_t^k$. Once a loan is obtained, the household transforms it costlessly together with assets (but not land, which is used as an input in farmer's production) into capital $k_t = a_t + x_t$. Together with land holdings, the capital is then used as collateral to secure the loan $x_t$. The household is free to default and walk away with his income and wealth at any time. In this case, collateral will be seized. I assume that the liquidation value of capital is uncertain at the time of contracting, similar to Jermann and Quadrini (2012). With probability $(1 - \frac{1}{\lambda_k})$, where $\lambda_k \geq 1$, intermediary recovers the full value of collateral, $k_t + q_t^l l_t$, where $q_t$ is the shadow price of land. However, it recovers nothing with probability $\frac{1}{\lambda_k}$. Hence, the amount of loan $x_t$ that intermediary is willing to provide is limited

[20] In the benchmark version of the model, land holdings are fixed for each household. Households are able to adjust the amount of land used in the production only by renting. In terms of allocation of land across farmers, the rental market is equivalent to the ability of households to buy or sell land. At the same time, the introduction of the market for land purchases will incentivize households to use the land as a saving tool. This additional mechanism would complicate the model substantially, and it is outside of the scope of this paper. In addition, the model is consistent with the limited land market in Tanzania, with most land being rented.

to $x_t \leq (1 - \frac{1}{\lambda_k})(k_t + q_t^l l_t)$.[21] The household's capital constraint in terms of his wealth and land holdings is then:

$$k_t \leq \lambda_k(a_t + q_t^l l_t) - q_t^l l_t$$

The parameter $\lambda_k$ measures the degree of credit frictions, with $\lambda_{k,l} = +\infty$ corresponding to a perfect credit market and $\lambda_k = 1$ to financial autarky where all capital is self-financed. This captures the common prediction from models with limited contract enforcement: credit is limited by an individual's wealth.

The land market is incomplete in the part of the economy with weak property rights. Land under customary tenure regime cannot be rented out and used as collateral. Land market imperfections amplify financial market frictions by making collateral constraint tighter:

$$k_t \leq \lambda_k(a_t + q_t^l l_{t,\mathbb{I}\{land=private\}}) - q_t^l l_{t,\mathbb{I}\{land=private\}}$$

That is, the collateral value of land only appears if land is private.

**Evolution of communal land** I assume that the communal land that belongs to the household brings zero value if not used. Moreover, communal land that is not used in the current period is subject to expropriation risk with some positive probability, $\pi_E$. This means, that $\pi_E > 0$, if $l_{i,\mathbb{I}\{land=communal\}} - l_i^d > 0$, where $l_i^d$ is farmer's land input. In addition, I assume that expropriation probability is independent of any other household characteristics.

Expropriated communal land is reallocated to other households via a lump-sum transfer $\eta_t$, which is endogenous. I assume that the reallocation probability $\pi_R$ is positive for households that engage in farming in the current period and zero otherwise. Similar to $\pi_E$, I

[21] $q_t^l$ is the shadow price of land in consumption units, and is defined as the present value of its expected future income flows in terms of the consumption numeraire. This means that there is endogenous general equilibrium effect on the tightness of collateral constraint, as $q_t^l$ is directly linked to the rental rate of land, $r_t^l$.

assume that reallocation probability and the value of a lump-sum transfer $\eta_t$ are independent of any other household characteristics.[22]

### 1.3.2 Household Problem

The state vector consists of level of wealth, amount of land owned, property rights regime, entrepreneurial ability, and agricultural productivity, $s_{it} \equiv (a_{it}, l_{it}, z^a_{it}, z^e_{it}, pr_i)$. I proceed in two steps to characterize the household problem. First, I write the household value function as the maximum across the value function conditional on occupational choice,

$$V_t(s_{it}) = \max \left\{ V_t^{Worker}(s_{it}), V_t^{Entrepreneur}(s_{it}), V_t^{Farmer}(s_{it}) \right\}.$$

Second, I consider the value function for different occupational choices, conditional on the property rights regime.

**Households under private property rights regime** Let $x_{it} \equiv (a_{it}, l_i, z^a_{it}, z^e_{it})$,[23] then the problem of households is the following:

$$\max_{c_{it}, a_{it+1}, k_{it}^{o\in\{E,F\}}, n_{it}^{o\in\{E\}}, l_{it,d}^{o\in\{F\}}} V_t(x_{it}) = \frac{c_{it}^{1-\sigma}}{1-\sigma} + \beta \mathbb{E}_t[V_{t+1}(x_{it+1}|x_{it})]$$

subject to the budget constraint

$$c_{it} + a_{it+1} \leq y^o_{it} + r^l_t l_i + (1 + r_t) a_{it},$$

[22] I assume that $\pi_R$ is constant across time, and $\eta_t$ depends on the amount of expropriated land and households' occupational choice. Alternately, $\eta$ can be fixed as in Ngai et al. (2019), implying $\pi_{t,R}$ to equalize expropriated and reallocated land. In their paper, Gottlieb and Grobovsek (2019) focus on the expropriation risk of communal land and model both $\eta$ and $\pi_R$ as state dependent variables.

[23] The amount of private land that household owns is fixed. In the model, I focus on the rental market as sale and purchase of land remain rare in Tanzania, with most land being inherited or allocated by local authorities.

the within period capital borrowing constraint (collateral)

$$k_{it} \leq \lambda_k a_{it} + (\lambda_k - 1) q_t^l l_i, \quad o \in \{Entrep, Farmer\},$$

and the across periods borrowing constraint

$$a_{it+1} \geq 0.$$

$y_{it}^o$ for each occupational choice is given by

$$\begin{aligned}
y_{it}^{Entrep} &= z_{it}^e k_{it}^{\alpha_e} n_{it}^{\gamma_e} - w_t n_{it} - r_t^k k_{it}, \\
y_{it}^{Worker} &= w_t, \\
y_{it}^{Farmer} &= z_{it}^a k_{it}^{\alpha_a} (l_{it}^d)^{\gamma_a} - r_t^k k_{it} - r_t^l l_{it}^d.
\end{aligned}$$

**Farmer under communal land property rights regime** For households living in the communal part of the economy, the amount of land endogenously evolves across periods. Given that communal land cannot be rented out and production function is increasing in land, farmers in the communal part of the economy would never use less land in production than their land holdings. Therefore, for farmers communal land is never subject to expropriation risk.

Letting $x_{it}' \equiv (a_{it}, l_{it}, z_{it}^a, z_{it}^e)$, the household problem for a farmer is:

$$\begin{aligned}
\max_{c_{it}, a_{it+1}, k_{it}, l_{it}^d} V_t^{Farmer}(x_{it}') = \frac{c_{it}^{1-\sigma}}{1-\sigma} \quad + \\
+ \beta \left\{ \pi_R \mathbb{E}_t [V_{t+1}(x_{it+1}', l_{it+1} = (l_{it} + \eta) | x_{it}')] + (1 - \pi_R) \mathbb{E}_t [V_{t+1}(x_{it+1}', l_{it+1} = l_{it} | x_{it}')] \right\}
\end{aligned}$$

subject to the budget constraint

$$c_{it} + a_{it+1} \leq y_{it} + (1 + r_t)a_{it},$$

the within period capital borrowing constraint (collateral)

$$k_{it} \leq \lambda_k a_{it},$$

and the across periods borrowing constraint

$$a_{it+1} \geq 0.$$

$y_{it}$ for the farmer is:

$$y_{it}^{Farmer} = z_{it}^{a} k_{it}^{\alpha_a} (l_{it}^{d})^{\gamma_a} - r_t^k k_{it} - r_t^l (l_{it}^d - l_{it}) \mathbb{I}_{\{l_{it}^d \geq l_{it}\}}.$$

**Entrepreneur and worker under communal land property rights regime** Workers and entrepreneurs in the communal part of the economy do not use land in production. Therefore, their entire land holdings are subject to expropriation risk. Their problem is:

$$\max_{c_{it}, a_{it+1}, k_{it}^{o\in E}, n_{it}^{o\in E}} V_t^{o\in\{Entrep,Worker\}}(x'_{it}) = \frac{c_{it}^{1-\sigma}}{1-\sigma} \quad +$$
$$+ \beta \left\{ \pi_E \mathbb{E}_t[V_{t+1}(x'_{it+1}, l_{it+1} = 0 | x'_{it})] + (1 - \pi_E)\mathbb{E}_t[V_{t+1}(x'_{it+1}, l_{it+1} = l_{it} | x'_{it})] \right\}$$

subject to the budget constraint

$$c_{it} + a_{it+1} \leq y_{it}^{o} + (1 + r_t)a_{it},$$

the within period capital borrowing constraint (collateral)

$$k_{it} \leq \lambda_k a_{it} \quad o \in \{Entrepreneur\},$$

and the across-period borrowing constraint

$$a_{it+1} \geq 0.$$

$y_{it}^{o}$ for each occupational choice is

$$y_{it}^{Entrep} = z_{it}^{e} k_{it}^{\alpha_e} n_{it}^{\gamma_e} - w_t n_{it} - r_t^k k_{it},$$
$$y_{it}^{Worker} = w_t.$$

### 1.3.3 Market Clearing

Let $\mathcal{F}_t(a, l, z^a, z^e, pr)$ denote the joint distribution of wealth, land ownership, property rights regime, and agricultural and entrepreneurial productivity at time $t$ over all households.

**The labor market clearing condition is:**

$$\int_{e=entrep} n_t d\mathcal{F}_t(a, l, z^a, z^e, pr) = \int \mathbb{I}\{e = worker\} d\mathcal{F}_t(a, l, z^a, z^e, pr).$$

That is, labor demand by entrepreneurs should be equal to the labor supply of workers to a wage job.

**The land market clearing is:**

$$\int l_{\mathbb{I}\{land=rent\_out\}} d\mathcal{F}_t(a,l,z^a,z^e,pr=private) = \int_{e=farmer} l_{\mathbb{I}\{land=rent\_in\}} d\mathcal{F}_t(a,l,z^a,z^e,pr)$$

The total amount of private land that is rented out should be equal to the amount of land rented in by farmers.

Also the amount of communal land that is reallocated should be equal to the amount of land that is expropriated:

$$\int l d\mathcal{F}_t(a,l,z^a,z^e,pr=communal) = \lambda_l L.$$

**The capital market clearing is:**

$$\int a_t d\mathcal{F}_t(a,l,z^a,z^e) = \int_{e=entrepreneur,farmer} k_t d\mathcal{F}_t(a,l,z^a,z^e).$$

The total supply of assets should be equal to the capital demand by entrepreneurs and farmers.

### 1.3.4 Competitive Equilibrium

Given an initial distribution of state variables $\mathcal{F}_t(a,l,z^a,z^e,pr)$ and a sequence of wages, interest rate of capital and land, and communal land reallocation $\{w_t, r_t^k, r_t^l, \eta_t\}_{t=0}^{\infty}$, a competitive equilibrium is given by a sequence of allocations $\{c_t(s), a_t(s), k_t(s), n_t(s), l_t^d(s)\}_{t=0}^{\infty}$ and occupational choices $\{e_t(s) = \{Worker, Entrepreneur, Farmer\}\}_{t=0}^{\infty}$ such that $(i)$ households maximize utility by solving value function maximization problem subject to budget constraint, within and across periods borrowing constraints, $(ii)$ the financial intermediary sector makes zero profits, $r_t^k = r_t + \delta$ and $(iii)$ there is market clearing in the labor market, capital market, and land market.

**Stationary competitive equilibrium** In addition, a stationary competitive equilibrium requires that the joint distribution of state space is a fixed point of the equilibrium mapping and that prices are constant over time.

$$\mathcal{F}(a, l, z^a, z^e, pr) = \mathcal{F}_t(a, l, z^a, z^e, pr) = \mathcal{F}_{t+1}(a, l, z^a, z^e, pr)$$

and

$$w_t = w, \quad r_t^k = r^k, \quad r_t^l = r^l, \quad \eta_t = \eta$$

I focus on a stationary competitive equilibrium when performing counterfactual exercises.

**Computational Algorithm** For a given set of parameter values, the solution algorithm involves first guessing a steady state prices, $w, r^k, r^l, \eta$. Given the prices, solve the policy functions for each set of state variables by value function iteration. Given the policy functions, find the stationary distribution. Check whether market clearing conditions are satisfied and update the guess of prices if needed. More details in the Section A.3 in the Appendix.[24]

## 1.4 Model Calibration and Underlying Mechanism

In this section, I present results from numerical exercises with the model. I start my analysis by calibrating the model to the economy of Tanzania. Then, I show how a household's wealth, land ownership, and productivity determine their occupational choices and land

[24]Given the dimensionality of state space and occasionally binding constraints, I use the computational resources provided for the Quest high-performance computing facility at Northwestern University to estimate the model.

usage decisions under different property rights regimes. This helps to illustrate how land property rights affect different people in different ways.

I use the calibrated model to conduct experiments to assess the effect of improvement in land property rights by moving from the economy with a positive share of land under the customary tenure system to the economy with only modern private land property rights. I first document the impact of such policy on a number of aggregate variables, like productivity and prices. Then, I decompose the effect of full-fledged land reform on the various channel by removing only one land market friction at a time and exploring the general equilibrium impact of such an experiment. In my third exercise, I use the model to compare the aggregate effect of financial reform relative to land reform by setting the parameter that governs the degree of financial friction to the level of an advanced economy. Finally, to analyze the short-run implication of land reform, I look at the transition path of the model economy from the initial steady state to a steady state after land reform took place.

### 1.4.1 Calibrating the Model to the Tanzanian Economy

The model has 15 parameters for which I need to specify values. Some of the parameters are standard in the literature, others recovered from the analysis of the data available for Tanzania. The remaining set of parameters is calibrated to match aggregate moments jointly. In addition to Household Panel Survey, I use the World Bank's Enterprise Survey and World Development Indicators to discipline the financial part of the model. All the data are for the period 2012-13.

**Access to finance** The use of bank financing by firms in Tanzania is still limited by international standards. According to the World Bank's enterprise survey, only 18% of firms used banks to finance investment, and around 17% of firms had a loan or a line of credit from a bank. From a list of fifteen items proposed in the same survey, respondents were asked to rank the most significant obstacle faced by the firm for its day-to-day operations. 38% of firms reported access to finance to be the biggest obstacle.

Excessive reliance on internal funds is a sign of potentially inefficient financial intermediation. Such inefficiencies are often reflected in a high value of collateral needed for a loan relative to the loan's value. According to the World Bank's enterprise survey, the level of this parameter in Tanzania is almost 250%, which is higher than the average value in low-income countries and Sub-Saharan Africa. Moreover, 96.2 percent of loans require collateral. Such a high collateral value accompanied by a low level of assets among households results in very limited access to finance. According to the model, private landholders can still get access to credit even when their financial assets are low by using land as collateral. This model feature is supported by the data on the land titling program in Tanzania. Based on information on one of the largest titling projects, Mkurabita, at least US$2.2 million had been loaned to some of the 110,000 villagers who obtained occupancy certificates under Mkurabita (Schreiber, 2017). Data from another pilot project also suggests that households used their documented land to get credit.

**Productivity** Productive skills of households are exogenous, independent from each other, and the evolution process is known to a household. Specifically, the logarithm of productive skills for each sector $s \in \{a, e\}$ follows a first-order autoregressive process

$$z_{s,t} = \rho z_{s,t-1} + \varepsilon_{s,t},$$

where $|\rho| < 1$ is the persistence in productivity and $\varepsilon_{s,t}$ is a white noise process with variance $\sigma^2_{\varepsilon,s}$, which represents idiosyncratic risk component.

**Technology** Entrepreneurs produce with a production function that combines entrepreneurial productive skill $z^e$, capital, and labor. The production function is increasing in all the arguments, strictly concave in capital and labor, and has a decreasing return to scale. In particular,

$$f(z^e, k, n)^e = \exp(z^e)(k^{\alpha_e} n^{1-\alpha_e})^{1-\nu},$$

where $0 < 1 - \nu < 1$ is the span of control as in Lucas (1978). Similarly, the agricultural production function has a decreasing return to scale and combines agricultural productivity skill $z^a$, capital and land with coefficient $\alpha_a$ and $\gamma_a$ obtained from the agricultural production function estimation.[25]

**Communal Land Evolution** I use simple functional forms for $\pi_R$ and $\pi_E$. $\pi_E \in (0, 1)$ if the fraction of land used by the household is smaller than land holdings,[26] and zero otherwise. $\pi_R \in (0, 1)$ if household decides to stay in agriculture in the current period,[27] and zero otherwise.

**Parameters invariant over time and across economies** The model is calibrated to a period of one year. I set the risk-aversion parameter $\sigma = 1.5$, and the one-year depreciation rate $\delta$ is set to 0.06 following Buera et al. (2021). The aggregate income share of capital for entrepreneur $\alpha_e$ is set to 0.33.

**Parameters derived from the data** Agricultural production function estimation, agricultural productivity is following AR(1) process in logs with persistence $\rho_a$ and normal innovations with variance $\sigma_a^2$. The autocorrelation coefficient, $\rho_a$ is estimated to be 0.533 for

[25] Labor input is not explicitly modeled but instead embedded in $z^a$ as almost all agricultural labor is coming within the household in the data. The production function is described by

$$f(z^a, k, l)^a = \exp(z^a) k^{\alpha_a} l^{\gamma_a}$$

[26] This means that only households that choose to be workers or entrepreneurs are subject to positive expropriation risk of land as those who are farmers would never decide to use less land than land holdings in equilibrium (production function is increasing in land; communal land can not be rented out).

[27] I also assume that for the households with $l_i = \max(l_i)$, reallocation probability is equal to zero, or $\pi_R = 0$. This assumption is made for computational reasons.

the model with non-anticipated shocks. I make a similar assumption about the productivity process for entrepreneurs, which is independent of the agricultural productivity process. To measure the autocorrelation coefficient, $\rho_e = 0.262$, I use values for net average monthly profit during the months when a non-farm enterprise is operating from the Household Panel Survey.

I set the share of communal land to be $\lambda_l = 80.7$ percent of total land, which is the share of households' land that does not have any official document that can prove ownership in years 2012-2013. I assume that the probability of land expropriation is constant for those households that decide to leave their land uncultivated. The share of land under weak property rights that cannot be left fallow without risk of expropriation identifies parameter $\pi_E = 9\%$.

**Parameters calibrated by matching moments** I have six remaining parameters, which are calibrated to match relevant moments shown in Table 1.5: the annual real interest rate; the share of hired workers, farmers, and entrepreneurs; and the distribution of land across households. The key parameter that captures financial frictions, $\lambda_k = 1.416$, is calibrated to match the average value of collateral needed for a loan as a percent of the loan amount, which is equal to 240.2% in Tanzania. Based on the data from Enterprise Survey, 96.2% of loans do require collateral, which is consistent with the model that assumes that every loan requires collateral.

Table 1.5: Calibration

| Target Moment | Data | Model | Parameter | Description |
|---|---|---|---|---|
| Real interest rate (%) | 3.8% | 3.75% | $\beta = 0.813$ | Discount factor |
| Share of hired workers (% of empl.) | 20.5% | 20.5% | $\nu = 0.535$ | Span of control |
| Share of farmers (% of empl.) | 61.0% | 61.1% | $\sigma_a = 0.09$ | S.d. of prod. shock (agriculture) |
| Share of entrepreneurs (% of empl.) | 18.5% | 18.4% | $\sigma_e = 0.75$ | S.d. of prod. shock (entrepreneurship) |
| Land distribution | | Figure A.5 | $\pi_R = 0.13$ | Probability of reallocation |
| Collateral/loan value | 240.2% | 240.4% | $\lambda_k = 1.416$ | Tightness of collateral constraint |

**Untargeted Moments** I also look at whether the model matches non-targeted measure of consumption inequality. Although consumption inequality in the model is slightly lower compared to the data, the overall pattern is similar (Figure A.6). In addition, the model matches well the level of land utilization, which is 88% in the data, and in the model it is 92%.

### 1.4.2 Discussion on the Mechanics of the Model

Using the baseline calibrated model, I compare household choices for the part of the economy that operates under customary land property rights with the part that operates under the modern property rights regime. Specifically, I describe how customary land tenure affects the economy through two channels: land misallocation and distortions in occupational choice. There are three main differences between the two property rights regimes: i) customary land is subject to expropriation risk in case it is not used by household, ii) customary land cannot be rented out, and iii) customary land is not allowed to be used as collateral to finance capital.

**Land property rights and land misallocation** Efficient allocation requires that the amount of land that the farmer uses is proportional to his productivity. However, the presence of land and financial markets distortions leads to the misallocation of inputs of production. First, financial frictions result in inefficient land usage for farmers both under modern and customary land tenure for financially constrained farmers. The reason for such inefficiency is that farmers are not able to obtain an efficient amount of capital and, hence, use the efficient amount of land.

Second, the presence of land market frictions leads to either over-usage or under-usage of land by farmers subject to these frictions. Figure 1.2 documents the ratio of farmer's operational land in the part of the economy without land frictions and the part of the economy with land frictions given different households characteristics. Under-usage of land is driven by the inability to use the land as collateral to finance the optimal amount of capital, which leads to a lower amount of both capital and land used by the farmer. This effect is the most pronounced for households with high agricultural productivity, low level of financial assets, and an amount of land holding that is positive but smaller than the efficient amount of land.

**Proposition 1.** *Denote optimal choices of land used by farmers who own the land under communal and private property rights regimes as* $l_c^*$ *and* $l_p^*$*, respectively. Then, if optimal land usage is larger than household land holding,* $l_p^* > l_p$*, and farmers' initial conditions in private and communal sectors of the economy are the same (i.e., the same amount of land, skills, and assets):*

$$l_c^* \leq l_p^*,$$

*and for assets holdings* $a_{small} < a_{large}$*, given everything else the same, the following is true*

$$l_p^*(a_{small}) - l_c^*(a_{small}) \geq l_p^*(a_{large}) - l_c^*(a_{large}),$$

*and for levels of agricultural productivity* $z_{small} < z_{large}$*, given everything else the same, we have*

$$l_p^*(z_{small}) - l_c^*(z_{small}) \leq l_p^*(z_{large}) - l_c^*(z_{large}),$$

*and for levels of land holdings* $l_{small} < l_{large}$*, given everything else the same, we get:*

$$l_p^*(l_{small}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_c^*(l_{large}).$$

*Proof* See Appendix A.4.

While under-usage is mainly driven by the inability to use the land as collateral, over-usage results from the inability to rent out land under customary tenure. Given that households that operate customary land do not receive any income if they decide not to use the land and the agricultural production function is increasing in land, they always prefer to operate the entire land holding. The effect will be the most pronounced for households with low agricultural productivity and large land holdings.

**Proposition 2.** *Denote optimal choices of land used by farmers who own the land under communal and private property rights regimes as $l_c^*$ and $l_p^*$, respectively. Then, if optimal land usage is lower than household land holding, $l_p^* < l_p$, and farmers' initial conditions in private and communal sectors of the economy are the same (i.e., the same amount of land, skills, and assets):*

$$l_c^* \geq l_p^*,$$

*and for the levels of agricultural productivity $z_{small} < z_{large}$, given everything else the same*

$$l_c^*(z_{small}) - l_p^*(z_{small}) \geq l_c^*(z_{large}) - l_p^*(z_{large}),$$

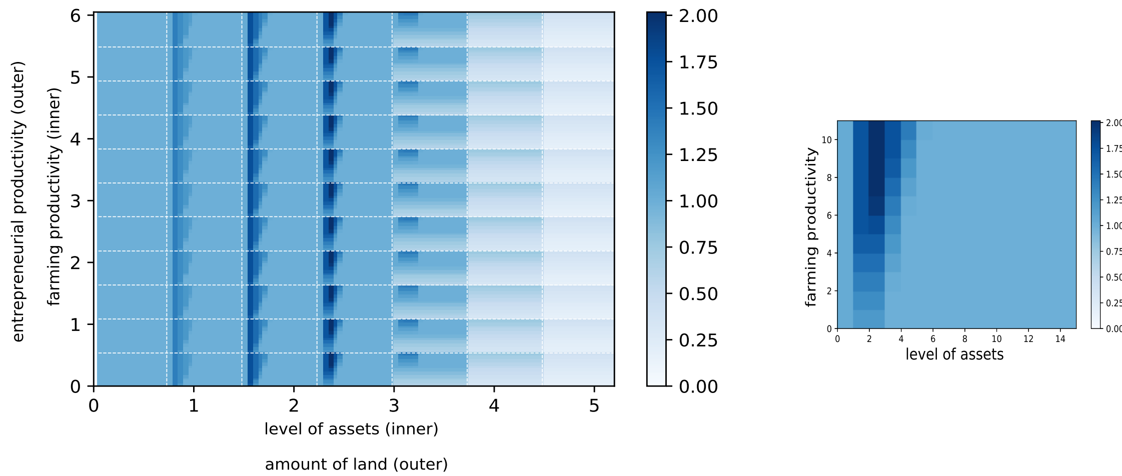
*and for the levels of land holdings $l_{small} < l_{large}$, given everything else the same, we get:*

$$l_c^*(l_{small}) - l_p^*(l_{small}) \leq l_c^*(l_{large}) - l_p^*(l_{large}),$$

*Proof* See Appendix A.4.

**Land property rights and occupational choice** Figure 1.3 documents occupational choices in parts of the economy under different land property rights regimes. In a frictionless world, households will choose their occupation based on the level of productivity in each sector. Similar to land misallocation, the presence of financial frictions distorts occupational choices for those households that are financially constrained irrespective of their land

Figure 1.2: Land Misallocation: Ratio of Land Usage by Farmers with Private Land Relative to Farmers with Communal Land




*Notes:* the right panel is a zoomed-in inner cell on the left (white dashed lines separate cells).

property rights regime. When the level of assets limits capital financing, those with high agricultural productivity might choose to become workers, and those with high entrepreneurial productivity might either stay in farming, which is less capital intensive, or become workers.

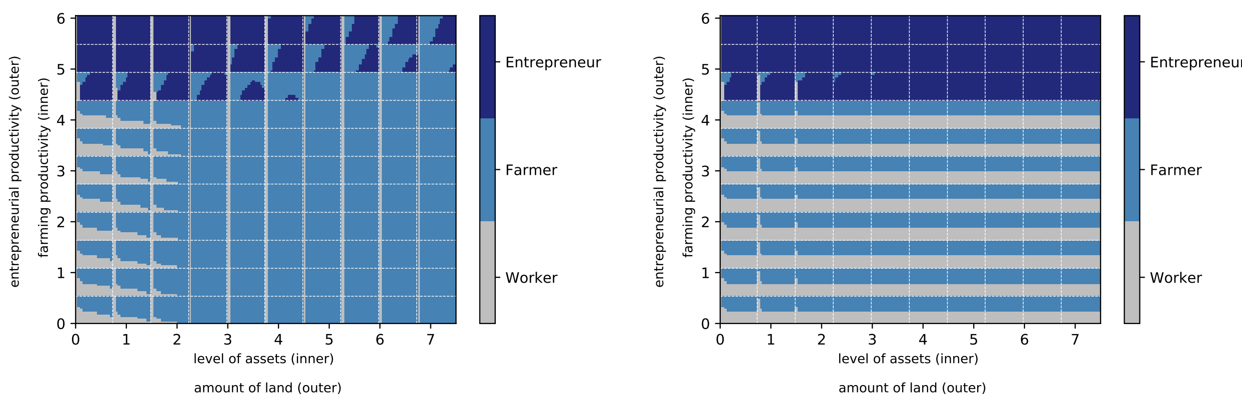
Land market imperfections would also lead to distortions in occupational choice in favor of farming, mainly through collateral and expropriation risk channels. The presence of expropriation risk prevents households from moving from farming to other sectors of the economy. The threshold of agricultural productivity when a household decides to move from agriculture to another occupation is much lower for those living under customary tenure relative to private tenure for all levels of assets and land. The risk of losing land in the next period and the probability of receiving the lump-sum land transfer incentivize households with relatively low agricultural productivity to remain in farming. Moreover, the agricultural productivity threshold goes down as the size of the owned land increases and, hence, potential land loss in case of expropriation. In the modern part of the economy, the agricultural productivity threshold is independent of the size of land owned by the household when their financial constraint does not bind.

Moving from worker or farmer to entrepreneur is limited by the collateral channel. Households with a low level of financial assets but sizable land holdings can finance their capital using land as collateral if their land is under a modern tenure system. This allows them to start their own business and switch to entrepreneurship. This option is not available for households whose land is under the customary system, so they are forced to stay in agriculture or become workers.

Finally, the inability to rent out your land leads to lower non-occupational income compared to the modern property rights regime, making non-agricultural occupations less attractive.

Figure 1.3: Occupational Choices

(a) Communal Land Holders (b) Private Land Holders




## 1.5 The Effect of Policy Interventions in Estimated Model

I now present a quantitative exploration of the aggregate and distributional impact of improvements in land property rights by moving from the economy with different tenure regimes, customary and modern, to an economy with only private land. In the model, customary land differs from private land in three different ways: i) it cannot be rented out, ii) it cannot be used as collateral, and iii) it is subject to expropriation risk. To better understand the impact of each channel on the economy, I conduct a set of experiments, where I remove only one type of friction at a time and explore the general equilibrium effects. I also compare the impact of land reform and financial reform, and finally, I look at the transition path of the model economy from the initial steady state to a steady state after land reform took place.

### 1.5.1 General Equilibrium Impact of Land Reform

Figure 1.4 presents the long-run general equilibrium effect of a land reform that transforms all communal land to private land. The four panels compare economic outcomes of the baseline calibrated economy with 80 percent of communal land and the economy after land reform. The impact of land reform is positive for both agricultural and non-agricultural output, as well as welfare, measured by real consumption. Moreover, it leads to a smaller share of labor remaining in agriculture and more entrepreneurs.

The top left panel documents changes in prices. An increase in the real interest rate is due to increased demand for capital as the budget constraint is relaxed for land owners under customary tenure before the reform. At the same time, the ability to rent out land results in higher land utilization and a drop in the rental rate of land. Finally, a wage increase is driven by increased demand for labor from entrepreneurs due to the higher amount of capital used as well as higher levels of entrepreneurship. At the same time, both farming and entrepreneurship become more attractive, putting pressure on the supply of workers and, hence, pressure on wages.

The left bottom panel presents the impact of land reform on labor shares for each occupation. Despite lower input price of land and, hence, higher attractiveness of agriculture, farmers' share in the economy decreases by 8.6%. A substantial increase in wage and absence of expropriation risk leads to an increase in the share of workers, while more relaxed collateral constraints increase entrepreneurship by 5.8%.

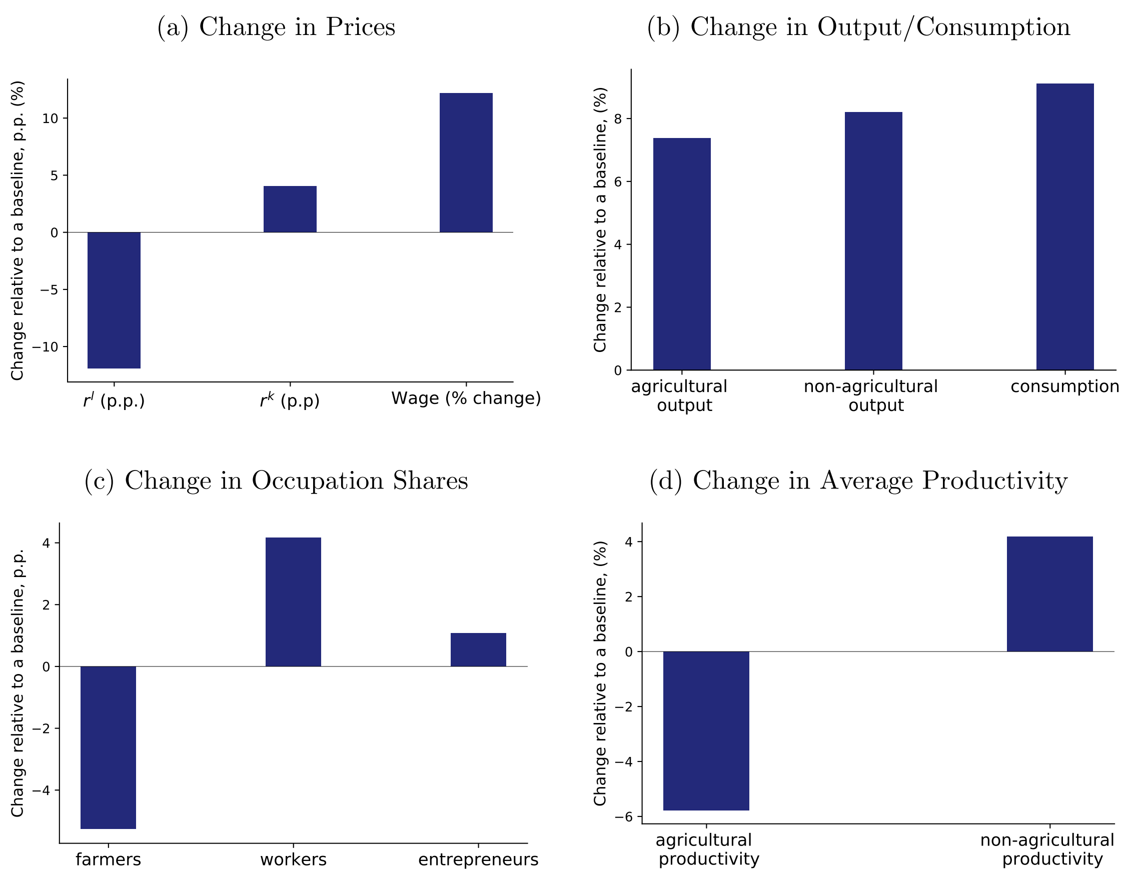
Output, both agricultural and non-agricultural, increases, as well as consumption. An increase in agricultural output by 7.4% is driven by higher land utilization and more efficient land allocation across farmers. Although the average agricultural skill of a farmer decreases, aggregate agricultural productivity measured by output per farmer increases by 17.5%.[28] Non-agricultural output increases by 8.2% due to both higher levels of inputs, labor and capital, and level of average entrepreneurial skill. Moreover, consumption increase is more significant than the increase in total output due to a lower level of households' savings driven by higher financial inclusion and better allocation of capital across households.

**Partial vs General Equilibrium** The importance of general equilibrium forces for aggregate effects of land reform is illustrated in the Figure A.7 in the Appendix. Agricultural output and employment decline substantially in partial equilibrium as households move to higher-income sectors. However, in general equilibrium, an increase in the interest rate of capital and wage encourage households with relatively high agricultural productivity to remain in agriculture. Moreover, a substantial decline in the rental rate of land makes agriculture more profitable, preventing the outflow of farmers to other sectors in general equilibrium setting.

**Distributional impact** While land reform leads to a higher level of consumption and welfare, these gains are not evenly distributed. Figure 1.5 shows the distribution of welfare gains and losses across households that were under customary and private land property rights before the reform. The gains are measured in equivalent consumption units. The

[28] Average agricultural skill of farmers decreases as households with both high agricultural and non-agricultural skills living in a communal part of the economy move from farming to entrepreneurship.

Figure 1.4: The Effects of Land Reform




(a) Change in Prices

(b) Change in Output/Consumption

(c) Change in Occupation Shares

(d) Change in Average Productivity

*Notes:* Plot (d) depicts the change in average productivity of employed farmers and entrepreneurs.

figure shows that a majority of households under the communal tenure system gain from land reform. There is empirical evidence that a significant fraction of households does realize economic gains of titled land. According to the last wave of the Household Survey, the majority of households that do not have any land certificate said that they would like to obtain one and are willing to pay for it (90.3% and 75.1%, respectively).

In the communal part of the economy, the gains are the largest for those with large land holdings. Now, they can use the land as collateral, receive rental income from unused land, and move to the occupation, where they are the most productive. Moreover, those gains are increasing in entrepreneurial productivity and decreasing in the level of financial assets.

Those with a low level of assets gain relatively more as they face a tighter financial constraint. Those with relatively large land holdings and high entrepreneurial productivity gain more than low productivity entrepreneurs, as now they switch from farming to entrepreneurship due to the absence of expropriation risk.

Precisely the opposite situation obtains for the initially private land holders: those with large land holdings experience welfare losses due to a drop in the land rental rate. For the originally private land holders, the most gains are observed for households with relatively little own land who stay in farming and need to rent in some land due to a decrease in the rental rate for land. The gain is higher for those with higher agricultural productivity.

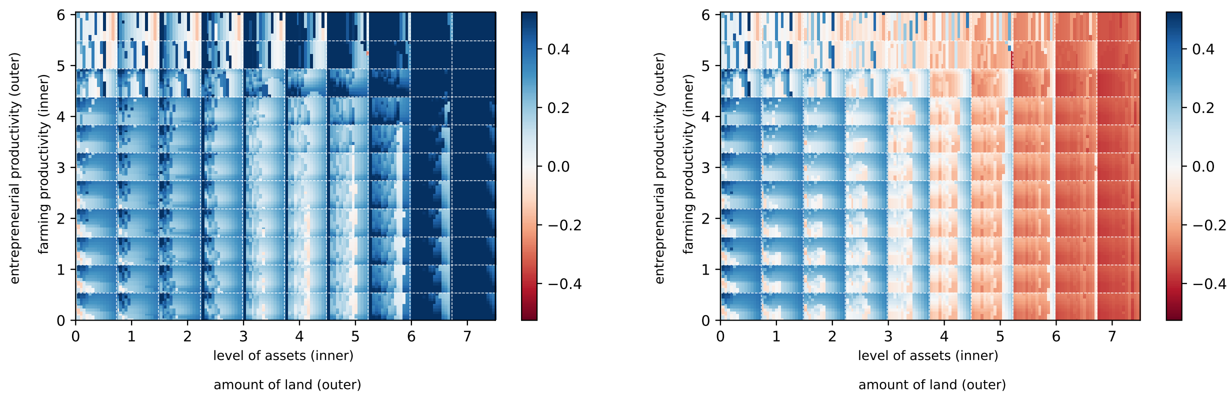
In sum, I find substantial welfare gains, especially for those in the communal part of the economy with a low level of assets. In addition, those with a high level of assets benefit from a higher rental rate of capital, while those with large holdings of private land experience losses. Moreover, consumption increases for many households due to higher levels of financial inclusion, and, hence, lower level of savings. Given that welfare gains are the largest among households initially belonging to the communal part of the economy, and consumption changes are favorable for the poorest households in terms of assets and land holdings, overall consumption inequality slightly decreases, with the Gini index declining from 30.9 to 29.6 for consumption.

### 1.5.2 Decomposing Impact of Land Reform

Given that there are three main differences between customary and modern land tenure regimes, I explore the effects of each channel separately. I perform a decomposition analysis of different channels of land reform by looking at the impact of removing only one friction at a time. Such a decomposition is important in the context of low-income countries as reform implementation often faces numerous challenges due to the presence of imperfections

Figure 1.5: Changes in Welfare Distribution

(a) Communal Land Holders (b) Private Land Holders




in other markets.[29] Three channels that are studied: i) expropriation risk, ii) inability to use land as collateral, and iii) inability to rent out land.

Figure 1.6 presents the general equilibrium effect of each channel of land reform on economic outcomes. Lower expropriation risk pushes households from agriculture to other occupations, leading to a higher rental rate of capital and lower wages. The increase in demand for workers, driven by households joining entrepreneurship, is smaller than the increase in the supply of labor driven by higher attractiveness to be a worker. A decrease in the number of farmers and lower average agricultural skills of farmers lead to a decline in agricultural output. An increase in average entrepreneurial productivity, and reduction of agricultural productivity, are driven by marginal entrepreneurs who have both relatively high agricultural and entrepreneurial productivity but remain in farming due to expropriation risk.

The ability to use the land as collateral creates demand for capital from farmers and entrepreneurs. As a result, the rental rate of capital increases, which pushes away some people from agriculture and business. Therefore, the supply of workers increases, but by a

[29] For example, the collateral channel might not work because banks would not be willing to accept land as collateral due to the limited land market.

smaller amount than the demand for workers driven by larger capital inputs of entrepreneurs. To clear the labor market, the wage increases. The effects on output and average productive skills are similar to the expropriation channel but larger in magnitude as the collateral channel has a more significant impact on capital and labor inputs.

Allowing households under customary tenure to rent out land increases land supply and land utilization. As a result of the larger supply, the rental rate of land drops, which attracts more households to agriculture. Higher land utilization also creates demand for capital, and the rental rate of capital slightly increases. To prevent the outflow of workers, the wage increases. The average productive skills of farmers increases as land is reallocated from less productive to more productive households. Higher inputs and average productivity increase agricultural output.

### 1.5.3 Land Reform vs Financial Reform

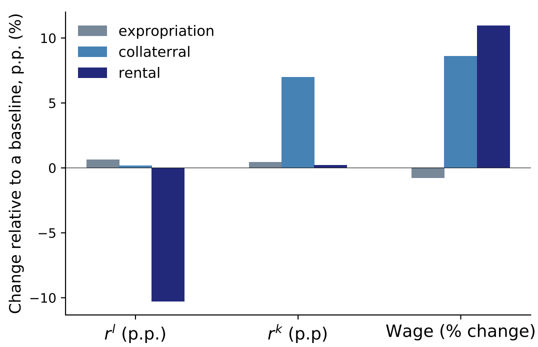
One of the channels through which land reform affects the economy is by allowing the use of private land as collateral. As a result, land reform also facilitates financial inclusion among poor households who own some land. Given the interaction between land property rights and the financial sector, I compare a land reform's impact on the economy with the effects of financial reform. To compute the effect of financial reform, I relax financial constraints in a way so that the loan to collateral value is equal to the level of the advanced economy – Sweden (83.9%).[30]

Figure A.8 compares the effects of land reform and financial reform. Given that it is impossible to perform two numerically equivalent reforms in different sectors, I cannot compare the magnitudes of economic outcomes changes. But it is worth exploring the direction of changes. In terms of prices, financial reform has a minor effect on land rental rate as land

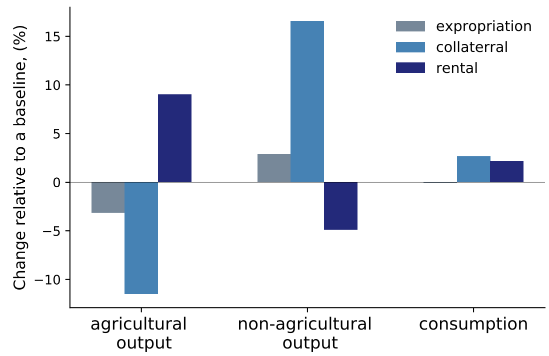
[30] I use Sweden to be consistent with the parameter I use for $\lambda_k$ in the baseline model, given that Sweden is the only advanced country that is present in the World Bank's enterprise survey.

Figure 1.6: Decomposition of Land Reform

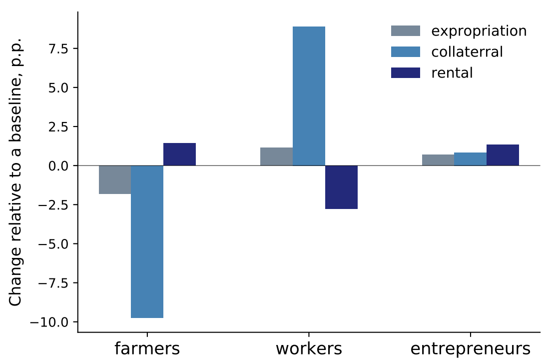
(a) Change in Prices



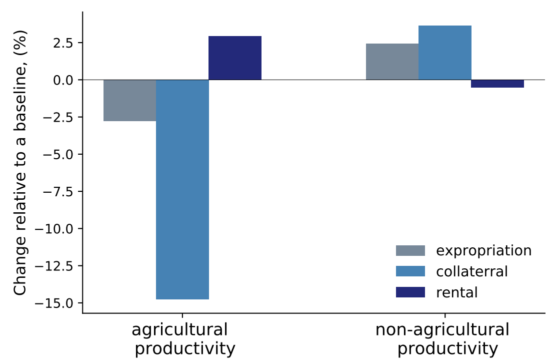

(b) Change in Output/Consumption




(c) Change in Occupation Shares




(d) Change in Average Productivity




supply does not change. The small drop in $r^l$ is driven by lower demand for land as some households move from agriculture to other sectors. Both consumption and non-agricultural output increase in the case of both reforms as households move from farming towards entrepreneurship and use more capital due to more relaxed financial constraints. However, financial reform leads to lower agricultural output as a lower share of households remains in agriculture, and average productivity in this sector decreases.

To sum up, the qualitative impact of financial reform on economic outcomes is the same as the impact of the collateral channel of land reform but differs from land reform as a whole. Moreover, distributional impacts are different (Figure A.9). In the case of financial reform, those who are marginal entrepreneurs and existing entrepreneurs with positive assets who are financially constraint benefit the most. In contrast, those operating communal land do not benefit significantly more than those operating private land, as we observe in the case of land reform.

### 1.5.4 Postreform Transition Dynamics

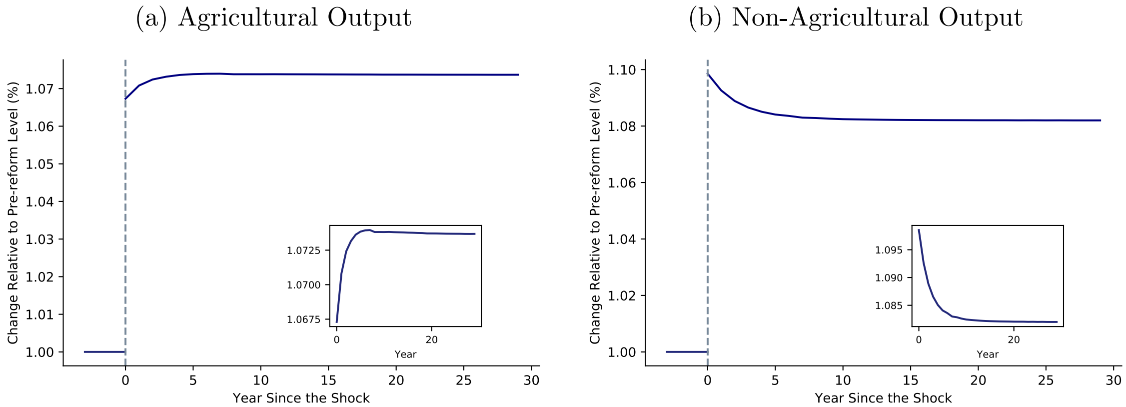
In this exercise, I study the transitional dynamics triggered by a sudden unexpected land reform that removes all land market frictions. I assume that financial frictions remains the same throughout the transition period.

Figure 1.7 shows the evolution of agricultural and non-agricultural output along with the transition to the new postreform steady state. The economy moves into the neighborhood of the new steady state in 20-25 years. However, the majority of changes happen in the first ten years after the reform. We observe a substantial initial increase in agricultural and non-agricultural output due to higher land utilization and relaxation of financial constraints, leading to more capital used in the production. While agricultural output continues to increase in the following years, non-agricultural output experiences some decline compared to the initial jump. The removal of land market frictions explains such dynamics that move labor from agriculture to other occupations, accompanied by a slow increase in prices of production for the non-agricultural sector, wage, and capital interest rate (Figure A.10).

## 1.6 Concluding Remarks

The prevalence of communal land tenure system in low-income countries is of first-order importance for the macroeconomic development of these economies. Such a system leads to both misallocation of factors of production and distortions in households occupational choices. Moreover, since communal land could not be used as collateral, such a tenure system amplifies financial market frictions widespread in developing countries. In this paper, I study

Figure 1.7: Postreform Transition Dynamics for Output




*Notes:* The output series are normalized by their respective prereform values.

the effect of land property rights on aggregate productivity and allocation of resources, and the impact of financial and land market frictions on the economic development of low-income countries.

To assess the aggregate and distributional impacts of economy-wide land reform, which eliminates the customary tenure system, I develop a general equilibrium model that features frictions of both land and financial markets. I leverage detailed panel household data from Tanzania in two ways: i) to discipline the model and ii) to show that the presence of insecure land property rights and limited access to credit is associated with resource misallocation in agriculture. Using a quantitative model, I find that land reform has positive effects on agricultural and non-agricultural output and leads to occupational shifts of households away from agriculture. Moreover, land reform increases the level of financial inclusion, especially among the poorest households with limited financial assets.

To sum up, this paper points to the significant potential gains from land reform that improves land property rights. Not only do stronger land property rights lead to higher welfare and more efficient allocation of resources, but they also help to create a more financially inclusive society. However, welfare gains of land reform, measured in consumption equivalent changes, are not evenly distributed, and some households experience losses. The main losers of such reform are large private landholders, suggesting that political economy barriers might prevent or slow the progress of land reform, despite its potential benefits.

# Chapter 2

# Global Innovation Spillovers and Productivity: Evidence from 100 Years of World Patent Data[1]

## 2.1 Introduction

Productivity is a key driver of economic growth within and across countries. Clark and Feenstra (2003) and Klenow and Rodríguez-Clare (1997) document that the majority of the divergence in income per capita over the 20th century can be attributed to cross-country differences in total factor productivity (TFP) growth. The endogenous growth literature, starting with the seminal contributions of Romer (1990) and Aghion and Howitt (1992), has emphasized the role of innovation and idea generation as a central driver of technology and, ultimately, productivity growth. However, from an empirical point of view, direct measures of innovation that cover a large number of technologies, countries, and time periods are scant.[2]

[1]We thank Isaac Baley, Joel David, Matthias Doepke, Ruben Gaetani, Giammario Impullitti, Ben Jones, Nan Li, Joel Mokyr, Dimitris Papanikolau, Sergio Petralia, Thomas Sampson, and seminar attendees at CREI, LSE, UCSD-UCLA-UCB trade conference, Midwest Macro, Nottingham, SED, UB, UCSD, NBER SI Growth Meeting, and Northwestern for helpful comments and discussions. We are also grateful to Bart Hobijn for his discussion of our paper. The views expressed are the authors' and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System. All errors are our own.

[2]See Comin and Mestieri (2014) and the references therein documenting the diffusion of major technologies since the Industrial Revolution. Comin and Mestieri (2018) show that the productivity transitional dynamics implied by the observed diffusion patterns match well the evolution of the distribution of cross-

In this paper, we use historical patent data spanning a vast range of countries over the past one hundred years to study the evolution of innovation across time and space. The use of patent data allows us to exploit a widely validated quantitative measure for the generation of new ideas and knowledge spillovers (i.e., how innovation builds on previous knowledge). We document a substantial rise of international knowledge spillovers since the 1990s mostly driven by the United States and Japan, as well as the rise of innovation related to computation, information and communication technologies (ICTs), and medicine. We also leverage the rich structure of citation linkages across time, space, and fields of knowledge (FoK) to propose an identification strategy to quantify the effect of innovation induced by knowledge spillovers on productivity and economic growth across countries and industries. To the best of our knowledge, our identification strategy is novel to the endogenous growth literature.

We build our measure of innovation using patent data collected from the European Patent Office Worldwide Patent Statistical Database (PATSTAT). PATSTAT contains bibliographical and legal status information on more than 110 million patents from the main patent offices around the world, covering leading industrialized countries, as well as developing countries, over the period 1782-2018. To avoid some of the arbitrariness of using broad patent technology classes (Keller, 2002), we classify patents into fields of knowledge that we obtain with a machine-learning approach. Based on the premise that knowledge is embedded in inventors, the algorithm first calculates the probability that the same inventor patents inventions in multiple technology classes. It then uses these probabilities to infer the proximity of technology classes in the knowledge space and to create knowledge clusters.[3]

Armed with our newly defined technology classes, we show that their significance—as

country income per capita in the past two centuries. Their analysis is circumscribed to 25 major technologies since 1780.

[3]As a robustness check, we also perform a clustering analysis where the strength of the linkages between different patent classes is based on citations and/or co-appearance of these classes on the same patent grant.

measured by the share of patents across fields of knowledge—has importantly evolved over time. The data reveal substantial technological waves in the past one hundred years. For example, mechanical engineering accrued the largest share of innovations near the beginning of the 20th century. Fields of knowledge related to chemistry and physics (e.g., macro-molecular compounds) were the most prominent fields around the mid-century mark, while inventions related to medicine and the digital economy appear to be the most prevalent at the end of the 20th century and over the most recent decades. We also show that while advanced economies account for the bulk of patenting activity, there is substantial variation in terms of countries' specialization across fields of knowledge. Moreover, these patterns of specialization are heterogeneous over time.[4]

Next, we turn our attention to knowledge spillovers. We measure knowledge spillovers through patent citations across fields of knowledge and countries. For this exercise, we focus on the post-1970 sample for which we have data for virtually all countries in the world. We show that for the average patent, citations tend to be biased towards domestic, as opposed to international, inventions and toward patents within the same field of knowledge. We also document that across all these categories, there is an upward trend over time in terms of citations. That is, new patents tend to cite other patents more.

A striking fact has emerged since the 1990s. Except for the US and Japan, international citations have grown *faster* than domestic citations. After the year 2000—excluding the US and Japan—international patents are cited more than twice as much as domestic patents. This finding suggests that the reliance on knowledge produced elsewhere—and particularly in the US and Japan—has markedly increased over this period of time. Even for technology leaders such as Germany and the United Kingdom (UK), foreign citations now account for most of the citations. The increase is mainly driven by a handful of fields of knowledge

[4] We also show that specialization in fields of knowledge tends to be clustered in space. Moreover, we document that inequality in patenting activity across countries has increased since the 2000s.

that are related to ICTs and medicine. This fact may be interpreted as a decline in the prominence of European inventions relative to their US and Japanese counterparts.

After having laid out these facts, we investigate the effect of innovation (as measured by patenting) on productivity and income using a novel instrumental variable strategy that we further discuss below. Our empirical specification is guided by a simple theoretical framework that incorporates patents and patent citations in a multi-sector growth model. Our baseline regression studies the effect of innovation induced by international spillovers on productivity in the latest part of the sample (2000-2014) for which we have high-quality data on cross-country sectoral value added and TFP, as well as factors of production. We use patent data starting in 1970 to construct our instrument for this exercise. We then extend our analysis back in time and study the effect of innovation on long-run income growth (for the periods 1980-2016 and 1960-2016), for which we use the full extent of our patent data to construct the instrument.

Simply correlating innovation and productivity or output per worker is problematic because of measurement error (which would generate attenuation bias), potential reverse causality, and the presence of unobserved factors affecting simultaneously patenting and the dependent variables. Examples of such factors include financial or external shocks that affect both the output of a country and the amount of innovation produced. We address these endogeneity concerns by constructing a shift-share instrument that leverages pre-existing knowledge linkages across countries and technologies and combines them with lagged foreign innovative output in other fields of knowledge and countries, in the spirit of Acemoglu et al. (2016) and Berkes and Gaetani (2022).

More precisely, we construct the instrument in two steps. First, we estimate the strength of the linkages across countries and fields of knowledge (measured by patent citations) in the pre-sample period. These linkages constitute our pre-determined *shares*. The *shifts* of our instrument for country and field of knowledge ($c_o$ and $k_o$, respectively) are given by the patents filed in all other countries ($c_d \neq c_o$) and fields of knowledge ($k_d \neq k_o$) in the sample. We are thus assuming that the probability that patents in ($c_d$, $k_d$) generate a patent in ($c_o$, $k_o$) can be inferred from the network of patent citations, and it is an increasing function of the strength of these links.[5] Applying this procedure recursively, we obtain a predicted number of patents for each country and field of knowledge.

Our main variable of interest is value added per worker by country and sector (measured from the World Input Output Database) over the 2000-2014 period. The regression model includes controls that vary at the country-sector-time level (e.g., sectoral capital and labor, along with differential country and sectoral trends). We find a robust effect of innovation on value added per employment growth. One standard deviation increase in patenting activity leads to a 0.078 standard deviation increase in output per worker growth (after partialling out the regression controls), implying an increase in output per worker growth of 1.1 percentage points. When we estimate the effect of innovation on TFP growth, we find a very similar result in magnitude, as implied by our theoretical framework.

We conduct a number of robustness checks to address concerns regarding the validity of the instrument, such as the existence of pre-trends or demand-pull anticipatory effects that might be correlated with the contemporaneous state of the local economy. To do this, among other things, we show that the pre-period productivity is uncorrelated with subsequent patent activity predicted by the instrument. We also "reverse" the network of citations that we used to measure knowledge spillovers and calculate the amount of innovation we would have expected to observe *in the past* if the patenting activity was driven only by future demand. Reassuringly, we find no evidence supporting this alternative hypothesis.

We conclude our paper by extending our framework to study the effect of innovation on long-run income per capita growth. In our first exercise, we reconstruct our shift-share

[5]In fact, we refine this procedure and extend this logic to higher-order linkages to create our main instrument (see Section 2.5).

instrument using pre-1980 patent data and estimate the effect of innovation on income per capita over the 1980-2016 period. Using pre-1980 data allows us to cover patenting activity of virtually all high-income and upper-middle-income countries (as defined by the World Bank). We perform a second exercise by estimating the effect on income per capita growth starting in 1960 and 1970. In this case, we construct our instrument using the pre-1960 and pre-1970 patent network, respectively. While covering a longer time span, we lose information on the patenting activity of many upper-middle-income countries. Despite this, we find a positive, significant effect of patenting on income per capita growth across these different time periods. An increase of one standard deviation in log patenting implies an increase in the growth of income per capita between 1.6 and 2.8 percentage points. The implied changes in growth rates represent 24% and 41% of a standard deviation in terms of income per capita growth, respectively.

**Related Literature** This paper relates to the vast and rich literature studying the link between innovation and productivity since, at least, the seminal work of Griliches (1979, 1986). Similar to Kogan et al. (2017), who find large positive effects of patented inventions on firm growth and productivity, we document positive effects of innovation on output and productivity growth at the country-sector level. Our instrumental variable approach leverages knowledge spillovers and the diffusion of technology as measured by patent citations. The existence of knowledge spillovers has been extensively documented (e.g., Jaffe et al., 1993, and Murata et al., 2014). However, most of this literature has focused on domestic spillovers, based on the premise that they are very localized. In this paper, we especially focus on international spillovers, which have also been documented to be quantitatively important (e.g., Eaton and Kortum, 1999; Keller, 2002; Keller and Yeaple, 2013; Buera and Oberfield, 2020; Keller, 2004; and Melitz and Redding, 2021 provide excellent surveys). We contribute to this strand of the literature by documenting an increase of international spillovers since the 1990s and by using international linkages to build our shift-share design and, ultimately, quantify the effect of innovation on productivity.

In addition, our paper contributes to a recent literature that uses historical patent data to shed light on various linkages between innovation and long-run outcomes, e.g., Nicholas (2010), Packalen and Bhattacharya (2015), Petralia et al. (2016), and Akcigit et al. (2017). One difference with most of this literature is that we extend our analysis beyond one country and aim to provide a global view. To the best of our knowledge, this is the first paper that uses the entire coverage of the PATSTAT database to study patenting activity. Regarding the goal of providing a global view, our work is perhaps closest to Bottazzi and Peri (2003), who use R&D and patent data for European regions in the 1977-1995 period to estimate research externalities.

This paper is also related to the growing literature that incorporates networks in the analysis of different aspects of economic growth and trade (e.g., Acemoglu et al., 2015; Oberfield, 2018; Liu, 2019; Baqaee and Farhi, 2019; and Kleinman et al., 2021). In this regard, our work complements the recent work by Ayerst et al. (2020) and Liu and Ma (2021), who use international patent data to study the diffusion of knowledge embedded in trade patterns and the design of optimal R&D policies in the presence of international knowledge spillovers, respectively. Finally, our network-based shift-share instrumental approach is related to a number of papers that have used the network structure of patent citations to construct shift-share instruments. Our approach is most similar to Berkes and Gaetani (2022), who construct a shift-share instrument leveraging the network of citations across US cities, and Acemoglu et al. (2016), who use a citation network to percolate sectoral innovations through the innovation network and illustrate how technological progress builds upon itself. Both papers focus on the United States.[6]

[6] A large number of papers have used more standard shift-share ("Bartik") instruments in the innovation and productivity literature. For example, Moretti et al. (2019) estimate the effects of R&D subsidies and Hornbeck and Moretti (2019) estimate the effect of TFP growth in manufacturing across US cities.

## 2.2 Data

### 2.2.1 Data Sources

In this paper, we measure new ideas through patent data and productivity through value added per worker and TFP. Patent data are collected from the European Patent Office's Worldwide Patent Statistical Database (PATSTAT, Autumn 2018 version). PATSTAT contains bibliographical and legal status information on more than 110 million patents from the main patent offices around the world, covering leading industrialized countries, as well as developing countries, over the period 1782-2018.[7] From PATSTAT, we collect information on patent filing years, inventor and assignee locations, citations, patent families, and technological classes. While PATSTAT provides the most comprehensive coverage of patenting activities worldwide, it also has some limitations (Kang and Tarasconi, 2016). The main limitation for our purposes is data availability in the earlier years. In fact, data along one or more dimensions are often missing for some countries in the years preceding 1970. We therefore split our sample into two groups of countries, which we use at different stages of our analysis. The first group is composed of six major technological leaders – the United States, the United Kingdom, France, Germany, the Soviet Union, and Switzerland – for which all the patents' characteristics required by our analysis are available at least since 1920.[8] The second group includes all the countries covered by PATSTAT and starts in 1970.[9] Appendix

[7]PATSTAT is increasingly popular in economics as it provides rich information on patents. Most of its use has focused on particular sectors, countries, or time periods. See, among others, Coelli et al. (2016); Aghion et al. (2016); Akcigit et al. (2018); Philippe Aghion and Melitz (2018); Bloom et al. (2020); and Dechezleprêtre et al. (2020).

[8]Note that to compare consistent geographical units over time, when appropriate, we aggregate the patents filed in the German Democratic Republic and the Federal Republic of Germany. Similarly, for the Soviet Union, we combine all the patents produced by Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan.

[9]For our empirical analysis, we exclude China from our sample because of a substantial rise in the number of Chinese patents since the third revision of the patent law in China in 2008. While we see a sharp increase in the total number of Chinese patents after the implementation of the new law, the same pattern is not

B.1 provides more information about the composition of the samples and summary statistics.

We assign each patent to a geographical unit according to the country of residence of its inventor(s). If this information is not available, we use instead the country of the assignee(s) or publication authority. When a given patent is associated with multiple inventors or applicants from different countries or territories, we assign weights to these patents. The weights are computed assuming that each inventor or applicant contributed equally to the development of the invention.[10] To avoid double-counting patents that are filed in more than one patent office, we restrict most of our analysis to patents that are the first in their (DOCDB) family. We further collect the full distribution of technology classes associated with each patent based on the International Patent Classification (IPC). For our analysis, we first consider all the fields at the four-digit level (e.g., A01B)—for a total of 650 classes—and we then cluster them into consistent groups following the machine-learning procedure outlined in Section 2.2.2. Finally, to capture when an idea was completed and abstract from potential bureaucratic delays that are orthogonal to innovative activities, in our analysis we use the patent filing years instead of the years in which patents were granted.[11]

We supplement the patent data with the World Input Output Database (WIOD, Timmer et al. 2015). This database provides data on prices and quantities of inputs, outputs, and trade flows covering 43 countries and the Rest of the World for the period 2000-2014. The data are classified according to the International Standard Classification revision 4 (ISIC) for a total of 56 sectors. Using the World Input-Output Tables (WIOT) for each set of countries, sectors, and years, we construct trade flows, gross output, intermediate purchases, and value added expressed in US dollars. Additionally, from the Socio-Economic Accounts (SEA) in

observed in the number of Triadic patents, which are made up of all the patents filed jointly in the largest patent offices, i.e., the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), and the Japan Patent Office (JPO). For more details, see Appendix B.1.1.

[10]For example, if a given patent has four inventors, one from the US and three from the UK, then the patent will be split between the US and the UK with weights of 0.25 and 0.75, respectively.

[11]We discuss in more detail our data construction procedure in Appendix B.1.1

the WIOD, we collect industry-level data on employment, capital stocks, gross output, and value added at current and constant prices. These data allow us to compute country-sector TFP paths and also to compute trade in intermediate and final goods across country-sector pairs.[12] Finally, we use data from the Maddison Project Database (Inklaar et al., 2018) for the historical analysis of income per capita growth presented in the Section 2.5.3.

### 2.2.2 Construction of Fields of Knowledge

Innovation is the process of creating new knowledge, potentially building on existing knowledge across different fields. To operationalize our goal of measuring innovation waves across time and space, we build on the vast existing literature that measures innovative activities through patent data. We propose grouping finely defined patent classes into broader "fields of knowledge," which taken together constitute what we refer to as the "technology space" of the world.[13]

We employ a novel approach to grouping patent technology classes based on inventors' information. Our procedure is based on the likelihood that the same inventor produces inventions associated with different patent subclasses. The idea is that because knowledge is embedded in people, it is possible to cluster fields of knowledge based on the IPC subclasses in which the same inventors tend to patent. More precisely, we build a probability matrix $T_{642\times 642}$,[14] where each element $(i, j)$ is the probability that an inventor patents in IPC subclass $i$ conditional on also having a patent assigned to subclass $j$.[15] For example,

[12]See details in Appendix B.1.2. In the Appendix, we also discuss the additional database we use (i.e., UNIDO INDSTAT2) for historical data on sectoral manufacturing output by country and the Penn World Data Tables.

[13]See Kay et al. (2014), Leydesdorff et al. (2014), and Nakamura et al. (2015) for alternative definitions of technology space based on patent technology classes.

[14]Eight IPC subclasses whose second level is 99 (i.e., "Subject Matter not otherwise Provided for in this Section") were excluded from the analysis because they are assigned to patents with no clear identified technology.

[15]The diagonal elements of the matrix, $i = j$, are set to be equal to one. Note that the so-obtained matrix does not need to be symmetric because different IPC codes might weight differently in terms of

a mechanical engineer specialized in brakes will most likely patent in IPCs B60T (Vehicle Brakes or Parts Thereof) and F16D (Clutches, Brakes), which our algorithm correctly bundles together.[16]

We then use a *k-medoids* clustering algorithm to group the IPC subclasses into knowledge clusters. We interpret each resulting cluster as a field of knowledge, and use this classification to analyze the evolution of patenting in the next section. The *k-medoids* algorithm minimizes the distance within clusters by comparing all possible permutations of subclasses, conditional on a specific number of clusters, $k$. To determine the optimal number of clusters, we first compute the optimal clustering for each possible $k$ and we then "score" each result using the silhouette coefficient. The score takes into consideration the distance between elements within a cluster, as well as the distance across clusters, while also penalizing the existence of singletons.[17] The optimal number of clusters implied by the silhouette coefficient is $k = 164$. Table **??** in the Appendix reports the subclasses assigned to each cluster.[18]

their importance and centrality relative to other IPC codes within a given field of knowledge. For example, according to the matrix, manufacture of dairy products (A01J) is closest to dairy product treatment (A23C), while dairy product treatment is closest to foods, foodstuffs, or non-alcoholic beverages (A23L).

[16]Other procedures for bundling patent classes have been proposed in the literature. One strand of the measures uses patent citation information (e.g., Zitt et al., 2000; von Wartburg et al., 2005; Leydesdorff and Vaughan, 2006; and Leydesdorff et al., 2014). We also conduct such grouping as a robustness check and find substantial overlap. Another strand of the measures uses the "co-classification" information of patents (Jaffe, 1986; Engelsman and van Raan, 1994; Breschi et al., 2003; Leydesdorff, 2008; Kogler et al., 2013; and Altuntas et al., 2015). Others used the likelihood of diversification as measures of distance (Hidalgo et al., 2007) and analysis of patent texts (Fu et al., 2012, and Nakamura et al., 2015).

[17]To apply the *k-medoids* algorithm requires the creation of a dissimilarity matrix $D$, which needs to be symmetric. To obtain such dissimilarity matrix, we apply the following transformation to the inventor probability matrix:

$$D_{ij} = 1 - (T_{ij} + T_{ji}) = D_{ji},$$

where each element in the dissimilarity matrix $D$ is interpreted as a measure of distance between subclass $i$ and subclass $j$. We use this matrix in our *k-medoids* clustering algorithm to group the IPC subclasses into clusters. More details on the procedure used to construct fields of knowledge can be found in Appendix B.1.4.

[18]As a robustness check, we also construct the proximity matrix based on the citation linkages instead, and apply the same procedure. The results are similar to the ones obtained with our proximity matrix: (i) the percentage of pairwise IPC subclasses that are in the same cluster is 50.6 (excluding singleton clusters, which accounts for 22.6 percent of all clusters); (ii) the percentage of pairwise IPC subclasses that are in the same cluster weighted by importance, measured by the number of patents in the respective subclass relative to all patents, in the sample is 51.9 (excluding singletons); (iii) the percentage of clusters' centers that are the same is 67.1.

## 2.3 Some Stylized Facts on World Innovation

We start our empirical analysis by presenting some stylized facts about the evolution of innovation and knowledge spillovers across time and space. We use the fields of knowledge created in Section 2.2.2 as our unit of analysis of the technology space.

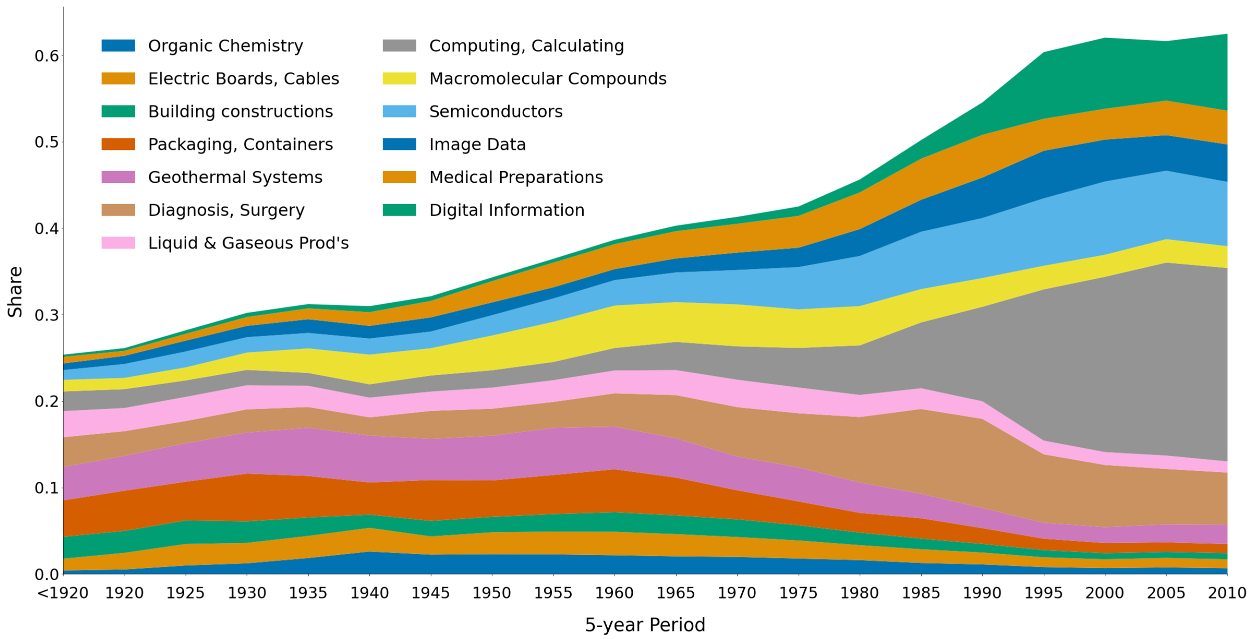
### 2.3.1 Evolution of Fields of Knowledge across Space and Time

We first document the evolution of the major fields of knowledge for the past hundred years and highlight how different countries contributed to their growth at different points in time. To measure the importance of each field of knowledge at any point in time, we compute the share of patents belonging to that field of knowledge. Each patent is weighted by the total number of forward citations.[19] We split our data set into nineteen 5-years periods from 1920 to 2015, plus a period prior to 1920 where we lump together all the patents filed before that year. For each time period, we rank every field of knowledge based on its relative contribution to the overall patent activity.

Figure 2.1 shows the evolution of the fields that were ever present in the top five fields at any point in time according to our measure. Two trends are readily noticeable. First, we observe a substantial increase in the concentration of innovation, especially around the 1990s – approximately 10% of the fields of knowledge account for 60% percent of all patent activity in the 2000s compared with 30% in the first half of the 20th century. Second, there is substantial heterogeneity in the evolution of fields of knowledge over time. At the beginning of the 20th century, fields of knowledge belonging to Mechanical Engineering and Transportation (Packaging & Containers; Geothermal Systems) are the most prominent fields. Starting in the 1950s, we observe a shift towards chemistry and physics (e.g., Macromolecular Com-

[19] Note that we are using only the first patent of the family. Moreover, if a patent belongs to multiple fields, we add a fraction of the patent to each field proportional to the number of IPC subclasses reported on the patents.

Figure 2.1: Evolution of Top Fields of Knowledge




*Notes:* This figure represents the share of each field of knowledge, measured by the number of first-in-the-family patents weighted by backward citations, in total patent activity across all fields in a given period of time. The width of the colored bars reflects the share of the knowledge field. Exact values for shares can be found in Table B.1.

pounds). Around the 1980s there was substantial increase in medical and veterinary science (e.g., Diagnosis and Surgery or Medical Preparation). Finally, and as expected, around the mid-1990s the fields of knowledge related to computing and communication techniques started playing a leading role in the innovation landscape.

We perform the same exercise using alternative measures of importance that address possible concerns related to, for example, heterogeneous patenting practices across countries or strategic patenting behavior that gained more prominence in the past few decades. To do this, we build importance measures that take into consideration country fixed effects or patents that were cited at least once. Table B.2 in the Appendix shows that these measures are highly correlated to our baseline.

Next, we turn to the spatial heterogeneity of innovation activities by studying the contribution of different countries to the growth of top fields of knowledge. We divide the sample into four periods: 1920-1945, 1945-1970, 1970-1995, and 1995-2015. We concentrate our analysis to the seven fields of knowledge that took a leading role based on the number of patents throughout the entire period of study. Similarly to what we did in Figure 2.1, we assess the contribution of each country by computing its patenting share in a certain field of knowledge.[20]

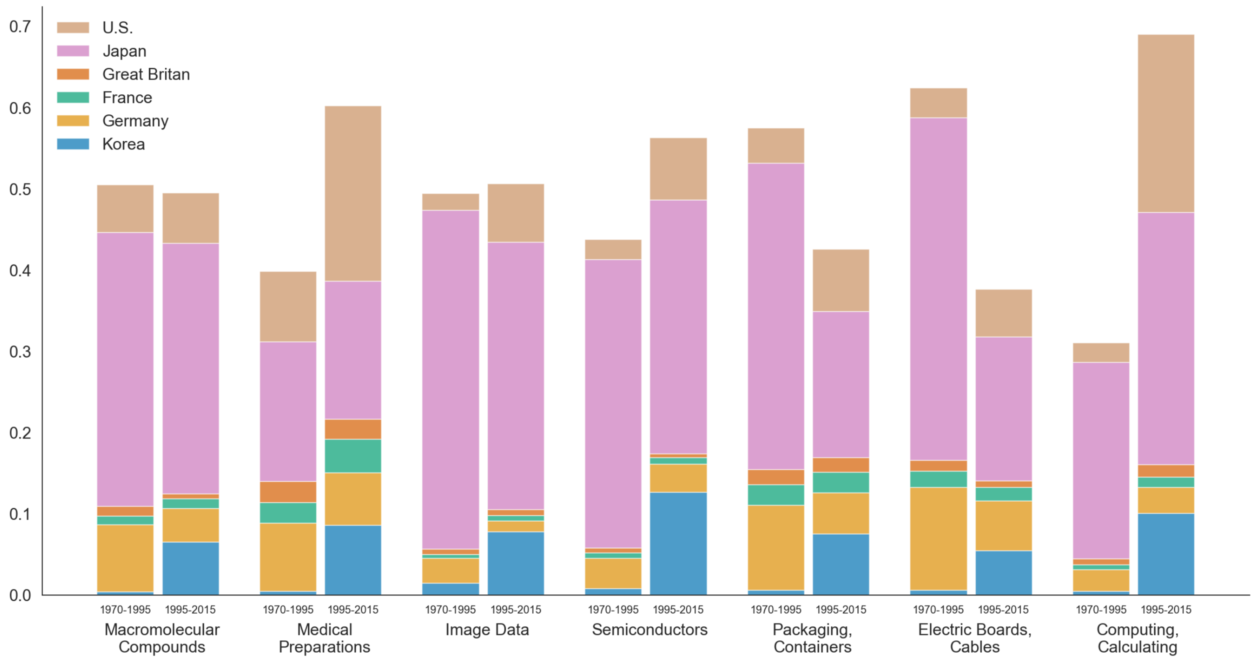
Because of data limitations, for the period 1920-1970, our sample is made up of just six countries: the US, the UK, France, Germany, the Soviet Union, and Switzerland. Figure B.5 in the Appendix shows that during this time period, the leading innovating role in major fields of knowledge was split between the US and Germany, followed by the UK and France. In fact, Germany overtook the US in every leading field in the period between the end of World War II and 1970.

In Figure 2.2, we consider the whole sample of countries in the years after 1970. Between 1970 and 1995, there are three clear technological leaders: Japan, the US, and Germany. The preponderant role played by Japan in the major fields of knowledge is remarkable. The US also gains substantial prevalence in the second part of the sample. After 1995 other Asian countries, such as South Korea, start rising to the forefront of the technological frontier. In this period, France experiences a decrease in importance in the innovation landscape. Asian countries dominate in the fields related to computing, engineering, and digital information, while their role in chemistry and medicine is less pronounced.

We extend our analysis beyond the top fields of knowledge and compute an overall ranking by averaging the country ranking across all fields of knowledge. This exercise paints a picture similar to the one in Figure 2.2. Japan and the US are the technological leaders from 1970 until 1995, with Japan falling behind after the 2000s. The Soviet Union's ranking is similar

[20]To account for potential differences in how countries assign patent citations, in this part of the analysis, we use the total number of patents without weighting by the number of citations for better comparability. We also verify that the results are robust to citation-weighted measures.

Figure 2.2: Countries Shares in Top Fields, 1970-2015




to the one of the US in 1970 and it declines subsequently, while Asian countries such as Taiwan gain prominence after the 2000s. See Section B.2 in the Appendix for further details and discussion of this exercise.[21]

### 2.3.2 Using Citations to Measure Spillovers across Time and Space

So far, we have shown that there is substantial time variation in terms of the composition of the technological output and in terms of the geographical contribution to worldwide innovation. We now turn our attention to knowledge spillovers. We measure spillovers through patent citations across fields of knowledge and countries. There is an abundant

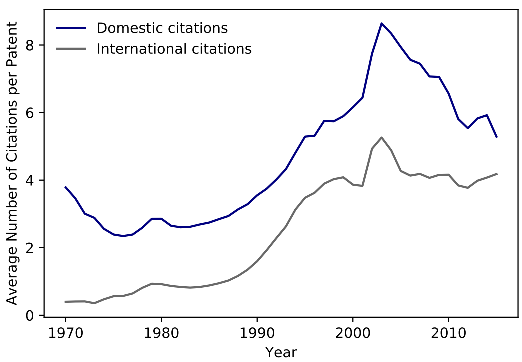
[21] In the Appendix, we report two additional results that shed more light on the spatial heterogeneity of innovative activities over time. First, we decompose inequality in innovation within and between countries, and find that the inequality in patenting across countries has increased since the 2000s, while the within component has remained mostly stable. Second, we use a gravity-type regression to estimate the relationship between gross domestic product (GDP) per capita, geographical distance, and production of technologies. We find that changes in patenting shares across fields of knowledge are correlated across countries that are geographically and linguistically close to each other.

literature studying within-country spillovers using patent citations (e.g., Jaffe et al., 1993, and Murata et al., 2014, for the United States), but the evidence on cross-country knowledge spillovers is more scarce. Despite being an imperfect measure of knowledge spillovers, patent citations provide a useful quantifiable benchmark that can be easily measured and used in our empirical analysis.

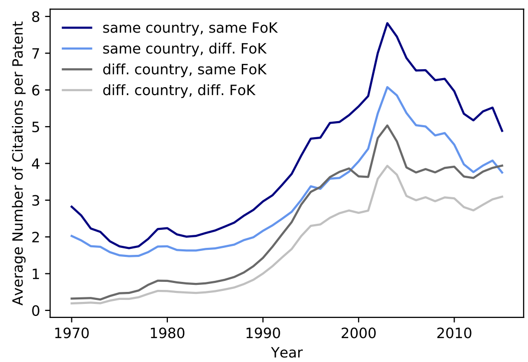
We focus our analysis on the post-1970 sample, for which we have data on filed patents for virtually all countries in the world. We compute the citations given by these patents to patents filed after 1900. Panel (a) in Figure 2.3 shows the evolution of the average number of citations given by patents filed after 1970. The average number of citations experiences an important increase starting around the 1980s. Domestic citations keep increasing up until 2002 and they then show a marked decline, whereas international citations plateau at about 4 international citations per patent in the late 1990s. A closer look at panel (a) further reveals that domestic patent citations tend to be more prominent than international patent citations: domestic patents are cited at roughly double the rate that of international patents are. Panel (b) breaks down these trends by additionally looking at whether citations belong to the field of knowledge (or FoK, as noted in Figure 2.3) of the citing patent.[22] The plot shows that citations tend to be concentrated not only geographically (i.e., domestic patents being cited relatively more), but also technologically (i.e., patents in the same field of knowledge being cited relatively more). Moreover, these gaps appear to have widened over the past decades.

An important pattern that is revealed by our analysis is that most knowledge (as measured by patent filings) is produced by a handful of countries – what we refer to as the "technological leaders." Specifically, as we have already seen in Figure 2.2, for the period 1970-2015 Japan and the United States are responsible for the largest share of patents pro-

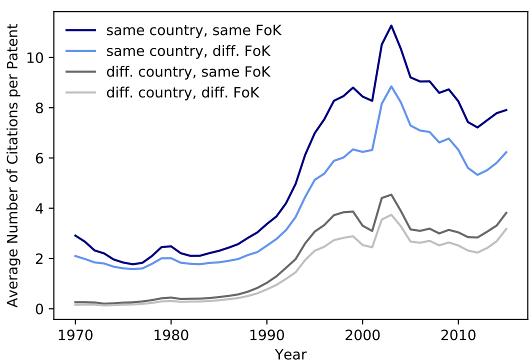
[22]The sum of the four lines in panel (b) is not equal to the total number of backward citations, since there is some double-counting due to the fact that cited patents belong to multiple fields of knowledge and (more rarely) to multiple countries.



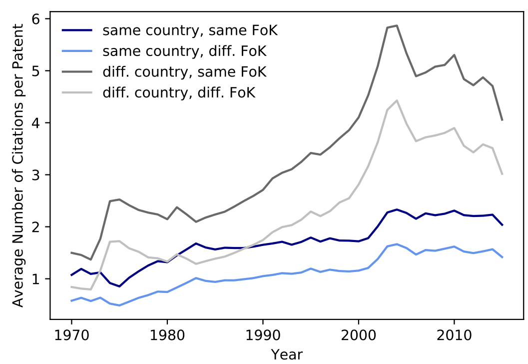

(a) Domestic vs. International Citations




(b) Citation Decomposition




(c) US and Japan

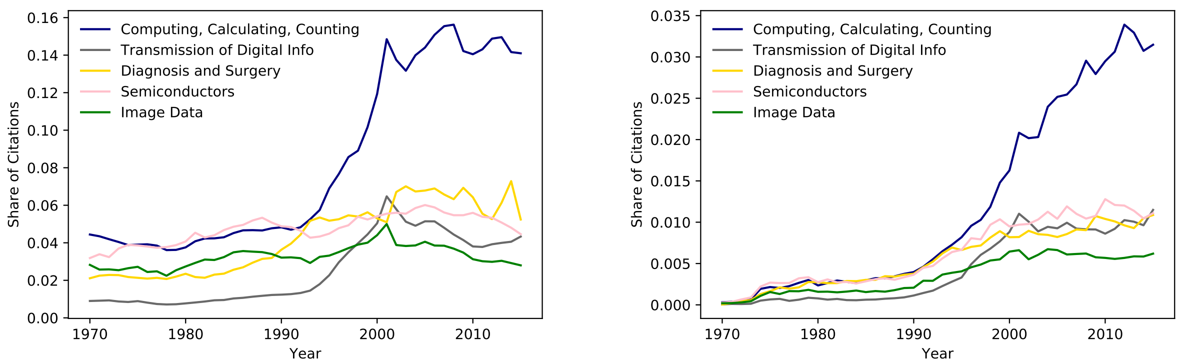



(d) All Countries except US and Japan

Figure 2.3: Citation Dynamics, 1970-2015

duced worldwide. Panels (c) and (d) of Figure 2.3 separately depict citation dynamics for Japan and the US and the rest of the world. While we observe an increase in the average number of citations per patent, there are two important differences between the two panels. First, the United States and Japan, on average, make more citations per patent than the rest of the world. Second, most of the citations in the US and Japan are given to domestic patents, while the rest of world mostly relies on knowledge produced in other countries, at least according to the data on patent citations.[23]

[23]Decomposition of citations for other countries, namely, Germany, France, and the UK, are reported in Figure B.6. The plots for these three frontier countries show how they moved from mostly relying on domestic knowledge in the early periods to foreign knowledge later in the sample.




(a) US and Japan

(b) All Countries except US and Japan

Figure 2.4: Share of citations to US and Japanese patents by FoK, 1970-2015. Each line in the plots represents the share of citations to US and Japanese patents that belong to a given field of knowledge. Panel (a) depicts the shares of domestic citations given by US and Japanese patents, and panel (b) depicts the shares of international citations received by patents filed in the US and Japan given by other countries.

Figure 2.3 depicts a rapid increase in the overall average number of citations per patent. To better understand what lies behind this increase, we concentrate on the backward citations received by the five leading fields of knowledge over the past five decades. Figure 2.4 shows that the substantial increase in the number of citations observed in Figure 2.3 is mainly driven by two fields of knowledge: Computing, Calculating, Counting and, to a lesser extent, Transmission of Digital Information. What is perhaps even more striking is the fact that most citations to this field of knowledge are given to US and Japanese patents, as illustrated by Figure 2.4.[24]

Taken together, the evidence presented in this section paints a picture consistent with the view that knowledge spillovers have increasingly become an important component of the innovation process in the past few decades. Although spillovers that originate from the same country and field of knowledge are still the most relevant, international knowledge spillovers have been steadily gaining importance over the past few decades. This increase is visible

[24]Similarly, Liu and Ma (2021) document a high reliance on domestic knowledge in both the US and Japan using Google Patents' global patent data for 40 countries during the period 1976-2020.

when considering spillovers coming both from the same field of knowledge and from other fields of knowledge, and it is mainly driven by a dramatic increase in the citations received by US and Japanese patents, especially in the fields of knowledge related to computing, information processing, and medicine.

## 2.4 Conceptual Framework

In this section, we present a framework that will guide our empirical analysis. This framework incorporates patents and patent citations to a standard, multi-sector growth model.[25] Importantly, our framework only specifies the production-side of the economy, and it does not assume the existence of a balanced growth path of output or productivity at the sectoral (or aggregate) level.[26]

Consider a world economy with $C$ countries, $S$ sectors, and $K$ fields of knowledge, where we index countries by $c$, sectors by $s$, fields of knowledge by $k$, and time by $t$. We denote by $N_{cskt}$ the stock of ideas available in country $c$, sector $s$, field of knowledge $k$, and time $t$. The state of world ideas at time $t$ is thus summarized by the vector $\mathbf{N}_t \equiv (N_{111t}, \ldots, N_{cskt}, \ldots, N_{CSKt})$. There is a production function for new ideas, $I(\cdot)$, that establishes the relationship between the flow of new ideas in a given field of knowledge and production sector, $\Delta N_{cskt}$; the current stock of knowledge, $\mathbf{N}_t$; and inputs devoted to generate new ideas, $R_{cskt}$;

$$\Delta N_{cskt} = I\left(S_{csk}(\mathbf{N}_t), R_{cskt}\right), \tag{2.1}$$

where $\Delta$ denotes the time difference operator between $t+1$ and $t$. The spillover function

[25] Our formulation builds upon previous studies that have been applied to the study of the patent network of citations, such as Acemoglu et al. (2016). Relative to Acemoglu et al. (2016), we present additional model elements to relate our results to TFP and output per capita and also extend the model to a multi-country setting.

[26] Unbalanced sectoral growth is indeed the empirically relevant case for the United States and other advanced economies (Comin et al., 2019).

$S_{csk}(\mathbf{N}_t)$ captures how the current world stock of knowledge $\mathbf{N}_t$ helps generate new ideas in country $c$, field of knowledge $k$, and sector $s$. We assume the spillover function to be

$$S_{csk}(\mathbf{N}_t) = \sum_{c' \in C} \sum_{s' \in S} \sum_{k' \in K} \alpha_{c's'k't} N_{c's'k't}, \tag{2.2}$$

where $\alpha_{c's'k't}$ captures the reliance of the production function of ideas in $csk$ on ideas from $c's'k'$ at time $t$. We leverage this structure to construct our instrumental variable. Note that we purposefully state Equation (2.1) generically so that it subsumes the first generation endogenous growth models as in Romer (1990) or Aghion and Howitt (1992), semi-endogenous growth models as in Jones (1995), Kortum (1997), or Segerstrom (1998), or second generation models as in Aghion and Howitt (1998), Young (1998), or Peretto (1998).[27]

Since ideas are to a large extent non-rival (Romer, 1990), the vast majority of endogenous growth theories resort to intellectual protection in the form of patents to ensure that investments in new ideas can be recovered with future profits.[28] This observation motivates our empirical strategy to proxy the generation of new ideas through patent filings. Patents provide a quantifiable measure over time and space that is arguably hard to obtain with other measures of ideas or innovation. Moreover, through citations, patents also provide an empirical measure of reliance on existing ideas across space and fields of knowledge. We rely on these spillover measures in our empirical analysis and, in particular, in our instrumental variables strategy. In practice, however, not all ideas are patented, and not all ideas which a patent builds on are cited. We thus think of patents as a *proxy* for new ideas, $\Delta N_{cskt}$, and citations as a *proxy* for spillovers. We discuss in the next section how our empirical specification addresses these potential discrepancies between idea generation and patenting.

In our framework, there is a representative firm in each country-sector that produces sec-

[27]For example, one specification extensively used in the literature (e.g., Romer, 1990, and Jones, 1995) ignores cross-country spillovers, and corresponds to having $S = K = 1$ and $S_c(\mathbf{N}_t) = N_{ct}$ and postulates a log-linear relationship, $I = N_{ct}^{\phi} R_{ct}$ with $\phi \leq 1$.

[28]See, among others, Aghion and Howitt (1998), Acemoglu (2009a), and the references therein.

toral output combining physical inputs (labor and capital) according to the best production methods available in that country-sector at time $t$, which are summarized by sectoral TFP, denoted $TFP_{cst}$. Sectoral value added per worker, $y_{cst}$, is given by the Cobb-Douglas production function $\log y_{cst} = \phi_{cst} + \log TFP_{cst} + \alpha \log k_{cst}$, where $k_{sct}$ denotes capital per worker, $0 < \alpha < 1$, and $\phi_{cst}$ denotes potential additional sources of variation of total productivity that are not captured by our framework. To obtain the baseline empirical specification, we assume that this effect denoted by $\phi_{cst}$ can be parameterized as a full set of dyadic fixed effects, $\phi_{cst} = \tilde{\delta}_{ct} + \tilde{\delta}_{st} + \tilde{\delta}_{cs}$. This parametrization captures the fact that the productivity of ideas (and/or other unmodeled sources of productivity differences) may differ across country-sector-time pairs because some country-sector pairs may be better suited at certain sectors than others (captured by $\tilde{\delta}_{cs}$), there may be some global technology trends affecting certain sectors (captured by $\tilde{\delta}_{st}$), or there may be some country-specific shocks (captured by $\tilde{\delta}_{ct}$).

Following the endogenous growth literature, we assume that the role of ideas is to increase firms' productivity by developing and improving methods of production (e.g., Acemoglu, 2009b). That is, we assume that there is a positive relationship between ideas produced and sectoral TFP growth. Moreover, as TFP grows and new production methods are implemented, we allow for the existence of fixed costs of adjustment scaling up with (a power function of) total output. This adjustment cost stands in for production disruptions related to the adoption of new technologies (e.g., as in Perla and Tonetti, 2014 or Comin and Gertler, 2006). In particular, our empirical specification assumes an iso-elastic relationship between TFP growth, ideas, and adjustment costs,

$$\log\left(\frac{TFP_{cst+1}}{TFP_{cst}}\right) = \phi_0 + \phi_N \log(1 + \Delta N_{cst}) - \phi_Y \log y_{cst}, \tag{2.3}$$

where $\phi_0, \phi_N, \phi_Y \geq 0$ and $\Delta N_{cst} = \sum_{k=1}^{K} \Delta N_{cskt}$ denotes the total number of ideas generated in country $c$ and sector $s$ at time $t$ across all fields of knowledge. Combining the idea production function, Equation (2.1), with the TFP Equation (2.3), we can readily verify that our framework nests a number of cases often considered in the literature, such as endogenous and semi-endogenous growth models.[29]

To derive our baseline empirical specification, we take the time difference in log-sectoral output between two adjacent time periods, $t$ and $t+1$. Combining the resulting expression with the law of motion for TFP, Equation (2.3), we find that

$$\log y_{cst+1} \quad = \quad \phi_N \log(1 + \Delta N_{cst}) + \phi_A \log y_{cst} + \delta_{ct} + \delta_{st}, \tag{2.4}$$

where $\delta_{ct}$ and $\delta_{st}$ denote country-time and sector-time fixed effects and $\phi_A = 1 - \phi_Y$. The focus of our analysis is on the effect of patenting on value added per worker. In our setting, this effect is captured by $\phi_N$, which corresponds to the elasticity of value added per worker growth on patenting. Note also that the country-sector fixed effect $\tilde{\delta}_{cs}$ appearing in our specification of the production function drops from Equation (2.4) because we take the time difference of log-sectoral output. In addition, note that the country-time fixed effect $\delta_{ct}$ absorbs the terms corresponding to sectoral capital-labor ratios (under the assumption of competitive markets for capital and labor across sectors). Since the assumption of factor markets being competitive may seem somewhat stringent, our we present empirical specifications that also

[29]Given our multi-sector, multi-country set-up, we find useful to separate the idea production function, Equation (2.1), which relates the evolution of the stock of knowledge across $cskt$ bins from the law of motion for TFP, Equation (2.3). Most models in endogenous growth theory do not present these equations separately. To relate our framework to the standard endogenous growth models, consider a one-country, one-sector and one-field of knowledge economy (or alternatively, a multi-country, multi-sector economy without spillovers across sectors and countries). Assume further that $TFP_{ct} = N_{ct}$, $\phi_0 = \phi_Y = 0$, $\beta_N = 1$ and that the idea production function (2.1) is $I = N_{ct}^{\phi} R_{ct}$ (as discussed in footnote 27). Then, we find that TFP growth is $\frac{N_{ct+1}}{N_{ct}} - 1 = N_{ct}^{\phi} R_{ct}$. For $\phi = 1$, the model generates the first-generation building-on-the-shoulders-of-giants dynamics (Romer, 1990), whereby the growth rate of $TFP_{cst}$ is directly controlled by the number of ideas produced at time $t$ with an elasticity of one. Letting $\phi < 1$ introduces the semi-endogenous growth fishing-out-of-the-same-pond effect so that increasingly more ideas become necessary to sustain constant TFP growth (Jones, 1995).

include as direct controls sectoral capital and (quality-adjusted) labor.[30]

## 2.5 Empirical Analysis

In this section, we empirically study the effect of innovation on productivity. We begin analyzing the effect of innovation on sectoral output per worker and TFP using cross-country panel data. We present our identification strategy in Section 2.5.1 and report our baseline results in Section 2.5.2. In Section 2.5.3, we extend our baseline estimation to a longer time horizon – at the expense of losing sectoral variation – where the dependent variable is output per capita.

### 2.5.1 Estimating Equations and Identification Strategy

Our baseline regression model closely follows Equation (2.4) and is specified as follows,

$$\overline{\log y_{cst+n}} = \phi_N \log\left(1 + pat_{cst}\right) + \phi_A \log y_{cst} + \phi_0 X_{cst} + \delta_{ct} + \delta_{st} + \epsilon_{cst}, \tag{2.5}$$

where $\overline{\log y_{cst+n}}$ is future annual output per worker in period $t+n$; $X_{cst}$ denotes a set of controls for country $c$, sector $s$, and time $t$; $\delta_{ct}$ and $\delta_{st}$ denote country-time and sector-time fixed effects; and $\epsilon_{cst}$ is the error term. The number of ideas in our model framework $\Delta N_{cst}$ is proxied by the number of first-in-the-family patents filed in $cst$. Thus, there is one departure relative to the model presented in the analytical framework. Rather than looking at one period ahead from $t$, we look at a measure of output per worker $n$ years ahead of period $t$. In particular, we take the three-year average annual output per worker as our

[30]Our framework implies that the lagged level of sectoral output per worker appears on the right-hand-side of Equation (2.4) with a coefficient $\phi_A = 1 - \phi_Y < 1$. This result follows from the lagged structure of the TFP, Equation (2.3), and it is not due to a log-linearization result around a steady state. The coefficient on lagged output per worker has been the focus of much of the cross-country growth literature. This coefficient is typically interpreted as proxying for convergence effects in regressions using aggregate data.

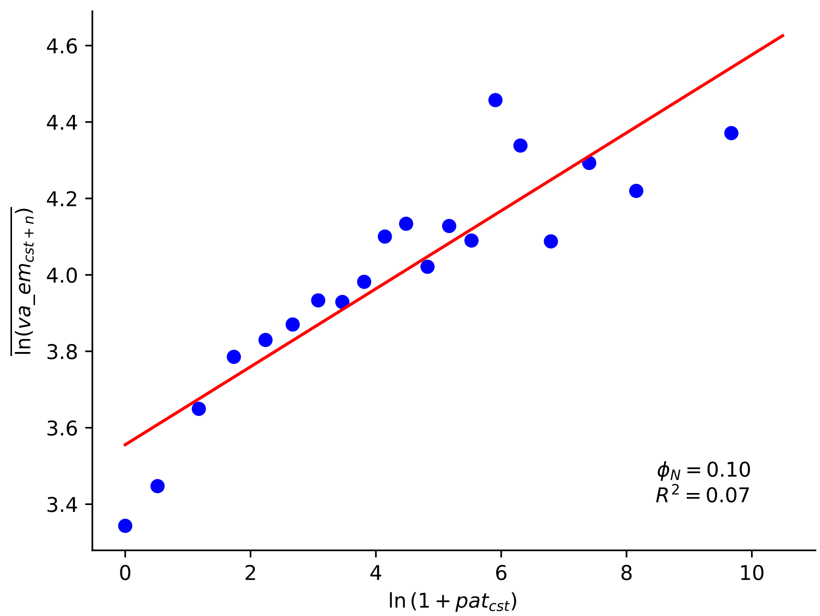
baseline measure (but we also show in the appendix that the results are robust to selecting any of these years in isolation, $n \in \{1, ..., 3\}$). We follow this approach, since it is common in the empirical growth literature to smooth out short-term fluctuations in the variable of interest and concentrate on longer-run trends (e.g., Arcand et al., 2015). Moreover, using this three-year average also alleviates the concern that the effect of a new patent may not be (fully) realized one year after its filing year.

The main coefficient of interest of our empirical equation, Equation (2.5), is the coefficient on patenting, $\phi_N$. It relates changes in the number of patents at the country-sector level in a given year to changes in output per worker in the following years, and it corresponds to the elasticity of output per worker growth to patenting. The presence of the fixed effects in Equation (2.5) follows from our conceptual framework. Intuitively, the inclusion of sector-year dummies controls for the fact that different industries may differently rely on innovation, as well as the fact that this relationship may vary over time. Sector-year dummies allow us to control for the presence of technological waves and other sectoral shocks that are common across all countries. Finally, the inclusion of country-year fixed effects controls, first, for the fact that different countries have different propensities to innovate, and, second, for any business cycles fluctuations at the country level (e.g., a financial crisis).[31]

Our main specification uses value added per worker from the World Input Output Database (WIOD). We also use TFP measures derived from the WIOD as part of our robustness exercises. The data used in our baseline analysis span from 2000 through 2014, and covers 36 countries and 20 sectors (see Appendix B.1 for more details). Figure 2.5 shows the binscatter plot of the raw correlation between patent activity, $\log(1 + pat_{cst})$, and value added per employment, $\overline{\log(va\_em_{cst+n})}$, over our sample period. In the cross-section of countries and sectors, a one percent increase in the number of patents is associated with a 0.10 percent increase in future output per worker averaged over the next three years. The

[31] As we showed in Section 2.4, country-sector fixed effects are differenced out.

Figure 2.5: Unconditional Correlation between Value Added Per Worker and Number of Patents




coefficient is statistically significant at conventional levels.[32]

To evaluate the strength of the causal relationship between innovation and productivity, we need to identify variation in patent activity that is orthogonal to unobserved factors that might affect innovation activity and productivity at the same time. There is a wide range of such possible factors and the direction of the bias is ex-ante ambiguous. An example of such factors is technological obsolescence of some industries. Reverse causality is also a concern – with higher productivity being the cause, rather than the consequence, of higher innovation activity in a given sector. Finally, estimates might be suffering from attenuation bias, due to presence of measurement error, given that patents are an imperfect measure of ideas and innovation.

[32]Standard errors are clustered at the country and sector level.

### Instrument Construction

To deal with these threats to identification, we build an instrumental variable for patenting activity in a given country and sector. Our instrument is based on a shift-share design that leverages pre-existing cross-country, cross-sector variation to predict the current level of patenting. We exploit the pre-determined network of patent citations during the period 1970-90 to identify knowledge links and construct the "share" component of our shift-share instrument. We then construct the "shifts" for the period 1990-2014 using a mix of the observed and predicted number of patents in other countries and sectors starting from the year 1980 on a rolling basis. Interacting the shares with the shifts and adding those up, we obtain the "predicted" number of patents in the period 2000-2014 as our shift-share instrument.[33] Thus, our instrument predicts patenting activity in the current period based on knowledge spillovers from other countries and sectors. In this sense, our shift-share design can be interpreted as a particular application of the linear knowledge spillover function presented in Equation (2.2) in Section 2.4.

Before delving into the details of the instrument, it is worth emphasizing that our proposed shift-share design differs from a more standard "Bartik" design. The reason is that we exploit the directed network of citations to construct linkages across country-sector pairs and then use shift terms that also vary at the country-sector level. In contrast, a standard "Bartik" would only use as sources of variation the own country-sector exposure (shares) and the world patenting activity in a sector (shifts). For our purposes, the standard Bartik design is unappealing, since it may confound innovation shocks with world industry or technological trends that also affect productivity.[34]

[33] As we will discuss in detail below, we only use "predicted" patents coming from lagged, pre-2000 patenting data as shifts to generate the instrument for our baseline sample.

[34] Consider, for example, a world where a few countries leaders determine in which sectors most of innovation activity is going to happen. In this case, the shift components that we would use in the construction of the instrument would not be orthogonal to either patent activity or productivity.

Next, we discuss in detail the steps we follow to construct our proposed instrument. To compute the "share" terms of our instrument, we gather patent information on the country of origin, technological field, backward and forward citations, and the sequence of the patent within its family (as described in Section 2.2) for all patents filed from $T_0^{share} = 1970$ to $T_1^{share} = 1990$. We use a correspondence from technological fields to industry codes to assign each patent to one or multiple sectors, with their respective weights in the latter case.[35] The underlying idea is to measure knowledge flows across countries and sectors through the share of citations that each patent produced in the country $c_o$ and sector $s_o$ of origin $o$ gives to patents in country of destination $d$, $c_d$, and sector, $s_d$. In particular, for each patent of sector $s_o$ belonging to country $c_o$ at time $t$, we calculate the share of citations given to patents produced in sector $s_d$ and country $c_d$ at time $t - \Delta$ for some citation lag $\Delta > 0$. We repeat this procedure for each time period $t$ between $T_0^{share}$ and $T_1^{share}$ and sum these shares to obtain the total number of citations over the $T_1^{share}$ to $T_0^{share}$ period. Importantly, to control for size effects due to the fact that some locations and/or sectors tend to patent more for idiosyncratic reasons, we normalize this measure by the total number of patents produced in the country-sector of the destination country $d$.

Formally, the entries of the adjacency matrix of the knowledge network for a citation lag $\Delta$ are given by,

$$m_{c_o,c_d,s_o,s_d,\Delta} = \frac{\sum_{t=T_0^{share}}^{T_1^{share}} \sum_{p \in \mathcal{P}(c_o,s_o,t)} s_{p \to (c_d,s_d,t-\Delta)}}{\sum_{t=T_0^{share}}^{T_1^{share}} |\mathcal{P}\left(c_d, s_d, t-\Delta\right)|}, \tag{2.6}$$

where $s_{p \to (c_d,s_d,t-\Delta)}$ denotes the share of citations that patent $p$ gives to patents of sector $s_d$ produced in country $c_d$ filed at time $t - \Delta$, $\mathcal{P}\left(s_o, c_o, t\right)$ denotes the set of patents in ($c_o$, $s_o$) at time $t$, and $|\mathcal{P}(\cdot)|$ denotes the total number of patents in the set (i.e., the set cardinality).

[35] We use Eurostat correspondence tables (Van Looy et al., 2014).

As the numerator shows, we add the citations of all patents originating in country-sector ($c_o$, $s_o$) at time $t$ over the time period from $T_0^{share}$ through $T_1^{share}$ going to patents filed in country-sector ($c_d$, $s_d$) at time $t-\Delta$, and normalize by the patent count in the destination country-sector at time $t-\Delta$. We use the resulting object $m_{c_o,c_d,s_o,s_d,\Delta}$ to construct the "shares" in our shift-share instrument.[36] Note that the "share" terms $m_{c_o,c_d,s_o,s_d,\Delta}$ do not need to add up to one, since their levels capture the number of citations from ($c_o$, $s_o$) that are typically received by patents filed in ($c_d$, $s_d$) with a lag $\Delta$.

Our network analysis also takes into account the fact that the speed at which ideas diffuse might differ across locations and sectors. We formally capture this effect by allowing the weights in our network to be time specific. We compute the citation shares at different time horizons, with citations lags $\Delta \in \{1, \cdots, 10\}$. In other words, we allow for the strength of the links to depend on how many years have passed between when the cited and citing patents were filed. In sum, our share terms are allowed to vary by country-sector citing-cited pairs and by time lag between cited and citing patents.

Our shift-share design is based on the idea that it is possible to predict the number of patents in a country and sector of interest based on pre-determined knowledge linkages. Intuitively, this approach mirrors the one of an input-output model for idea production except that it recognizes the non-rival nature of ideas (an idea in one country-sector can potentially spillover to multiple country-sector pairs). To carry out this approach, we then use as shift terms patents filed $\Delta$ years before the period of interest $t$ in other countries and sectors (or predicted patents as we explain below), and use the strength of the linkages to predict the number of patents filed in the country-sector of interest. We assume that

[36] As in Section 2.2, we restrict our sample to patents that are the first in their family to avoid double-counting of the same idea and capture only knowledge creation originated in a particular country-sector. However, for cited patents, we count all cited patents irrespective of whether they are the first or not in their family to capture all innovations on which any given patent builds on. We also note that Berkes and Gaetani (2022) show that the network of patents in the United Stated is stable in the time frame they consider, which roughly coincides with ours.

the strength of knowledge spillovers between country-sector dyads is mediated through how ideas in other country-sectors (as measured by our shift terms) diffuse through the knowledge network (as measured by the linkages $m_{c_o,c_d,s_o,s_d,\Delta}$). By interacting the shift and share terms and summing across countries, sectors, and diffusion lags, we then obtain a predicted number of patents $\widehat{pat}_{c_o,s_o,t}$ in country $c_o$, sector $s_0$ and time $t$.

Formally, our baseline shift-share design is constructed iteratively as follows. For 1990, we obtain predicted patents as

$$\widehat{pat}_{c_o,s_o,1990} = a_{1990} \sum_{s_d \in \mathcal{S} \setminus s_o} \sum_{c_d \in \mathcal{N} \setminus c_o} \sum_{\Delta=1}^{10} m_{c_o,c_d,s_o,s_d,\Delta} \cdot pat_{c_d,s_d,1990-\Delta},$$

where $a_t$ is a rescaling term that ensures that predicted number of patents is equal to the actual number of patents in period $t$ worldwide and $pat_{c_d,s_d,1990-\Delta}$ is the actual number of patents filed in $c_d, s_d, 1990-\Delta$.[37] Between 1991 and 1999 we construct the predicted number of patents using the previously computed predicted number of patents for years *since* 1990, and the observed patenting activity *prior* to 1990. That is, for $t \in (1990, 2000)$ we have that

$$\widehat{pat}_{c_o,s_o,t} = a_t \sum_{s_d \in \mathcal{S} \setminus s_o} \sum_{c_d \in \mathcal{N} \setminus c_o} \left( \sum_{\Delta=1}^{t-1990} m_{c_o,c_d,s_o,s_d,\Delta} \cdot \widehat{pat}_{c_d,s_d,t-\Delta} + \sum_{\Delta=t-1990}^{10} m_{c_o,c_d,s_o,s_d,\Delta} \cdot pat_{c_d,s_d,t-\Delta} \right),$$

where $\widehat{pat}_{c_o,s_o,t}$ denotes predicted patenting. Finally, starting in year 2000, we construct predicted patenting only leveraging the *predicted* patenting computed in the 1990s described above:

$$\widehat{pat}_{c_o,s_o,t} = a_t \sum_{s_d \in \mathcal{S} \setminus s_o} \sum_{c_d \in \mathcal{N} \setminus c_o} \sum_{\Delta=1}^{10} m_{c_o,c_d,s_o,s_d,\Delta} \cdot \widehat{pat}_{c_d,s_d,t-\Delta}.$$

To mitigate endogeneity concerns, the proposed shift-share design avoids using contemporaneous shares and shifts. First, to construct the share terms, we use the pre-sample

[37] Figure B.11 in the appendix shows a simple example of this procedure.

period 1970-1990 to construct the knowledge network. Second, when constructing the shift terms, we diffuse the observed patents filed pre-1990 over the period 1990-1999 to predict the patenting activity in the 1990s. We then use this predicted patenting activity to predict patenting activity over the sample period (2000-2014). Last but not least, we discard citations coming from the same country and from the same sector when we construct predicted patents. In other words, when calculating the $m_{c_o,c_d,s_o,s_d,\Delta}$ terms in Equation (2.6), we set the own-country and own-sector terms to 0,

$$m_{c_o,c_d,s_o,s_d,\Delta} = \begin{cases} 0 & c_o = c_d \\ 0 & s_o = s_d. \end{cases}$$

We exclude own country and sector to avoid endogeneity concerns arising from the fact that the links that connect the same country or sector might be correlated with future shocks (despite being at least 10 years apart).[38]

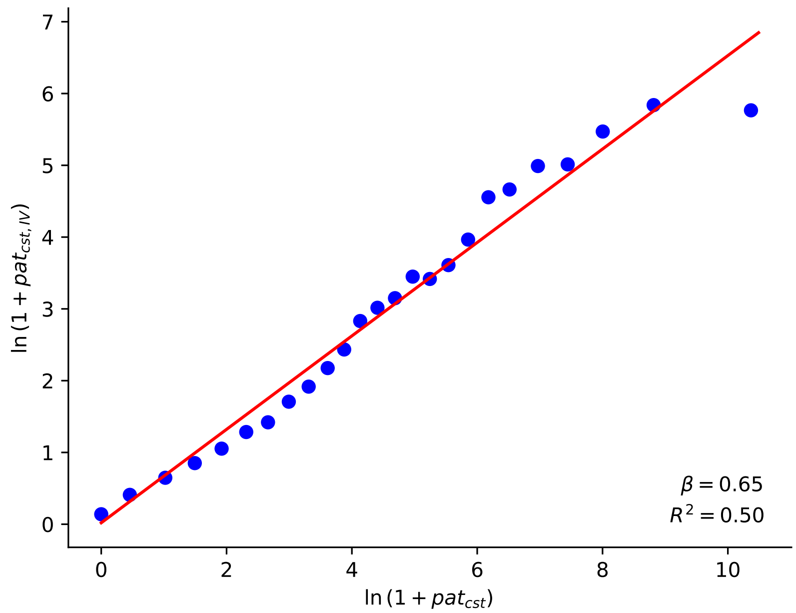
Figure 2.6 visually compares the actual and predicted number of patents through a binscatter plot. The two variables are strongly but not perfectly correlated: the coefficient of the regression is 0.65 and $R^2 = 0.50$. The Kleibergen-Paap Wald F statistics in the benchmark regression is 34, which rules out weak instrument concerns.

To provide evidence in support of our instrument, we report in the next section tests for a number of assumptions underlying the identification restrictions of shift-share designs, along the lines of Tabellini (2020).[39] First, the validity of the shift-share instrument rests on the assumption that countries and sectors giving more citations (to other sectors and countries) in the period between 1970 and 1990 are not on different trajectories in terms of the evolution

[38]For example, Cai and Li (2019) document the importance of multi-sector firm innovation using US patents, suggesting that some firms are able to internalize knowledge spillovers across sectors.

[39]The analysis of the validity of our instrument falls within the shift-share instrumental variable framework and it must rely on some assumptions about the exogeneity of the shift terms, exposure shares, or both; see Borusyak et al. (2018) and Goldsmith-Pinkham et al. (2020) for a technical discussion of those assumptions.

Figure 2.6: Unconditional Correlation between Actual and Predicted Patents




of output per worker in the analysis period (2000-2014). We test this assumption in two ways: i) by regressing productivity in 1990 against average patent activity in the period of 2000-14 predicted by the instrument and ii) by checking that results are unchanged when controlling for an average level of patent activity in the period 1970-90.[40]

Second, we rule out the possibility that the links of knowledge diffusion used to construct the instrument capture demand pull factors from the destination country and sector, rather than a supply push from the origin country and sector. We do so by directly controlling by a shift-share variable constructed analogously to our instrument but with the timing reversed, so that it predicts the number of patents that should have been produced in the past in other countries and sectors to generate the current level of patenting in other country-sector pairs. More precisely, we start by constructing the pre-determined network of citations, this time

[40] We use value added per employment obtained from United Nations Industrial Development Organization (UNIDO) data as a measure of productivity.

using forward citations instead. Then, using the patenting activity across country-sector pairs during our sample period (2000-2014) and the forward citation network generated in the previous step, we infer the number of patents in the period 1970-1990 that would have been necessary to rationalize the 2000-2014 period. Finally, we include this predicted number of patents in our baseline regression as an additional control. These are patents that should have been filed in the period of 1970-1990 to generate patent activity in the period 2000-2014 that we observe in the data according to our idea generation empirical model.

### 2.5.2 Innovation and Productivity

In this section, we explore the effect of innovation on productivity. As we have just discussed, our identification strategy relies on pre-determined network knowledge linkages. They allow us to predict country- and sector-specific shocks to innovation activity (measured by patent filings) due to knowledge created in other geographical areas and sectors.

Table 2.1 shows our benchmark estimates of the relationship between value added per employment and innovation instrumented with predicted innovation. As discussed above, we use a three-year average of output per worker to remove short-term business cycle fluctuations.[41] Our benchmark regression uses data from the years 1970-90 to compute pre-determined network linkages, and the period of our analysis is 2000-2014. The first two columns report the estimated results when we only include lagged value added as a control, as well as country-year and sector-year fixed effects. In the third and fourth columns, we add to our empirical model lagged capital and employment as controls, to account for differences in inputs across countries and sectors. We find similar results to the regressions in columns (1) and (2). We use the specification in columns (3) and (4) as our a baseline.[42]

[41] Our baseline specification $\log(1+pat)$ allows us to retain the observations with zero patenting. The results are robust to using the inverse hyperbolic sine transformation of the number of patents instead of $\log(1+pat)$. Results for alternative log transformation of patents and forward lags for the dependent variable are reported in Table B.6 in the Appendix.

[42] Results with both lagged and contemporaneous capital and employment as controls are very similar.

Finally, in the fifth and sixth columns, we exploit the trade linkages given by the world input-output structure and add as controls the value of intermediates imported by each country-sector pair to explore the possibility that foreign imports of intermediates may disproportionately contribute to value added per worker, perhaps because of diffusion of ideas or intangible knowledge (Ayerst et al., 2020). We find no support for this hypothesis: the estimated coefficient on patenting hardly changes relative to our baseline.

The coefficient on innovation activity is positive, and statistically significant across the board. The magnitude of the two-stage least squares regressions is also stable across specifications. The coefficient in column (4) suggests that a 1% increase in patenting leads to 0.017% increase in value added per employment. Using the structure of our simple framework, we can rewrite the estimating equation by subtracting the current level of log value added per worker to also conclude that the estimated elasticity implies that a 1% increase in patenting leads to a 0.017% increase in the growth of value added per worker. This estimated elasticity implies that a one residual standard deviation increase in log patenting generates an increase in value added per employment growth of 1.1 percentage points. This change in valued added growth represents 7.8% of the standard deviation in output per worker growth in our sample.[43] To have a sense of magnitudes, one standard deviation increase corresponds to an increase in innovation activity in the pharmaceutical sector from the level of innovation observed in Canada to the level observed in the US in 2000. This also approximately corresponds to an increase in innovation activity in computer and electronic products from the level of innovation observed in Australia or France to the level observed in the US in

They are reported in Table B.7 in the Appendix. The fact that the inclusion of these controls does not change the estimated coefficient on patenting is consistent with our conceptual framework – which suggests that, with competitive factor markets, capital labor ratios across sectors are equalized and thus absorbed by the country-time fixed effects.

[43] Note that these results are calculated using residual standard deviations, that is, standard deviations obtained after partialling out the full set of controls in column (4). Without doing that, we would obtain larger effects. In fact, a one standard deviation increase in log patents implies an increase in log value added per employment (or value added per employment growth) of 4.4 percentage points.

Table 2.1: 2SLS Estimates: 2000-2014

| | $\overline{\log(va\_em_{cst+n})}$ $n \in \{1,2,3\}$ | | | | | |
|---|---|---|---|---|---|---|
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $\log(1+pat_{cst})$ | 0.006 | 0.019 | 0.004 | 0.017 | 0.005 | 0.019 |
| | (0.003) | (0.008) | (0.003) | (0.008) | (0.003) | (0.007) |
| $\log(va\_em_{cst})$ | 0.919 | 0.917 | 0.942 | 0.937 | 0.934 | 0.928 |
| | (0.012) | (0.012) | (0.016) | (0.016) | (0.016) | (0.015) |
| $\log(capital_{cst})$ | | | -0.016 | -0.014 | -0.015 | -0.014 |
| | | | (0.008) | (0.008) | (0.008) | (0.008) |
| $\log(employ_{cst})$ | | | 0.020 | 0.015 | 0.012 | 0.006 |
| | | | (0.010) | (0.010) | (0.011) | (0.010) |
| $\log(int\_import_{cst})$ | | | | | 0.009 | 0.010 |
| | | | | | (0.009) | (0.009) |
| Country-Year FE | Y | Y | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y | Y | Y |
| # obs. | 8,357 | 8,357 | 8,357 | 8,357 | 8,357 | 8,357 |
| # countries | 36 | 36 | 36 | 36 | 36 | 36 |
| | First-stage estimates | | | | | |
| Predicted | | 0.496 | | 0.461 | | 0.461 |
| $\log(1+pat_{cst})$ | | (0.082) | | (0.079) | | (0.079) |
| F-stat | | 36.7 | | 33.9 | | 34.3 |

*Notes:* Period of the analysis is 2000-14 using pre-determined matrix based on the data from 1970-90. First-stage estimates include all the controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Columns (1), (3), and (5) report the results using OLS, and Columns (2), (4), and (6) report the results obtained with 2SLS. Kleibergen-Paap Wald F-stat is reported for the first stage.

2000.

If we used interquartile range changes to quantify our results instead of standard deviation changes, we would obtain similar results. A one interquartile range increase in the log of the number of patents implies an increase of 10.4% of the interquartile range in value added per employment growth. Looking at countries at the bottom quartile of the patenting distribution in our sample, our estimated elasticity implies that, ceteris paribus, if Mexico in 2000 innovated in computer and electronic products and pharmaceuticals at the level of the US, output per worker in these sectors would have been higher by 3.1% and 2.9%, respectively.

The estimated 2SLS coefficients are larger than the ones obtained with the OLS regression. This increase is consistent with the likely scenario in which our OLS estimates suffer from attenuation bias because patents are an imperfect measure of innovation activity. Another possible explanation for the downward bias could be an increase in market concentration—a trend observed in most advanced countries since the 2000s. In particular, Akcigit and Ates (2021) and Olmstead-Rumsey (2019) have argued that higher market concentration leads to a slowdown in aggregate productivity growth while stimulating the innovation activity of market leaders to maintain their technological advantage.

**First-Stage Estimates and Knowledge Spillovers** Before turning to the robustness checks, we discuss the first-stage results reported in Table 2.1. We find positive and significant coefficients across the board of predicted patents constructed using our shift-share design on actual patenting. These estimates inform us directly about the average knowledge spillovers from other country-sector pairs on a given country-sector pair. The estimated coefficient implies an elasticity of 0.46 between the predicted patents from our shift-share design and the actual patenting activity. In terms of magnitude, a one standard deviation increase in predicted patents outside country-sector $(c, s)$ implies a 0.46 increase in actual patenting in country sector $(c, s)$ in a sample period.[44,45]

**Alternative Growth Specification and TFP Regressions** To assess the robustness of our findings, we extend our analysis to using TFP growth instead of output per worker as our dependent variable.[46] Table 2.2 shows our estimates for two measures of TFP growth, as well as value added per employment growth (rather than in levels, as in our baseline specification). The coefficient on innovation activity is positive, statistically significant across different measures, and quantitatively consistent with our baseline results.[47] Moreover, when comparing the coefficient on patenting, $\phi_N$, across different specifications, e.g., columns (3), (6), and (9), we see that, as implied by our simple framework, its magnitude is similar regardless of whether we use value added or TFP as the dependent variable.[48]

**Robustness Checks**

As discussed above, the validity of our shift-share design rests on country-sector pairs that give more citations pre-1990 not being on different trajectories in the terms of output per worker post-2000. This assumption is violated if the characteristics of countries and sectors that give more citations to particular countries and sectors in the period 1970-90 had persistent effects on patenting activity, as well as on changes in the outcomes of interest,

[44]We residualize all variables with all regression controls before computing the standard deviations. An analogous exercise without partialling out the controls would imply a 0.43 standard deviation increase.

[45]It is also possible to further investigate knowledge spillovers across countries and sectors by relaxing the restriction we impose in our baseline exercise by also including spillovers from the same countries and sectors. Of course, this is at the expense of endogeneity concerns. However, since we include country-time and sector-time fixed effects, a large array of potential concerns is taken care of by these. We find that if we include the own sector or own country or both, we obtain significant estimates implying a similar quantitative effect.

[46]We obtain measures of TFP growth at a country-sector level at a given period of time using "dual" and "primal" approaches as in Hsieh (1999) and Hsieh (2002).

[47]As in our baseline specification, the results reported in Table 2.2 are robust to using the inverse hyperbolic sine transformation of the number of patents instead of $\log(1 + pats)$ and adding forward lags as controls. See Tables B.9 and B.8 in the Appendix.

[48]We also find similar results to our baseline $\phi_N$ when estimating Equation (2.5) assuming $\phi_A = 1$ (and, thus, having the growth rate as a dependent variable). These are reported in Table B.5 in the Appendix.

Table 2.2: 2SLS Estimates: 2000-2014 TFP and VA/EMP growth

| | $\overline{\Delta \log(y_{cst+n})}$ $n \in \{1,2,3\}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | VA/EMP | | | Primal TFP | | | Dual TFP | | |
| | 2SLS (1) | 2SLS (2) | 2SLS (3) | 2SLS (4) | 2SLS (5) | 2SLS (6) | 2SLS (7) | 2SLS (8) | 2SLS (9) |
| $\log(1+patent_{cst})$ | 0.011 | 0.009 | 0.009 | 0.007 | 0.010 | 0.010 | 0.004 | 0.008 | 0.008 |
| | (0.004) | (0.004) | (0.003) | (0.004) | (0.006) | (0.005) | (0.004) | (0.003) | (0.003) |
| $\log(y_{cst})$ | -0.044 | -0.031 | -0.033 | -0.017 | -0.010 | -0.011 | -0.018 | -0.009 | -0.009 |
| | (0.009) | (0.007) | (0.007) | (0.010) | (0.009) | (0.008) | (0.009) | (0.005) | (0.005) |
| $\log(capital_{cst})$ | | -0.005 | -0.005 | | -0.023 | -0.022 | | -0.031 | -0.031 |
| | | (0.003) | (0.003) | | (0.003) | (0.003) | | (0.003) | (0.004) |
| $\log(employ_{cst})$ | | 0.005 | 0.002 | | 0.021 | 0.022 | | 0.026 | 0.023 |
| | | (0.004) | (0.005) | | (0.004) | (0.004) | | (0.004) | (0.003) |
| $\log(int\_import_{cst})$ | | | 0.003 | | | -0.002 | | | 0.001 |
| | | | (0.004) | | | (0.005) | | | (0.003) |
| Country-Year FE | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| # obs. | 8,834 | 8,357 | 8,357 | 7,931 | 7,931 | 7,931 | 8,554 | 8,336 | 8,336 |
| # countries | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 |
| | First-stage estimates | | | | | | | | |
| Predicted | 0.468 | 0.461 | 0.461 | 0.498 | 0.470 | 0.472 | 0.498 | 0.472 | 0.473 |
| $\log(1+pat_t)$ | (0.085) | (0.079) | (0.079) | (0.081) | (0.080) | (0.080) | (0.085) | (0.083) | (0.083) |
| F-stat | 30.5 | 33.9 | 34.3 | 34.5 | 32.5 | 32.4 | 38.1 | 35.0 | 34.9 |

*Notes:* Period of the analysis is 2000-14 using the pre-determined matrix based on the data from 1970-90. First-stage estimates include all controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. $y_{cst}$ is a respective measure of productivity (in columns (1)-(3) $y_{cst}$ is value added per employment, and in columns (4)-(9) $y_{cst}$ stands for TFP measured using either the primal or dual approach). In the case of primal TFP for our baseline specification (Column (5)), the main coefficient of interest is significant at the 10% level with p=0.09. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table 2.3: Checking for Pre-trends

| | $\log(\overline{va\_emp_{cs}})$ | | | |
|---|---|---|---|---|
| | Sample Period | | Pre-Sample Period | |
| | (1) | (2) | (3) | (4) |
| $\log(\overline{1+pat_{cs2000-14}})$ | 0.080 | 0.102 | 0.032 | 0.014 |
| | (0.033) | (0.046) | (0.064) | (0.053) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Country FE | Y | Y | Y | Y |
| Sector FE | Y | Y | Y | Y |
| # obs. | 641 | 433 | 433 | 424 |

*Notes:* Columns (1) and (2) use average value added per employment in the period 2000-14 as a dependent variable computed with WIOD and UNIDO data, respectively. The latter one is included for better compatibility with results in Columns (3) and (4), where the dependent variable is the average value added per employment computed with UNIDO data for the periods 1981-90 and 1971-90, respectively. All regressions include average (log) values for capital, employment, and intermediate imports in the period 2000-14. Standard errors (in parentheses) are two-way clustered at a country and sector levels.

and these are not captured by our controls. We test this assumption in a variety of ways. First, we test for pre-trends by showing that the pre-period productivity is uncorrelated with subsequent patent activity predicted by the instrument. Table 2.3 presents the results of regressing the average value of productivity during the pre-sample period against the average annual number of patents in the period 2000-14.[49] The coefficients of this regression, reported in Columns (3) and (4), are not statistically significant—and also of a different magnitude compared with the estimates obtained for the period used in the main exercises (which are indeed significant), reported in Columns (1) and (2).

Second, in Column (2) of Table B.10, we check that our results hold when controlling for the average level of patenting activity in the period 1970-90. The results remain virtually unchanged. The coefficient of interest becomes larger in magnitude (in absolute value), but

[49]As a measure of productivity we use value added per employment data from UNIDO database, since data for historical periods is not available in WIOD. We also averaged all the variables in order to suppress the time dimension as the left-hand side and right-hand side of our regression belong to different time periods.

it is statistically indistinguishable from the baseline level because the standard error also increases.

Next, we present evidence consistent with ruling out the possibility that the links of knowledge diffusion used to construct the instrument capture a demand pull factors from the destination country and sector, rather than a supply push from the origin. To do that, we include in our baseline regression as a control the number of patents that should have been filed in the pre-sample period to explain the actual number of patents observed in the sample in the period of study given the citations linkages in the pre-sample.[50] The results presented in Column (3) of Table B.10 are stable. The coefficient of interest remains statistically significant and quantitatively close to the baseline. Column (4) includes both controls simultaneously, i.e., the historical patent activity and the demand-driven number of patents in the past in the baseline regression. The coefficient remains significant and has a similar magnitude.

Finally, to check for whether some outliers are driving our results, we repeat our baseline regression excluding one country or sector at a time. We find that our results remain stable and are essentially unchanged across all these regressions.[51]

### 2.5.3 Innovation and Long-term Development

Our baseline analysis studied value added per worker after the year 2000. This section extends our analysis to a longer time frame. One challenge of looking at long-term outcomes is that high-quality value added per employment or TFP panel data spanning a large number of countries and sectors are not readily available. To circumvent this problem, we adapt

[50] We construct this variable by using the "reverse" matrix procedure described in the end of Section 2.5.1. To deal with the time dimension of data, we include in the regression the predicted number of patents that should have been filed 30 years in the past. The results hold for other choices of lags.

[51] The largest change in magnitude that we obtain in $\phi_N$ is when we exclude the sector called Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials. In this case, it increases from 0.017 to 0.023. These results are available upon request.

Table 2.4: 2SLS Estimates: Robustness

| | $\overline{\log(va\_em_{cst+n})}$ $n \in \{1,2,3\}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $\log(1 + pat_{cst})$ | 0.017 | 0.029 | 0.025 | 0.030 |
| | (0.008) | (0.010) | (0.010) | (0.010) |
| $\log(1 + \overline{pat_{cs1970-90}})$ | | -0.009 | | -0.009 |
| | | (0.005) | | (0.005) |
| $\log(1 + \widehat{pat}_{cst-30})$ | | | -0.006 | -0.001 |
| | | | (0.007) | (0.006) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Country-Year FE | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y |
| # obs. | 8,357 | 8,357 | 8,357 | 8,357 |
| | First-stage estimates | | | |
| Predicted | 0.461 | 0.264 | 0.388 | 0.305 |
| $\log(1 + pat_{cst})$ | (0.079) | (0.058) | (0.065) | (0.056) |
| F-stat | 33.9 | 20.9 | 35.5 | 29.3 |

*Notes:* Column (1) shows the results of our baseline regression. Column (2) and (3) show the regression results when including separately the historical levels of average patent activity and the predicted number of patents driven by demand pull factors, respectively; and Column (4) shows the regression results when including them together. All regressions include (log) values for value added per employment, capital, and employment as controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

our empirical strategy to study the relationship between innovation activity and GDP per capita at the aggregate country level since 1980 (and later extend it back to 1960), using real GDP per capita data from the Maddison Project Database (Inklaar et al., 2018). We therefore depart from our baseline exercise along two dimensions. First, we abstract from sectoral variation both when we construct our instrument and when we conduct the regression analysis. Second, we use GDP per capita rather than output per worker as our outcome variable.

The choice of the time period for our analysis is the result of a balancing act. On the one hand, since we are interested in long-run growth, we would like to study a long time period. On the other hand, given that comprehensive patent data for the period prior to 1970 mostly covers advanced economies and given that for most developing countries we observe little to no innovation activity measured in terms of patents prior to 1970, our shift-share design may miss a part of the variation we are interested in capturing. For these reasons, we choose the years 1980-2016 as our baseline time period, while we use the pre-1980 data to construct our instrument (so that we include the 1970s, for which there are data on a substantial number of patents for middle-income economies). The set of countries that we consider are the ones categorized as high-income and upper-middle-income countries according to the World Bank classification, for which we have substantial variation in patenting activity.[52]

To obtain our shift-share instrument in this cross-country setup, we use only country-time variation in citations to generate the pre-determined matrix of linkages. Each element

[52] Our patent data cannot distinguish between zero patenting activity and missing data in a given country, sector, and year. Throughout our analysis, we assume that no records of patenting activity are treated as zero patents. Under this assumption, the average annual number of patents in the period 1960-80 is 21,264 and 1,227 patents for high-income and upper-middle-income countries, respectively. At the same time, the average number of annual patents for the same period for lower-middle-income and low-income countries is 45 patents and 1 patent, respectively. Given the little variation in patenting activity for the historical time period in less developed economies, we focus on high-income and upper-middle-income countries for our long-term analysis. We report a number of robustness checks at the end of the section.

of the matrix is computed as

$$m_{c_o,c_d,\Delta} = \frac{\sum_{t=T_0^{share}}^{T_1^{share}} \sum_{p \in \mathcal{P}(c_o,t)} s_{p \to (c_d, t-\Delta)}}{\sum_{t=T_0^{share}}^{T_1^{share}} |\mathcal{P}\left(c_d, t-\Delta\right)|},$$

and we thus abstract from sectoral variation.[53] We use the citation data observed in the period prior 1980 to construct the pre-existing linkages across countries, along with countries' patenting activity during the period starting in 1970, as shifts to construct our instrument for patenting activity during the period 1980-2016.[54]

The empirical specification we run corresponds to Equation (2.4) in our motivating framework (without sectoral variation). As a reminder, it is obtained from combining a Cobb-Douglas aggregate production function and our law of motion for TFP. The following specification is used in the analysis:

$$\overline{\log(gdp\_cap_{ct+n})} = \phi_N \log(1 + pat_{ct}) + \phi_A \log(gdp\_cap_{ct}) + \delta_t + \delta_c + \varepsilon_{ct}, \tag{2.7}$$

where on the left-hand side we use the average level of GDP per capita over $n = 3$ years after $t$ to smooth out variation driven by business cycles and other idiosyncratic shocks.

[53]As a robustness check, we also compute our shift-share instrument using cross country and sector variation and then aggregate up the sectoral variation. That is, we compute the linkages at the country-sector level as in our baseline regression and then create our shift-share instrument at the country-sector level first. Then, we aggregate the predicted number of patents across sectors within a country (and year) to construct the instrument. We find very similar results with this alternative procedure.

[54]Similar to our baseline instrument, we use a mix of actual and predicted patents as shifts. We also do not take into account domestic spillovers when constructing the instrument, i.e., $m_{c_o,c_d} = 0$, when $o = d$ and consider citation lags $\Delta \in \{1, \cdots, 10\}$. However, we no longer have the intermediate 10-year period between the pre-determined matrix and instrument as in our baseline. This is to ensure both that we have a sufficiently long sample size of GDP growth rates and that we include patenting activity of the 1970s to construct our shift-share. We also performed as a robustness check analysis where we use all citations available before 1960/70 to construct the pre-determined matrix of citation linkages, along with 1970/80-2016 as a period for the regression analysis, obtaining similar results.

Table 2.5 shows our results. As in the previous section, the 2SLS estimates reported in columns (2) and (4) imply a higher elasticity of patenting on income compared with the OLS estimates in columns (1) and (3). In our preferred specification, which includes country and year fixed effects, we find a positive, significant coefficient that is similar in magnitude to the elasticity of patents to sectoral output per worker that we find for the period 2000-2014. The elasticity of patenting to income per capita is 0.034.[55] Quantitatively, this elasticity implies that one standard deviation increase in the logarithm of the annual number of patents leads to 0.41 standard deviations increase in the logarithm of annual GDP per capita, implying an increase of 2.8 percentage points in the growth of GDP per capita.

**Income per capita growth over longer horizons.** We extend the period of analysis to longer time horizons. Columns (1)-(4) in Table B.11 in the Appendix report the results of running the same specification, Equation (2.7), using income per capita data spanning the periods 1960-2016 and 1970-2016. In each case, we construct our shift-share instrument in an analogous way to what we have done so far in this section, but now with patenting data pre-1960 or pre-1970, respectively. In both cases, we find a positive and significant first stage, despite our network of innovation being more sparse. We estimate a positive and significant effect of innovation on income per capita growth in both regressions. The implied magnitudes suggest that that one standard deviation increase in the logarithm of the annual number of patents generates an increase of 1.64 and 2.15 percentage points in GDP per capita growth for the periods 1960-2016 and 1970-2016, respectively.

[55] If we run our regression for all countries in our sample rather than only middle and upper income countries, we find an almost identical coefficient of 0.31. However, the first stage is weak and the estimated coefficient is not significant at conventional levels. See columns (5) and (6) of Table B.11 in the Appendix.

Table 2.5: 2SLS Estimates: Innovation and Long-term Development: 1980-2016

| | Dependent Variable is: $\overline{\log(gdp\_cap_{ct+n})}$ | | | |
|---|---|---|---|---|
| | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) |
| $\log(1 + pat_{ct})$ | 0.013 | 0.086 | 0.005 | 0.034 |
| | (0.004) | (0.021) | (0.003) | (0.012) |
| $\log(gdp\_cap_{ct})$ | 0.906 | 0.735 | 0.852 | 0.804 |
| | (0.026) | (0.052) | (0.025) | (0.028) |
| Country FE | Y | Y | Y | Y |
| Year FE | N | N | Y | Y |
| # obs. | 1,985 | 1,985 | 1,985 | 1,985 |
| # countries | 60 | 60 | 60 | 60 |
| | First-stage estimates | | | |
| Predicted | | 0.771 | | 1.884 |
| $\log(1 + pat_{ct})$ | | (0.199) | | (0.695) |
| F-stat | | 15.0 | | 7.3 |

*Notes:* Period of the analysis is 1980-2016 using pre-determined matrix based on the data for the pre-1980 period. Standard errors (in parentheses) are clustered at the country level. Columns (1) and (3) present the results for OLS, and Columns (2) and (4) presents the results obtained with 2SLS. In regressions (1) and (2) only country fixed effects are used. To account for a trend in the number of patents, the regressions in columns (3) and (4) also include year fixed effects. Kleibergen-Paap Wald F-stat is reported for the first stage.

## 2.6 Concluding Remarks

In this paper, we use a panel of historical patent data spanning the past hundred years and a large range of countries to study the evolution of innovation across time and space and its effect on productivity. In the first part of the paper, we propose a clustering algorithm to classify finely defined patent classes into fields of knowledge based on inventors' patent activity. We then turn to documenting some salient facts of patenting activity since the beginning of the 20th century. We document broad technological waves over the 20th century and in the early decades of the 21st century, and the heterogeneous contribution of countries to these waves. We also document a substantial rise of international knowledge spillovers, as measured by patent citations since the 1990s. This rise is mostly accounted by an increase in citations to US and Japanese patents in fields of knowledge related to computation, information processing, and medicine.

After documenting these facts, we propose a shift-share approach that leverages the directed network of knowledge spillovers across fields of knowledge and countries (to construct the shift terms) and the heterogeneity in exposure of countries to technological waves (to construct the share terms). We then utilize our proposed instrument to estimate the effect of innovation on output per worker and TFP growth in a panel of country-sectors over the period 2000-2014, with our instrument using historical patent data spanning the years 1970 through 2000. We find that, on average, an increase of one standard deviation in patenting implies a 1.1 percentage point increase of output per worker growth.

Finally, we estimate the effect of innovation on long-run income per capita growth and find a positive effect, similar in magnitude to our baseline results. We believe that our shift-share design can be applied to other settings in which the effect of innovation or productivity are of interest. For example, our empirical strategy can be employed in a multi-sectoral Ricardian trade model as in Costinot et al. (2012) to estimate the elasticity of trade flows to productivity differences.

# Chapter 3

# Gendered Impacts of Covid-19 in Developing Countries[1]

## 3.1 Introduction

The Covid-19 pandemic and the associated shutdowns, social distancing measures, and school closures have resulted in a global recession that sharply reduced output and employment in nearly all countries. In many high-income economies, one of the most unusual characteristics of this recession has been a disproportionate impact on women in the labor market (Alon et al., 2022). In the United States, for example, the unemployment rate increased by three percentage points more for women compared to men. This marks a sharp deviation from the usual pattern of recent recessions in high-income economies, which have affected men's employment more than women's.

In this paper, we explore how the Covid-19 recession has affected women's versus men's employment in developing countries. While the impact of school closures is similar, we argue that differences in the distribution of job characteristics and in the role of income effects have limited the employment reductions experienced by women in low-income economies. As a case study, we show how these factors play out in Nigeria, the most populous country in Sub-Saharan Africa.

[1]We thank Olga Stoddard for helpful suggestions and the German Research Foundation (through the CRC-TR-224 project A3 and the Gottfried Wilhelm Leibniz-Prize) and the National Science Foundation for their financial support.

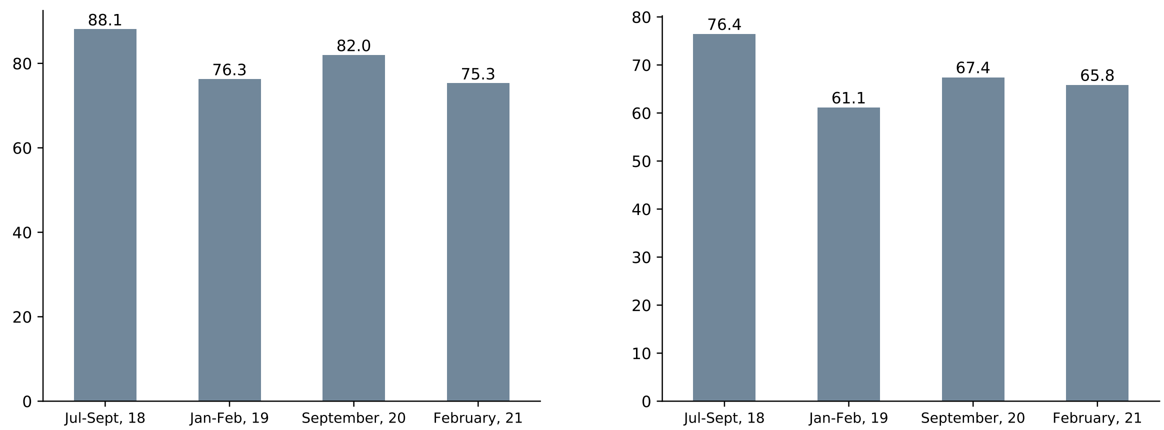
## 3.2 Origins of Gender Differences in the Pandemic

The literature on the gendered impact of the Covid-19 pandemic has pointed out two primary reasons why women in advanced economies experienced unusually large employment reductions. The first is the distribution of job characteristics of employed women and men. In the Covid-19 recession, employment losses were concentrated in contact-intensive occupations in the service industry, such as wait staff in restaurants and workers in hotels and entertainment. In many countries, these sectors and occupations have high female employment shares, which contributed to large job losses for women during the pandemic (Albanesi and Kim, 2021; Alon et al., 2022).

While developing countries also employed shutdowns and social distancing measures, contact-intensive service industries account for a small share of women's employment (see Figure C.1 in the Appendix). Especially in the poorest economies, many more women work in family-based agriculture and in non-farm household enterprises, where there are only small employment changes over the cycle. Hence, the distribution of job characteristics for women and men in the economy is one explanation for why the impact of the pandemic on women's employment was different in low-income countries.

The second reason underlying women's reduced labor supply in high-income economies was the impact of increased childcare needs during closures of schools and daycare centers. A number of studies document that during school closures parents, and in particular mothers, spent much more time on childcare and home schooling tasks (Adams-Prassl et al., 2020). Correspondingly, reductions in labor supply were particularly large among mothers of school-age children (Alon et al., 2022).

School closures during the pandemic were widely adopted in high- and low-income economies alike, and while the duration of school closures varies widely across countries, there is no clear correlation with income levels (Alon et al., 2022). Nevertheless, the effects of these clo-




(a) Men

(b) Women

Figure 3.1: Share of Working Adults in Nigeria by Gender

*Notes:* The share of adults of age 21-55 that worked in the past week (at time when interview was conducted). Sample includes $\approx 9,000$ and $\approx 4,000$ individuals for pre-Covid and Covid interviews, respectively.

sures on women's and men's labor supply may still depend on local conditions. The need for additional childcare is reduced if informal modes of childcare are available, for example, if an extended family is living together and grandparents can look after children during closures. The need for spending time on home schooling also depends on how much remote schooling actually takes place. If no remote schooling is available and families decide that kids will simply take a break from learning, parental time needs are lower. The evidence indeed suggests learning activities during school closures were reduced even more in low-income compared to high-income countries (see Figure C.2 in the Appendix), which is consistent with a lower impact of closures on parents' time needs.

Another factor determining the impact of school closures on labor supply is the extent to which spending time on childcare and home schooling interferes with work. Alon et al. (2022) show that among parents who can work from home (e.g., workers with office jobs who can connect remotely) there is no gender gap in the impact of increased childcare needs on labor supply. It is mothers with jobs that have to be done at a specific workplace (such as a

Table 3.1: Impact of Covid-19 on Employment and Hours of Work for Adults

| | Employment Status | | | | Weekly Working Hours | | | |
|---|---|---|---|---|---|---|---|---|
| | Sept. | Sept. | Febr. | Febr. | Sept. | Sept. | Febr. | Febr. |
| Covid | -0.045 | -0.025 | 0.036 | 0.004 | -2.766 | -4.136 | 4.859 | 3.197 |
| | (0.013) | (0.014) | (0.031) | (0.033) | (0.969) | (1.228) | (2.272) | (2.224) |
| Covid × Female | | -0.035 | | 0.058 | | 2.784 | | 3.242 |
| | | (0.018) | | (0.024) | | (1.264) | | (1.210) |
| # Obs | 12,229 | 12,229 | 12,444 | 12,444 | 9,634 | 9,634 | 8,519 | 8,519 |
| R-squared | 0.20 | 0.20 | 0.22 | 0.22 | 0.23 | 0.23 | 0.25 | 0.25 |
| Mean Pre-Covid | 0.817 | 0.817 | 0.680 | 0.680 | 34.3 | 34.3 | 31.6 | 31.6 |
| Age FE | Y | Y | Y | Y | Y | Y | Y | Y |
| LGA FE | Y | Y | Y | Y | Y | Y | Y | Y |
| Control Variables | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are clustered at the state level. Controls include gender, urban, number of HH members, access to electricity & internet, ownership of different assets (radio, car, land, etc.), access to finance, consumption quantile before the pandemic, education and literacy of the individual, marriage status, whether individual is a head of household, a geographic fixed effect (LGA), and a dummy for pre-covid interview held in January. In regressions for weekly working hours only working adults are included. Results for weekly working hours that combine both intensive and extensive margins are reported in Table C.1 in the Appendix.

manufacturing plant or a retail store) who reduce labor supply a lot when childcare needs go up. In low-income economies, a large share of employment is done in or around the home, such as family based agriculture and other forms of self-employment. This fact suggests, once again, that the impact of school closures on labor supply in general and on women's labor supply in particular may be smaller in low-income economies.

In what follows, we document how these factors shape the impact of the pandemic on women's employment in Nigeria.

## 3.3 The Employment Impact of the Covid-19 Pandemic in Nigeria

Nigeria was one of the first African countries that reported Covid-19 cases. As in many other countries, the government implemented strict measures to contain the spread of the virus, including travel restrictions and school closures. We use data from the Nigeria COVID-19 National Longitudinal Phone Survey (Covid-19 NLPS) to assess the impact of the pandemic on employment. We focus on data collected in September 2020, covering outcomes when school closures and other containment measures were still in effect, and in February 2021, when schools were open again. For these survey waves, we can compare outcomes to data collected around the same months two years prior in Nigeria's General Household Survey. Comparing outcomes for the same season is important given that employment in Nigeria varies over the planting and harvesting seasons.

In both September 2020 and February 2021, a variety of Covid mitigation measures were in place (see Figure C.3 in the Appendix for a timeline). Measures of people's mobility had mostly recovered by September 2020; restrictions and shutdown measures were the most stringent in April and May of 2020 and gradually relaxed afterwards. However, school closures were still ongoing in September 2020; most schools fully reopened only in November 2020 (see Figure C.4 in the Appendix). Hence, the comparison of outcomes for September 2020 and February 2021 is informative about the impact of school closures.

Figure 3.1 shows how overall employment of prime-age adults (ages 21 to 55) varies across the survey waves for women and men. Comparing the levels in July-September of 2018 and September of 2020, we observe a substantial drop in the share of employed adults. Women's employment drops by 9.0 percentage points, much larger than the drop of 6.1 percentage points for men. Hence, the initial impact mirrors the observation from high-income

Table 3.2: Role of Childcare for Impact of Covid-19 on Employment and Hours of Work

| | Employment Status | | Weekly Working Hours | |
|---|---|---|---|---|
| | Sept. | Febr. | Sept. | Febr. |
| Covid × Female × Young Kids | 0.028 | 0.070 | 0.374 | -3.007 |
| | (0.029) | (0.035) | (1.975) | (2.102) |
| Covid × Female × School-Age Kids | -0.058 | 0.031 | 2.768 | 6.701 |
| | (0.028) | (0.035) | (1.542) | (2.134) |
| Covid × Female × No kids | -0.035 | -0.025 | 1.137 | -0.878 |
| | (0.048) | (0.052) | (2.196) | (3.093) |
| # Obs | 12,229 | 12,444 | 9,634 | 8,519 |
| Mean Pre-Covid | 0.817 | 0.680 | 34.3 | 31.6 |
| Age FE | Y | Y | Y | Y |
| LGA FE | Y | Y | Y | Y |
| Control Variables | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are clustered at the state level. Controls include gender, urban, number of HH members, access to electricity & internet, asset ownership, access to finance, consumption quantile before the pandemic, education and literacy, marriage status, head of household status, a geographic fixed effect (LGA), and a dummy for pre-covid interview held in January. In regressions for weekly working hours only working adults are included.

economies that women's employment was disproportionately affected by the pandemic. However, this picture is reversed by February 2021: here we observe a substantial increase of women's employment by 4.7 percentage points compared to the pre-pandemic period, versus a moderate decline of one percentage point in men's employment. Similarly, in terms of weekly hours worked conditional on being employed, there is a sharp rise in women's labor supply in February 2021 compared to before the pandemic (see Figure C.5 in the Appendix).

Table 3.1 displays individual-level regression results of the impact of the pandemic on employment by gender that include individual and household controls and geographic fixed effects (LGA). Regressions for September pool data for September 2020 with the July–September survey in 2018, and regressions for February include data for February 2021 and January–February 2019. "Covid" is an indicator variable equal to 1 for September 2020 and February 2021, respectively, and zero for the pre-pandemic period.

The regressions confirm that women lost substantially more employment in the early phase of the pandemic, but also experienced an expansion of employment later in the recovery, both relative to men and in absolute terms. On the intensive margin, women who continued working worked more hours both in September and February compared to the pre-pandemic period.

## 3.4 The Role of Childcare

To examine the possible role of childcare needs during school closures for employment changes, we expand the regressions displayed in Table 3.1 by including indicator variables for the presence of children in the household. Following the empirical setting in Alon et al. (2022), we distinguish between households with at least one child under the age of five, households where the youngest child is of school age (here defined as 5 to 14, as compulsory education in Nigeria is completed at age 14), and households who either don't have children or only have older children. These indicator variables are interacted with the Covid indicator variable and gender. Table 3.2 displays the coefficient estimates for the double interaction of Covid with the female indicator variable and the child variables. For September 2020, the regressions confirm the finding of Alon et al. (2022) for high-income economies that employment declined the most among mothers of school-age children. Given that schools were still closed in September 2020 but not in February 2021, this finding strongly suggests that as in high-income countries, increased childcare needs during school closures were an important driver of women's employment declines during the pandemic.

Overall, the aspect of increased childcare needs for school-age children is the main parallel between the experience of women in high-income economies during the pandemic and women in Nigeria. However, even among parents of school-age children we do not observe a statistically significant gender gap in working hours during school closures conditional on continued employment. This may reflect that in low-income countries, a smaller share of children continued learning activities during school closures, which reduces the need for parental time. Moreover, unlike in high-income countries, we do not observe statistically significant gender differences in initial employment changes among those without children. This observation is consistent with the notion that in low-income countries, the industry composition of employment did not favor one gender over the other in the pandemic recession. Likewise, there are no statistically significant gender differences among those with young children, which may be due to lower initial use of formal childcare, the fact that a lot of work takes place at home, and the availability of informal childcare.

A final major difference between the employment outcomes of women in high- and low-income economies is that in many high-income economies, women's employment losses have been persistent; in the United States, for example, labor force participation remained well below pre-pandemic levels even after schools reopened and unemployment rates fell to historic lows. In contrast, in Nigeria we observe that women's employment not only recovered quickly, but actually rose above pre-pandemic levels once schools reopened.

For explaining the rise in women's employment in the later phase of the pandemic, based on Alon et al. (2021) we conjecture that income effects play a role. In the United States and other high-income countries, governments provided generous transfer payments during the crisis, making many households less dependent on the next paycheck. In low-income countries, households received few transfers and were much poorer to begin with. The need to make up for income losses during the economic downturn caused by the pandemic may have induced many women to work more or to take on additional jobs. Given that women's labor supply was initially lower than that of men, women had more room to expand labor supply to increase household income. The income channel is supported by the observation that the positive effect of the pandemic on women's labor supply in February 2021 is concentrated among poorer households (see Table C.3 in the Appendix for regression results that split the sample by consumption quantiles). This mechanism resembles the insurance role of women's labor supply analyzed by Alon et al. (2020), but here the main impact is during the recovery rather than at the height of the pandemic.

## 3.5 Concluding Remarks

Compared to high-income economies, the gender differences in the employment impact of the pandemic that we document for the case of Nigeria are muted. A channel that is potentially more important in developing countries is the impact of the pandemic on children's education. Early indications are that learning losses in developing countries are larger than in high-income economies, and that many older children dropped out of school and started working during the pandemic (see the Appendix for evidence on the impact of the pandemic on adolescents' labor supply). These changes can have long-run repercussions for children's future earnings as well as for outcomes such as marriage and childbearing. We examine the impact of the pandemic on children's education in low-income economies in Alon et al. (2022).

# Bibliography


Acemoglu, D. (2009a). *Introduction to modern economic growth*. Princeton, NJ [u.a.]: Princeton Univ. Press.

Acemoglu, D. (2009b). *Introduction to Modern Economic Growth*. Princeton University Press.

Acemoglu, D., U. Akcigit, and W. Kerr (2015). Networks and the macroeconomy: An empirical exploration. In *NBER Macroeconomics Annual 2015, Volume 30*, pp. 273–335. National Bureau of Economic Research, Inc.

Acemoglu, D., U. Akcigit, and W. R. Kerr (2016). Innovation network. *Proceedings of the National Academy of Sciences 113*(41), 11483–11488.

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

Adamopoulos, T., L. Brandt, J. Leight, and D. Restuccia (2017). Misallocation, selection and productivity: A quantitative analysis with panel data from china. Working papers, University of Toronto, Department of Economics.

Adamopoulos, T. and D. Restuccia (2014). The size distribution of farms and international productivity differences. *American Economic Review 104*(6), 1667–97.

Adamopoulos, T. and D. Restuccia (2020). Land reform and productivity: A quantitative analysis with micro data. *American Economic Journal: Macroeconomics 12*(3), 1–39.

Adams-Prassl, A., T. Boneva, M. Golin, and C. Rauh (2020). Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics 189*, Article 104245.

Aghion, P., A. Dechezlepretre, D. Hemous, R. Martin, and J. van Reenen (2016). Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. *Journal of Political Economy 124*(1), 1 – 51.

Aghion, P. and P. Howitt (1992). A model of growth through creative destruction. *Econometrica 60*(2), 323–51.

Aghion, P. and P. Howitt (1998). *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Agyei-Holmes, A., N. Buehren, M. P. Goldstein, R. D. Osei, I. Osei-Akoto, and C. Udry (2020). The effects of land title registration on tenure security, investment and the allocation of productive resources: Evidence from ghana. Policy Research Working Paper Series 9376, The World Bank.

Akcigit, U. and S. T. Ates (2021, January). Ten facts on declining business dynamism and lessons from endogenous growth theory. *American Economic Journal: Macroeconomics 13*(1), 257–98.

Akcigit, U., S. Baslandze, and F. Lotti (2018, October). Connecting to power: Political connections, innovation, and firm dynamics. Working Paper 25136, National Bureau of Economic Research.

Akcigit, U., J. Grigsby, and T. Nicholas (2017, January). The Rise of American Ingenuity: Innovation and Inventors of the Golden Age. CEPR Discussion Papers 11755, C.E.P.R. Discussion Papers.

Albanesi, S. and J. Kim (2021). Effects of the covid-19 recession on the us labor market. *Journal of Economic Perspectives 35*(3), 3–24.

Alon, T., S. Coskun, M. Doepke, D. Koll, and M. Tertilt (2022). From mancession to shecession: Women's employment in regular and pandemic recessions. Forthcoming, NBER Macroeconomics Annual.

Alon, T., M. Doepke, K. Manysheva, and M. Tertilt (2022). What is the aggregate impact of pandemic school closures in low-income countries? Unpublished Manuscript, Northwestern University.

Alon, T., M. Doepke, J. Olmstead-Rumsey, and M. Tertilt (2020). This time it's different: The role of women's employment in a pandemic recession. NBER Working Paper 27660.

Alon, T., M. Kim, D. Lagakos, and M. VanVuren (2021). Macroeconomic effects of covid-19 across the world income distribution. Unpublished Manuscript, UCSD.

Altuntas, S., T. Dereli, and A. Kusiak (2015). Analysis of patent documents with weighted association rules. *Technological Forecasting and Social Change 92*, 249–262.

Arcand, J.-L., E. Berkes, and U. Panizza (2015). Too much finance? *Journal of Economic Growth 20*(2), 105–148.

Asker, J., A. Collard-Wexler, and J. De Loecker (2011). Productivity volatility and the misallocation of resources in developing economies. CEPR Discussion Papers 8469, C.E.P.R. Discussion Papers.

Ayerst, S., F. Ibrahim, G. MacKenzie, and S. Rachapalli (2020). Trade and Diffusion of Embodied Technology: An Empirical Analysis.

Banerjee, A. and B. Moll (2010). Why does misallocation persist? *American Economic Journal: Macroeconomics 2*(1), 189–206.

Banerjee, A. V. and E. Duflo (2005). Growth theory through the lens of development economics. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth* (1 ed.), Volume 1, Part A, Chapter 07, pp. 473–552. Elsevier.

Baqaee, D. and E. Farhi (2019). The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem. *Econometrica 87*(4), 1155–1203.

Baqaee, D. and E. Farhi (2020). Productivity and misallocation in general equilibrium*. *The Quarterly Journal of Economics 135*(1), 105–163.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2013). Cross-country differences in productivity: The role of allocation and selection. *American Economic Review 103*(1), 305–34.

Bau, N. and A. Matray (2020). Misallocation and capital market integration: Evidence from india. NBER Working Papers 27955, National Bureau of Economic Research, Inc.

Beg, S. (2021). Digitization and development: Formalizing property rights and its impact on land and labor allocation. Mpra paper, University Library of Munich, Germany.

Berkes, E. and R. Gaetani (2022). Income Segregation and the Rise of the Knowledge Economy. *American Economic Journal: Applied*. forthcoming.

Besley, T. and M. Ghatak (2010). Property rights and economic development. Volume 5, Chapter Chapter 68, pp. 4525–4595. Elsevier.

Bick, A., F.-S. Nicola, and D. Lagakos (2016). How do average hours worked vary with development? cross-country evidence and implications. CEPR Discussion Papers 11092, C.E.P.R. Discussion Papers.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020, April). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–44.

Borusyak, K., P. Hull, and X. Jaravel (2018). Quasi-experimental shift-share research designs. NBER Working Papers 24997, National Bureau of Economic Research, Inc.

Bottazzi, L. and G. Peri (2003, August). Innovation and spillovers in regions: Evidence from European patent data. *European Economic Review 47*(4), 687–710.

Breschi, S., F. Lissoni, and F. Malerba (2003). Knowledge-relatedness in firm technological diversification. *Research Policy 32*(1), 69–87.

Bromley, D. (2010). The emergence of land markets in africa: Impacts on poverty, equity, and efficiency, edited by stein t. holden, keijiro otsuka, and frank m. place. *Journal of Regional Science 50*(5), 1006–1008.

Buera, F., J. Kaboski, and Y. Shin (2014). Macro-perspective on asset grants programs: Occupational and wealth mobility. *American Economic Review 104*(5), 159–64.

Buera, F., J. Kaboski, and Y. Shin (2021). The macroeconomics of microfinance. *Review of Economic Studies 88*(1), 126–161.

Buera, F., J. Kaboski, and R. Townsend (2021). From micro to macro development. NBER Working Papers 28423, National Bureau of Economic Research, Inc.

Buera, F. and E. Oberfield (2020). The global diffusion of ideas. *Econometrica 88*(1), 83–114.

Buera, F. J., J. Kaboski, and Y. Shin (2015). Entrepreneurship and financial frictions: A macrodevelopment perspective. *Annual Review of Economics 7*(1), 409–436.

Cai, J. and N. Li (2019). Growth through inter-sectoral knowledge linkages. *Review of Economic Studies 86*(5), 1827–1866.

Chari, A., E. Liu, S.-Y. Wang, and Y. Wang (2017). Property rights, land misallocation and agricultural efficiency in china. NBER Working Papers 24099, National Bureau of Economic Research, Inc.

Chen, C. (2017). Untitled land, occupational choice, and agricultural productivity. *American Economic Journal: Macroeconomics 9*(4), 91–121.

Clark, G. and R. C. Feenstra (2003). Technology in the Great Divergence. In *Globalization in Historical Perspective*, NBER Chapters, pp. 277–322. National Bureau of Economic Research, Inc.

Coelli, F., A. Moxnes, and K. H. Ulltveit-Moe (2016, September). Better, Faster, Stronger: Global Innovation and Trade Liberalization. NBER Working Papers 22647, National Bureau of Economic Research, Inc.

Comin, D. and M. Gertler (2006, June). Medium-term business cycles. *American Economic Review 96*(3), 523–551.

Comin, D., D. Lashkari, and M. Mestieri (2019). Structural Transformation of Innovation. 2019 Meeting Papers 1394, Society for Economic Dynamics.

Comin, D. and M. Mestieri (2014). Technology Diffusion: Measurement, Causes, and Consequences. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth*, Volume 2 of *Handbook of Economic Growth*, Chapter 2, pp. 565–622. Elsevier.

Comin, D. and M. Mestieri (2018, July). If Technology Has Arrived Everywhere, Why Has Income Diverged? *American Economic Journal: Macroeconomics 10*(3), 137–178.

Costinot, A., D. Donaldson, and I. Komunjer (2012). What goods do countries trade? a quantitative exploration of ricardo's ideas. *Review of Economic Studies 79*(2), 581–608.

de Brauw, A. and V. Mueller (2012). Do limitations in land rights transferability influence mobility rates in ethiopia? *Journal of African Economies 21*(4), 548–579.

de Janvry, A., K. Emerick, M. Gonzalez-Navarro, and E. Sadoulet (2015). Delinking land rights from land use: Certification and migration in mexico. *American Economic Review 105*(10), 3125–49.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica 79*(5), 1407–1451.

Dechezleprêtre, A., D. Hémous, M. Olsen, and C. Zanella (2020, February). Automating Labor: Evidence from Firm-Level Patent Data. CEP Discussion Papers dp1679.pdf, Centre for Economic Performance, LSE.

Deininger, K. and J. S. Chamorro (2004). Investment and equity effects of land regularisation: the case of nicaragua. *Agricultural Economics 30*(2), 101–116.

Di Tella, R., S. Galiant, and E. Schargrodsky (2007). The formation of beliefs: Evidence from the allocation of land titles to squatters. *The Quarterly Journal of Economics 122*(1), 209–241.

Eaton, J. and S. Kortum (1999). International technology diffusion: Theory and measurement. *International Economic Review 40*(3), 537–570.

Engelsman, E. and A. van Raan (1994). A patent-based cartography of technology. *Research Policy 23*(1), 1–26.

Field, E. (2007). Entitled to work: Urban property rights and labor supply in peru. *The Quarterly Journal of Economics 122*(4), 1561–1602.

Fu, K., J. Chan, J. Cagan, K. Kotovsky, C. Schunn, K. Wood, and P. Pillar (2012, 08). The meaning of "near" and "far": The impact of structuring design databases and the effect of distance of analogy on design output. Volume 135.

Galiani, S. and E. Schargrodsky (2010). Property rights for the poor: Effects of land titling. *Journal of Public Economics 94*(9-10), 700–729.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

Gollin, D., D. Lagakos, and M. Waugh (2014). Agricultural productivity differences across countries. *American Economic Review 104*(5), 165–70.

Gollin, D. and C. Udry (2021). Heterogeneity, measurement error, and misallocation: Evidence from african agriculture. *Journal of Political Economy 129*(1), 1 – 80.

Gottlieb, C. and J. Grobovsek (2019). Communal land and agricultural productivity. *Journal of Development Economics 138*(C), 135–152.

Griliches, Z. (1979, Spring). Issues in Assessing the Contribution of Research and Development to Productivity Growth. *Bell Journal of Economics 10*(1), 92–116.

Griliches, Z. (1986, March). Productivity, R&D, and the Basic Research at the Firm Level in the 1970's. *American Economic Review 76*(1), 141–154.

Guner, N., G. Ventura, and Y. Xu (2008). Macroeconomic implications of size-dependent policies. *Review of Economic Dynamics 11*(4), 721–744.

Hidalgo, C., B. Klinger, A.-L. Barabasi, and R. Hausmann (2007, 08). The product space conditions the development of nations. *Science (New York, N.Y.) 317*, 482–7.

Hopenhayn, H. A. (2014). Firms, misallocation, and aggregate productivity: A review. *Annual Review of Economics 6*(1), 735–770.

Hornbeck, R. and E. Moretti (2019, April). Estimating Who Benefits from Productivity Growth: Direct and Indirect Effects of City Manufacturing TFP Growth on Wages, Rents, and Inequality. IZA Discussion Papers 12277, Institute of Labor Economics (IZA).

Hsieh, C.-T. (1999). Productivity growth and factor prices in east asia. *American Economic Review 89*(2), 133–138.

Hsieh, C.-T. (2002). What explains the industrial revolution in east asia? evidence from the factor markets. *American Economic Review 92*(3), 502–526.

Hsieh, C.-T. and P. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Inklaar, R., H. de Jong, J. Bolt, and J. van Zanden (2018). Rebasing 'maddison': new income comparisons and the shape of long-run economic development. GGDC Research Memorandum GD-174, Groningen Growth and Development Centre, University of Groningen.

Jaffe, A. (1986). Technological opportunity and spillovers of r&d: Evidence from firms' patents, profits, and market value. *American Economic Review 76*(5), 984–1001.

Jaffe, A., M. Trajtenberg, and R. Henderson (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics 108*(3), 577–598.

Jermann, U. and V. Quadrini (2012). Macroeconomic effects of financial shocks. *American Economic Review 102*(1), 238–71.

Jones, C. I. (1995, August). R&D-Based Models of Economic Growth. *Journal of Political Economy 103*(4), 759–784.

Kalemli-Ozcan, S. and B. Sorensen (2012). Misallocation, property rights, and access to finance: Evidence from within and across africa. NBER Working Papers 18030, National Bureau of Economic Research, Inc.

Kang, B. and G. Tarasconi (2016). Patstat revisited: Suggestions for better usage. *World Patent Information 46*, 56–63.

Kay, L., N. Newman, J. Youtie, A. L. Porter, and I. Rafols (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology 65*(12), 2432–2443.

Keller, W. (2002, March). Geographic localization of international technology diffusion. *American Economic Review 92*(1), 120–142.

Keller, W. (2004, September). International Technology Diffusion. *Journal of Economic Literature 42*(3), 752–782.

Keller, W. and S. R. Yeaple (2013, June). The Gravity of Knowledge. *American Economic Review 103*(4), 1414–1444.

King, R. and R. Levine (1993). Finance, entrepreneurship and growth: Theory and evidence. *Journal of Monetary Economics 32*(3), 513–542.

Kleinman, B., E. Liu, and S. Redding (2021). Dynamic spatial general equilibrium. Lse research online documents on economics, London School of Economics and Political Science, LSE Library.

Klenow, P. and A. Rodríguez-Clare (1997). The Neoclassical Revival in Growth Economics: Has It Gone Too Far? In *NBER Macroeconomics Annual 1997, Volume 12*, NBER Chapters, pp. 73–114. National Bureau of Economic Research, Inc.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics 132*(2), 665–712.

Kogler, D., D. Rigby, and I. Tucker (2013, 09). Mapping knowledge space and technological relatedness in us cities. *European Planning Studies 21*, 1374.

Kortum, S. S. (1997, November). Research, Patenting, and Technological Change. *Econometrica 65*(6), 1389–1420.

Levinsohn, J. and A. Petrin (2000). Estimating production functions using inputs to control for unobservables. NBER Working Papers 7819, National Bureau of Economic Research, Inc.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science and Technology 59*(10), 1582–1597.

Leydesdorff, L., D. Kushnir, and I. Rafols (2014, March). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics 98*(3), 1583–1599.

Leydesdorff, L. and L. Vaughan (2006). Co-occurrence matrices and their applications in information science: Extending aca to the web environment. *Journal of the American Society for Information Science and Technology 57*(12), 1616–1628.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics 134*(4), 1883–1948.

Liu, E. and S. Ma (2021). Innovation networks and innovation policy. NBER Working Papers 29607, National Bureau of Economic Research, Inc.

Macours, K., A. de Janvry, and E. Sadoulet (2010). Insecurity of property rights and social matching in the tenancy market. *European Economic Review 54*(7), 880–899.

Melitz, M. and S. Redding (2021). Trade and innovation. CEPR Discussion Papers 16264, C.E.P.R. Discussion Papers.

Mendoza, E. (2010). Sudden stops, financial crises, and leverage. *American Economic Review 100*(5), 1941–66.

Moretti, E., C. Steinwender, and J. V. Reenen (2019, November). The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers. NBER Working Papers 26483, National Bureau of Economic Research, Inc.

Murata, Y., R. Nakajima, R. Okamoto, and R. Tamura (2014, December). Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach. *The Review of Economics and Statistics 96*(5), 967–985.

Nakamura, H., S. Suzuki, I. Sakata, and Y. Kajikawa (2015). Knowledge combination modeling: The measurement of knowledge similarity between different technological domains. *Technological Forecasting and Social Change 94*(C), 187–201.

Ngai, L. R., C. Pissarides, and J. Wang (2019). China's mobility barriers and employment allocations. *Journal of the European Economic Association 17*(5), 1617–1653.

Nicholas, T. (2010). The role of independent invention in u.s. technological development, 1880–1930. *The Journal of Economic History 70*(01), 57–82.

Oberfield, E. (2013). Productivity and misallocation during a crisis: Evidence from the chilean crisis of 1982. *Review of Economic Dynamics 16*(1), 100–119.

Oberfield, E. (2018). A theory of input–output architecture. *Econometrica 86*(2), 559–589.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–97.

Olmstead-Rumsey, J. (2019, April). Market Concentration and the Productivity Slowdown. MPRA Paper 93260, University Library of Munich, Germany.

Packalen, M. and J. Bhattacharya (2015, January). Cities and ideas. Working Paper 20921, National Bureau of Economic Research.

Pande, R. and C. Udry (2005). Institutions and development:a view from below. Working papers, Economic Growth Center, Yale University.

Peretto, P. (1998). Technological change and population growth. *Journal of Economic Growth 3*(4), 283–311.

Perla, J. and C. Tonetti (2014). Equilibrium imitation and growth. *Journal of Political Economy 122*(1), pp. 52–76.

Petralia, S., P.-A. Balland, and D. L. Rigby (2016, 08). Unveiling the geography of historical patents in the united states from 1836 to 1975. *Scientific Data 3*, 160074.

Philippe Aghion, Antonin Bergeaud, M. L. and M. J. Melitz (2018, April). The Impact of Exports on Innovation: Theory and Evidence. Working papers 678, Banque de France.

Piza, C. and M. J. S. B. de Moura (2016). The effect of a land titling programme on households' access to credit. *Journal of Development Effectiveness 8*(1), 129–155.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous plants. *Review of Economic Dynamics 11*(4), 707–720.

Restuccia, D. and R. Rogerson (2013). Misallocation and productivity. *Review of Economic Dynamics 16*(1), 1–10.

Restuccia, D. and R. Rogerson (2017). The causes and costs of misallocation. *Journal of Economic Perspectives 31*(3), 151–74.

Restuccia, D. and R. Santaeulalia-Llopis (2017). Land misallocation and productivity. NBER Working Papers 23128, National Bureau of Economic Research, Inc.

Romer, P. (1990). Endogenous technological change. *Journal of Political Economy 98*(5), S71–102.

Santaeulàlia-Llopis, R. and Y. m. o. r. a. c. u. i. a. l. e. f. e. d. a. c. H. Zheng (2016). The price of growth: Consumption insurance in china 1989-2009. Economics Working Papers ECO2016/13, European University Institute.

Schreiber, L. A. (2017). Registering rural rights: Village land titling in tanzania, 2008-2017.

Segerstrom, P. S. (1998, December). Endogenous Growth without Scale Effects. *American Economic Review 88*(5), 1290–1310.

Tabellini, M. (2020). Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration. *Review of Economic Studies 87*(1), 454–486.

Timmer, M., E. Dietzenbacher, B. Los, R. Stehrer, and G. de Vries (2015). An illustrated user guide to the world input–output database: the case of global automotive production. *Review of International Economics 23*(3), 575–605.

URT (2016). Draft national land policy 2016. *United Republic of Tanzania, Ministry of Lands, Housing and Human Settlements Development, Dar es Salaam*.

Van Looy, B., C. Vereyen, and U. Schmoch (2014). Patent statistics: Concordance ipc v8 – nace rev.2. *EUROSTAT*.

von Wartburg, I., T. Teichert, and K. Rost (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy 34*(10), 1591–1607.

Young, A. (1998). Growth without scale effects. *Journal of Political Economy 106*(1), 41–63.

Zegarra, E., J. Escobal, and U. Aldana (2011). Titling, credit constraints and rental markets in rural peru: Exploring channels and conditioned impacts. IDB Publications (Working Papers) 3707, Inter-American Development Bank.

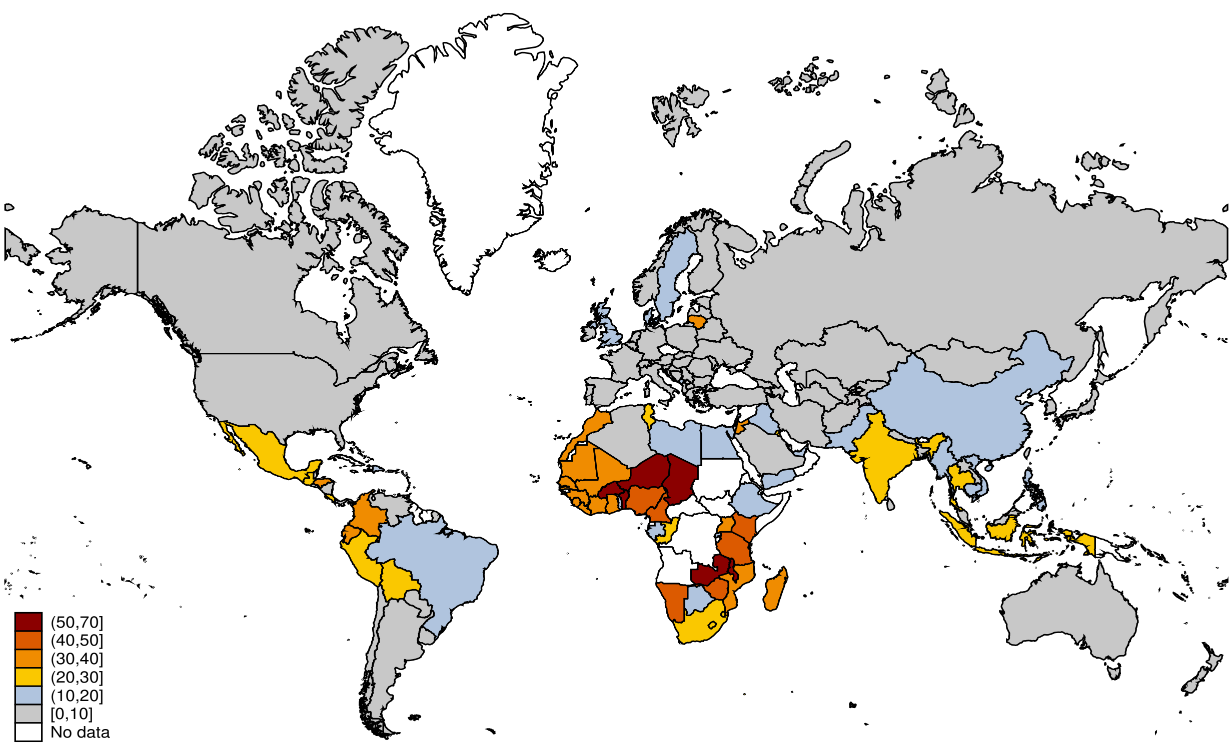
Zitt, M., E. Bassecoulard, and Y. Okubo (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics 47*(3), 627–657.

# Appendix A

# Appendix to Chapter One

## A.1 Additional Tables and Figures

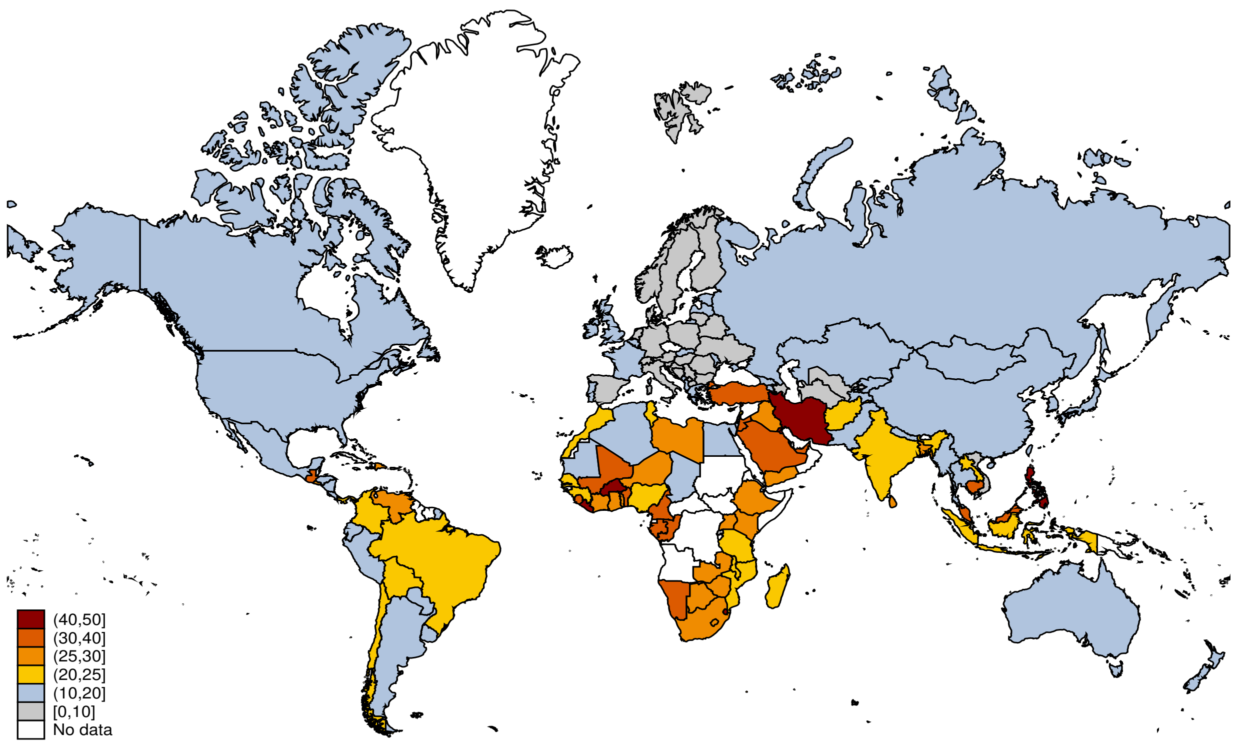
Figure A.1: Share of Land with No Official or Unofficial Document (2020)




*Data Source:* Prindex.

*Notes:* Legend reflects the share of land with no documentation.

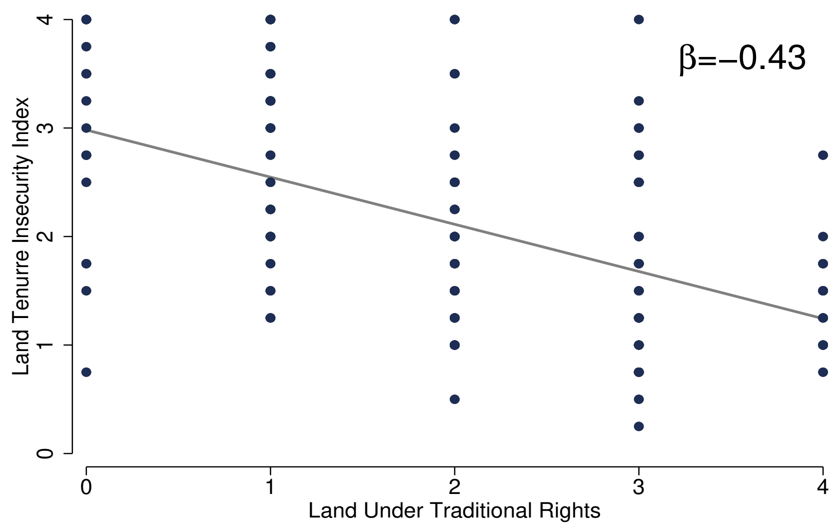
Figure A.2: Share of Adults that Feel Insecure about Their Property (2020)




*Data Source:* Prindex.

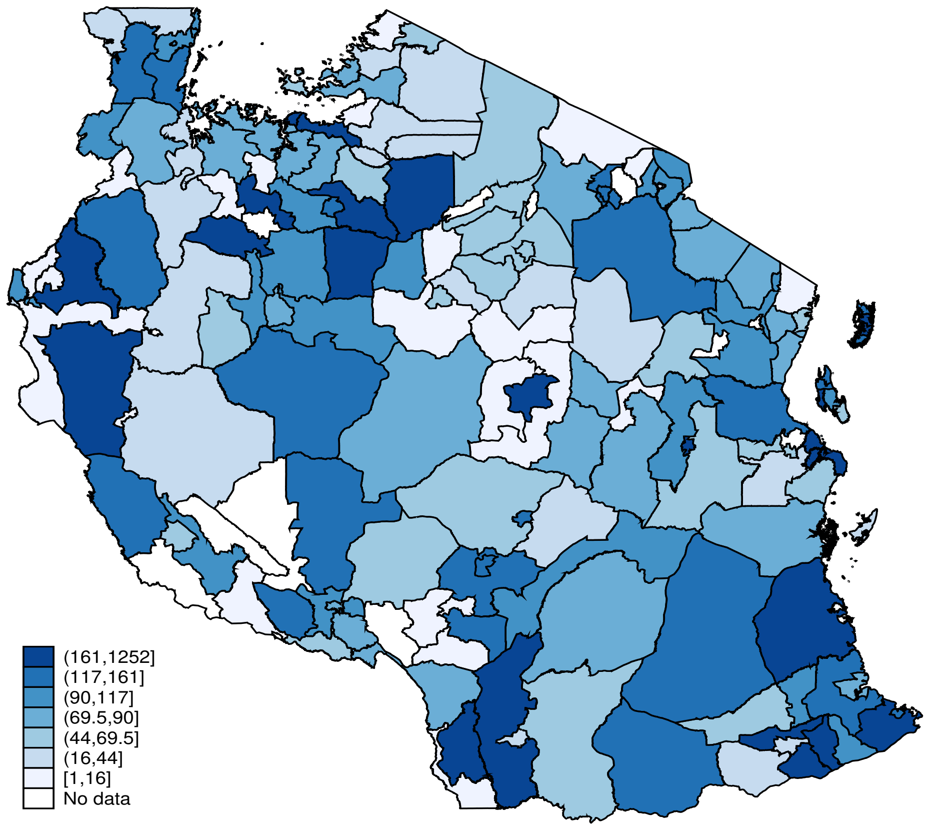
*Notes:* Legend reflects the share of surveyed adults that feel insecure about their property.

Figure A.3: Share of Traditional Land and Land Tenure Insecurity




*Notes:* The land tenure insecurity index ranges from 0 to 4, with 0 being the highest level of land insecurity. Land under traditional system measures the share of rural land under the traditional rights system, and ranges from 0 to 4, with 0 indicating that there is no land under traditional system. Both indicators are obtained from The Institutional Profiles Database (IPD) of the Centre d'Etudes Prospectives et d'Informations (CEPII), and are a composite measures of several factors.

Figure A.4: Sample coverage




*Notes:* Legend reflects the number of households surveyed in a given district.

Table A.1: Summary statistics (TPNS 2008-2015)

| Variable | Mean | Median | Std. Dev. |
|---|---|---|---|
| Total harvest (ths TZS) | 722.9 | 164.4 | 25,460 |
| Yield (ths TZS/acre) | 163.3 | 62.5 | 2,288 |
| Land cultivated (acres) | 5.5 | 2.8 | 12.3 |
| Land available (acres) | 6.2 | 3.0 | 14.9 |
| Total labor (per-day) | 172.9 | 116.0 | 185.7 |
| HH labor (per-day) | 158.6 | 104.0 | 178.2 |
| Hired labor (per-day) | 14.3 | 0 | 37.9 |
| Daily wage (ths TZS) | 3.8 | 2.5 | 4.7 |
| Capital (ths TZS) | 1,887.9 | 13.5 | 7,850.4 |
| Chemicals (ths TZS) | 2.5 | 0 | 7.6 |
| Variable | % of obs | | |
| HH own/cultivate plot | 65.4 | - | - |
| Plots cultivated | 85.0 | - | - |
| Land utilization | 85.2 | - | - |
| Hire workers | 43.1 | - | - |
| Use chemicals | 35.5 | - | - |
| Can leave plot | 86.5 | - | - |
| Right sell/coll | 68.4 | - | - |
| Title/certificate | 12.5 | - | - |
| Took loan (1 yr) | 10.5 | - | - |
| Took loan (ag) (1 yr) | 1.3 | - | - |
| Took loan (bus) (1 yr) | 2.7 | - | - |

*Notes:* Average exchange rate in 2013 was ≈ 1,600 TZS per 1 USD.

Table A.2: Descriptive Statistics for Plots Based on Title

| | Title | | No Title | | *Differences* | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Diff. | p-value |
| | | | **Panel A: Land Characteristics** | | | |
| Soil quality | 1.19 | 0.75 | 0.19 | 0.75 | 0.00 | 0.98 |
| % Slope flat bottom | 0.51 | 0.50 | 0.52 | 0.50 | 0.00 | 0.69 |
| % Slope flat top | 0.07 | 0.25 | 0.06 | 0.23 | 0.01 | <0.01 |
| % Slightly sloped | 0.24 | 0.43 | 0.25 | 0.43 | -0.01 | 0.11 |
| % Slope steep | 0.03 | 0.16 | 0.03 | 0.18 | -0.01 | 0.05 |
| % Soil clay | 0.15 | 0.36 | 0.15 | 0.36 | 0.01 | 0.43 |
| % Soil loam | 0.50 | 0.50 | 0.53 | 0.50 | -0.02 | 0.01 |
| % Soil sandy | 0.17 | 0.37 | 0.16 | 0.37 | 0.01 | 0.21 |
| | | | **Panel B: Plot Characteristics** | | | |
| Plot area (acres) | 3.77 | 17.7 | 2.73 | 6.36 | 1.04 | <0.01 |
| Distance to home (km) | 7.26 | 35.8 | 4.93 | 25.4 | 2.33 | <0.01 |
| Distance to market (km) | 10.0 | 14.5 | 9.72 | 13.5 | 0.30 | 0.27 |
| Distance to road (km) | 2.17 | 4.61 | 2.32 | 5.09 | -0.15 | 0.14 |
| % Erosion control | 0.1 | 0.3 | 0.09 | 0.29 | 0.01 | 0.10 |
| % Irrigation system | 0.02 | 0.14 | 0.02 | 0.12 | 0.00 | 0.10 |
| | | | **Panel C: Agricultural Practices** | | | |
| % Use inorganic fertilizer | 0.10 | 0.30 | 0.09 | 0.29 | 0.01 | 0.26 |
| % Use organic fertilizer | 0.13 | 0.35 | 0.09 | 0.29 | 0.05 | <0.01 |
| % Hire labor (outside HH) | 0.29 | 0.45 | 0.25 | 0.43 | 0.04 | <0.01 |
| % Use input on credit | 0.01 | 0.10 | 0.01 | 0.10 | 0.00 | 0.71 |
| % Use pesticides | 0.07 | 0.26 | 0.08 | 0.27 | 0.01 | 0.17 |
| % Use animal traction | 0.22 | 0.42 | 0.28 | 0.45 | -0.05 | <0.01 |
| % Use mechanization | 0.06 | 0.23 | 0.05 | 0.21 | 0.01 | 0.25 |
| Labor per acre (person-days) | 102.9 | 388.9 | 101.3 | 503.2 | 1.54 | 0.87 |
| % Use credit for agriculture | 0.04 | 0.19 | 0.02 | 0.13 | 0.02 | <0.01 |
| Land utilization (%) | 0.93 | 0.18 | 0.93 | 0.19 | 0.00 | 0.50 |
| | N=3,030 | | N=19,808 | | | |

Table A.3: Descriptive Statistics for Plots Based on the Right to Sell/Use as Collateral

| | Has Right to Sell | | Does Not Have | | *Differences* | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Diff. | p-value |
| **Panel A: Land Characteristics** | | | | | | |
| Soil quality | 1.19 | 0.76 | 1.19 | 0.71 | 0.00 | 0.76 |
| % Slope flat bottom | 0.50 | 0.50 | 0.55 | 0.50 | -0.04 | <0.01 |
| % Slope flat top | 0.06 | 0.23 | 0.06 | 0.24 | 0.00 | 0.25 |
| % Slightly sloped | 0.25 | 0.43 | 0.25 | 0.43 | -0.01 | 0.39 |
| % Slope steep | 0.04 | 0.19 | 0.03 | 0.16 | 0.01 | <0.01 |
| % Soil clay | 0.15 | 0.36 | 0.14 | 0.34 | 0.01 | <0.01 |
| % Soil loam | 0.51 | 0.50 | 0.55 | 0.50 | -0.04 | <0.01 |
| % Soil sandy | 0.16 | 0.36 | 0.17 | 0.37 | -0.01 | 0.06 |
| **Panel B: Plot Characteristics** | | | | | | |
| Plot area (acres) | 3.20 | 8.7 | 1.99 | 8.9 | 1.20 | <0.01 |
| Distance to home (km) | 5.36 | 27.5 | 4.93 | 25.5 | 0.43 | 0.28 |
| Distance to market (km) | 10.0 | 13.7 | 9.01 | 13.3 | 0.93 | <0.01 |
| Distance to road (km) | 2.43 | 5.39 | 1.95 | 3.09 | 0.48 | <0.01 |
| % Erosion control | 0.09 | 0.29 | 0.09 | 0.28 | 0.01 | 0.13 |
| % Irrigation system | 0.02 | 0.12 | 0.02 | 0.13 | 0.00 | 0.07 |
| **Panel C: Agricultural Practices** | | | | | | |
| % Use inorganic fertilizer | 0.10 | 0.30 | 0.08 | 0.27 | 0.02 | <0.01 |
| % Use organic fertilizer | 0.10 | 0.30 | 0.10 | 0.29 | 0.00 | 0.56 |
| % Hire labor (outside HH) | 0.25 | 0.43 | 0.29 | 0.45 | -0.04 | <0.01 |
| % Use input on credit | 0.01 | 0.10 | 0.01 | 0.08 | 0.01 | <0.01 |
| % Use pesticides | 0.09 | 0.28 | 0.06 | 0.23 | 0.03 | <0.01 |
| % Use animal traction | 0.28 | 0.45 | 0.24 | 0.43 | 0.04 | <0.01 |
| % Use mechanization | 0.05 | 0.21 | 0.05 | 0.31 | -0.01 | 0.26 |
| Labor per acre (person-days) | 91.1 | 497.7 | 129.2 | 466.3 | -38.1 | <0.01 |
| % Use credit for agriculture | 0.02 | 0.15 | 0.01 | 0.12 | 0.01 | <0.01 |
| Land utilization (%) | 0.93 | 0.19 | 0.94 | 0.17 | -0.02 | <0.01 |
| | N=16,590 | | N=6,246 | | | |

Table A.4: Descriptive Statistics for Plots Based on Ability to Leave Land Fallow without Fear to Lose Land

| | Can Leave Fallow | | Can Not Leave | | *Differences* | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Diff. | p-value |
| | | | **Panel A: Land Characteristics** | | | |
| Soil quality | 1.19 | 0.75 | 1.23 | 0.73 | -0.04 | <0.01 |
| % Slope flat bottom | 0.51 | 0.50 | 0.57 | 0.50 | -0.05 | <0.01 |
| % Slope flat top | 0.06 | 0.23 | 0.06 | 0.24 | -0.01 | 0.09 |
| % Slightly sloped | 0.25 | 0.43 | 0.24 | 0.43 | 0.01 | 0.26 |
| % Slope steep | 0.03 | 0.18 | 0.02 | 0.16 | 0.01 | <0.01 |
| % Soil clay | 0.15 | 0.36 | 0.18 | 0.38 | -0.03 | <0.01 |
| % Soil loam | 0.52 | 0.50 | 0.55 | 0.50 | -0.03 | <0.01 |
| % Soil sandy | 0.16 | 0.37 | 0.15 | 0.35 | 0.01 | 0.08 |
| | | | **Panel B: Plot Characteristics** | | | |
| Plot area (acres) | 2.93 | 9.1 | 1.97 | 3.5 | 1.20 | <0.01 |
| Distance to home (km) | 5.19 | 27.0 | 8.60 | 40.1 | -3.40 | <0.01 |
| Distance to market (km) | 9.77 | 14.1 | 10.3 | 16.5 | -0.53 | 0.05 |
| Distance to road (km) | 2.31 | 7.04 | 2.61 | 5.96 | -0.29 | 0.02 |
| % Erosion control | 0.09 | 0.29 | 0.10 | 0.30 | -0.01 | 0.34 |
| % Irrigation system | 0.02 | 0.12 | 0.03 | 0.17 | -0.01 | <0.01 |
| | | | **Panel C: Agricultural Practices** | | | |
| % Use inorganic fertilizer | 0.10 | 0.29 | 0.11 | 0.31 | -0.01 | 0.02 |
| % Use organic fertilizer | 0.10 | 0.30 | 0.08 | 0.28 | 0.01 | 0.01 |
| % Hire labor (outside HH) | 0.25 | 0.43 | 0.34 | 0.47 | -0.08 | <0.01 |
| % Use input on credit | 0.01 | 0.10 | 0.01 | 0.10 | 0.00 | 0.62 |
| % Use pesticides | 0.08 | 0.27 | 0.08 | 0.28 | 0.00 | 0.52 |
| % Use animal traction | 0.27 | 0.44 | 0.30 | 0.46 | -0.03 | 0.03 |
| % Use mechanization | 0.05 | 0.21 | 0.08 | 0.26 | -0.03 | <0.01 |
| Labor per acre (person-days) | 101.7 | 506.9 | 94.6 | 220.8 | 7.04 | 0.43 |
| % Use credit for agriculture | 0.02 | 0.14 | 0.03 | 0.16 | -0.01 | 0.07 |
| Land utilization (%) | 0.93 | 0.19 | 0.96 | 0.15 | -0.03 | <0.01 |
| | N=20,960 | | N=3,283 | | | |

Table A.5: Descriptive Statistics for Plots Based on Whether Land Was Obtained/Used for Free

| | Not free | | For Free | | *Differences* | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Diff. | p-value |
| | | | **Panel A: Land Characteristics** | | | |
| Soil quality | 1.19 | 0.75 | 1.24 | 0.69 | -0.06 | <0.01 |
| % Slope flat bottom | 0.51 | 0.50 | 0.60 | 0.49 | -0.09 | <0.01 |
| % Slope flat top | 0.06 | 0.23 | 0.05 | 0.22 | 0.01 | 0.06 |
| % Slightly sloped | 0.25 | 0.43 | 0.24 | 0.42 | 0.01 | 0.15 |
| % Slope steep | 0.04 | 0.18 | 0.02 | 0.14 | 0.02 | <0.01 |
| % Soil clay | 0.15 | 0.36 | 0.15 | 0.35 | 0.01 | 0.29 |
| % Soil loam | 0.52 | 0.50 | 0.56 | 0.50 | -0.04 | <0.01 |
| % Soil sandy | 0.16 | 0.36 | 0.17 | 0.38 | -0.02 | 0.02 |
| | | | **Panel B: Plot Characteristics** | | | |
| Plot area (acres) | 2.99 | 9.04 | 1.52 | 3.0 | 1.47 | <0.01 |
| Distance to home (km) | 5.61 | 29.1 | 5.91 | 29.8 | -0.29 | 0.61 |
| Distance to market (km) | 9.98 | 14.5 | 8.8 | 14.4 | 1.19 | <0.01 |
| Distance to road (km) | 2.11 | 4.08 | 2.39 | 7.21 | 0.28 | 0.04 |
| % Erosion control | 0.10 | 0.29 | 0.06 | 0.25 | 0.03 | <0.01 |
| % Irrigation system | 0.02 | 0.13 | 0.02 | 0.12 | 0.00 | 0.45 |
| | | | **Panel C: Agricultural Practices** | | | |
| % Use inorganic fertilizer | 0.10 | 0.30 | 0.09 | 0.28 | 0.01 | 0.09 |
| % Use organic fertilizer | 0.10 | 0.30 | 0.06 | 0.24 | 0.04 | <0.01 |
| % Hire labor (outside HH) | 0.26 | 0.44 | 0.29 | 0.45 | -0.03 | <0.01 |
| % Use input on credit | 0.01 | 0.10 | 0.01 | 0.08 | 0.00 | 0.02 |
| % Use pesticides | 0.09 | 0.28 | 0.05 | 0.22 | 0.03 | <0.01 |
| % Use animal traction | 0.29 | 0.45 | 0.23 | 0.42 | 0.06 | <0.01 |
| % Use mechanization | 0.05 | 0.22 | 0.06 | 0.24 | -0.01 | 0.41 |
| Labor per acre (person-days) | 96.1 | 486.6 | 113.4 | 412.4 | -37.2 | <0.01 |
| % Use credit for agriculture | 0.02 | 0.15 | 0.01 | 0.11 | 0.01 | <0.01 |
| Land utilization (%) | 0.93 | 0.19 | 0.96 | 0.15 | -0.02 | <0.01 |
| | N=21,265 | | N=2,279 | | | |

Table A.6: Production function estimates

| | (OLS) | (OLS FE) | (DP) |
|---|---|---|---|
| log(Land) | 0.347 | 0.266 | 0.280 |
| | (0.018) | (0.027) | (0.042) |
| log(Labor) | 0.411 | 0.348 | 0.446 |
| | (0.027) | (0.030) | (0.081) |
| log(Capital) | 0.111 | 0.048 | 0.036 |
| | (0.008) | (0.010) | (0.020) |
| $\beta_l$ | | | 0.268 |
| $\beta_n$ | | | 0.421 |
| $\beta_k$ | | | 0.049 |
| $\rho$ | | | 0.371 |
| Return to scale | 0.87 | 0.66 | 0.74 |
| Test on common factor restrictions | | | 0.832 |
| # obs. | 8,949 | 6,073 | 3,641 |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household level. Regressions include year FE, OLS regressions - district-year FE.

Table A.7: Factor ratios: Capital

| | ln(land) | | | | |
|---|---|---|---|---|---|
| | | leave fallow | right to sell | title | obtain free |
| ln(Capital) | 0.177 | 0.147 | 0.145 | 0.173 | 0.181 |
| | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) |
| ln(Capital) × land_rights | | 0.033 | 0.043 | 0.022 | -0.048 |
| | | (0.003) | (0.002) | (0.004) | (0.003) |
| ln(Capital) × credit | | 0.034 | 0.032 | 0.033 | 0.033 |
| | | (0.007) | (0.007) | (0.007) | (0.007) |
| # obs. | 10,047 | 10,047 | 10,047 | 10,047 | 10,047 |
| # households | 5,513 | 5,513 | 5,513 | 5,513 | 5,513 |
| Wave#District FE | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels.

Table A.8: Factor ratios: Labor

| | ln(land) | | | | |
|---|---|---|---|---|---|
| | | leave fallow | right to sell | title | obtain free |
| ln(Labor) | 0.586 | 0.528 | 0.515 | 0.576 | 0.583 |
| | (0.013) | (0.015) | (0.015) | (0.013) | (0.013) |
| ln(Labor) × land_rights | | 0.055 | 0.072 | 0.042 | -0.076 |
| | | (0.006) | (0.005) | (0.008) | (0.007) |
| ln(Labor) × credit | | 0.054 | 0.050 | 0.050 | 0.051 |
| | | (0.014) | (0.014) | (0.014) | (0.014) |
| # obs. | 10,054 | 10,054 | 10,054 | 10,054 | 10,054 |
| # households | 5,515 | 5,515 | 5,515 | 5,515 | 5,515 |
| Wave#District FE | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels.

Table A.9: CES Production Function Estimates

| | (1) | (2) |
|---|---|---|
| $\epsilon$ | 1.186 | 1.186 |
| | (0.041) | (0.042) |
| $\sigma$ | 0.851 | 0.841 |
| | (0.015) | (0.015) |
| $\alpha$ | 0.602 | 0.602 |
| | (0.039) | (0.039) |
| $\beta$ | 0.364 | 0.364 |
| | (0.030) | (0.030) |
| # obs. | 8,959 | 8,959 |
| Unexpected shocks | | ✓ |

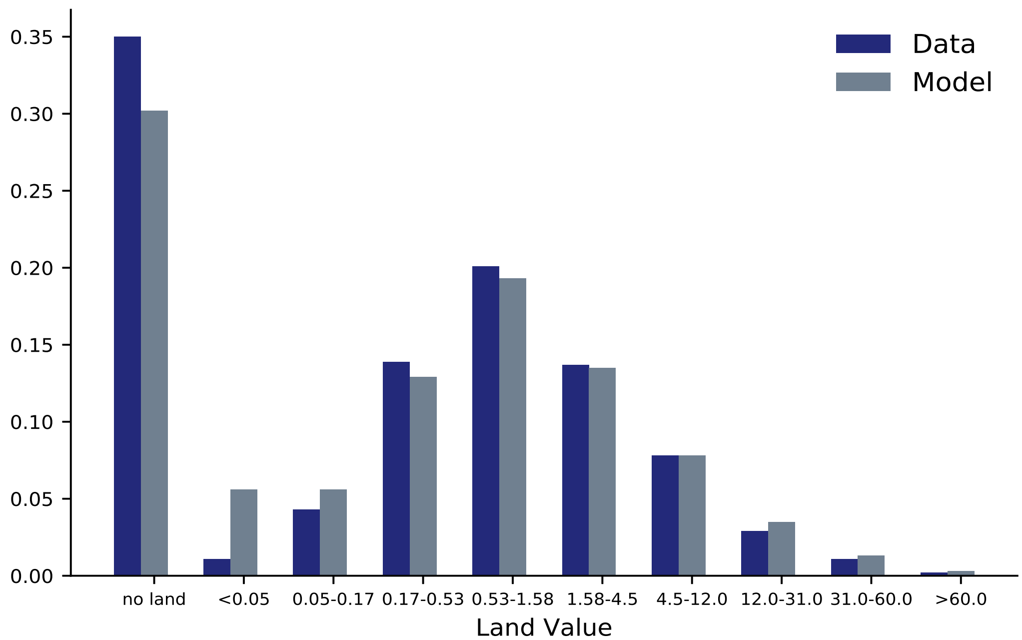
*Notes:* Estimated using fixed-effects nonlinear least-squares. Robust standard errors (in parentheses) are two-way clustered at the district and household levels.

Table A.10: Land Misallocation: Across Time Variation

| | ln(land) | | | |
|---|---|---|---|---|
| | leave fallow | right to sell | title | obtain free |
| HH productivity | -0.007 | -0.010 | -0.006 | -0.002 |
| | (0.010) | (0.009) | (0.008) | (0.008) |
| HH productivity × land_rights | 0.002 | 0.009 | 0.010 | -0.023 |
| | (0.005) | (0.004) | (0.005) | (0.008) |
| HH productivity × credit | 0.024 | 0.024 | 0.024 | 0.025 |
| | (0.010) | (0.010) | (0.010) | (0.010) |
| # obs. | 6,043 | 6,043 | 6,043 | 6,043 |
| # households | 2,218 | 2,218 | 2,218 | 2,218 |
| Wave#District FE | ✓ | ✓ | ✓ | ✓ |
| HH FE | ✓ | ✓ | ✓ | ✓ |
| $R^2$ | 0.833 | 0.833 | 0.833 | 0.833 |

*Notes:* Robust standard errors (in parentheses) are two-way clustered at the district and household levels.

Figure A.5: Distribution of Land: Model and Data




*Notes:* The distribution is based on price of land in mln TZS such that it is equispaced on a log scale.

Figure A.6: Lorenz Curve for Consumption

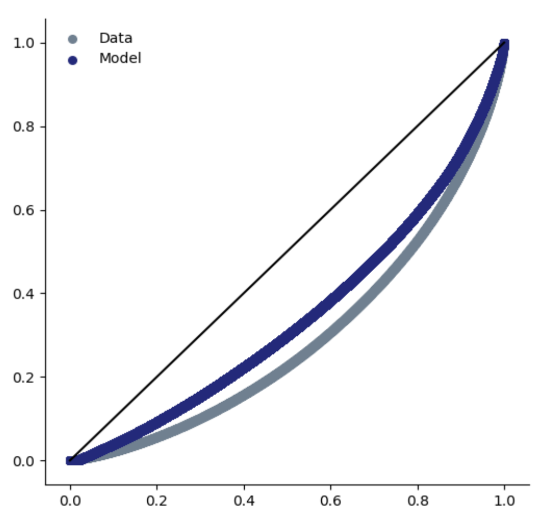
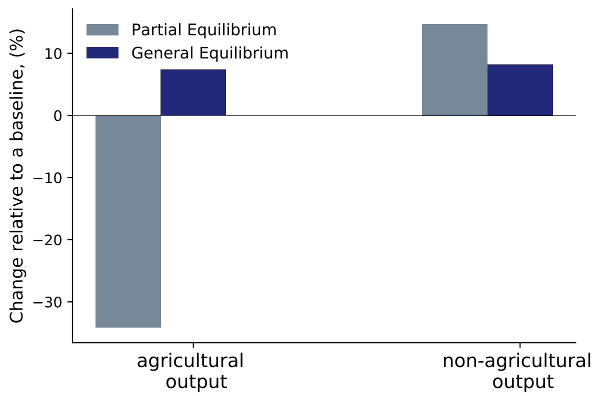
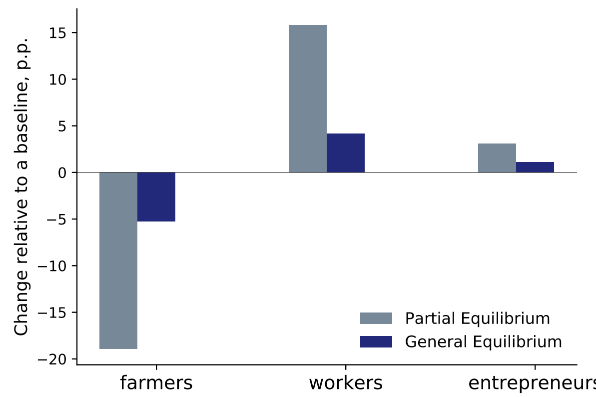


Figure A.7: The Effects of Land Reform

(a) Change in Output




(b) Change in Occupation Shares




*Notes:* The effects of land reform in partial equilibrium are estimated keeping all prices fixed.

## A.2 Land Tenure System in Tanzania

The current land tenure and administration system in Tanzania has evolved from the Germans and British colonial rules and incorporates the features of pre-colonial, colonial and post-colonial tenures.

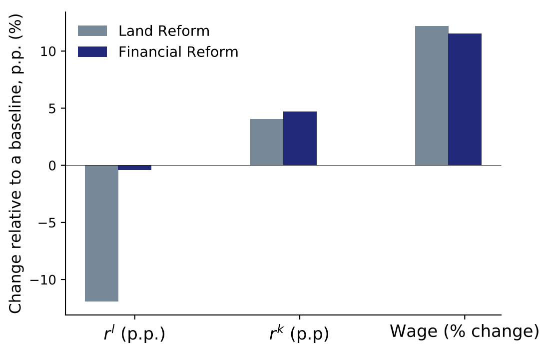
### A.2.1 Brief Historical Context

Prior to colonial era all land belonged to different tribes and the general characteristics of land holdings were based on the culture of each tribe. The common principal of most tribes was that land belongs to its user, which means that when the family is no longer using the land, it is reallocated to another family.

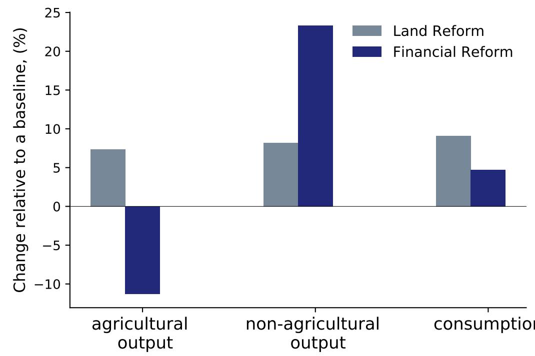
Colonial period can be split into two sub-periods – the German Era (1884-1917) and the British Era (1918-1961). The Germans imposed a declaration in 1895 that all land in German East Africa to be unowned Crown Land vested in the German Empire. The only exception was land where proof of ownership could be shown either though documentations,

Figure A.8: The Effects of Land and Financial Reforms

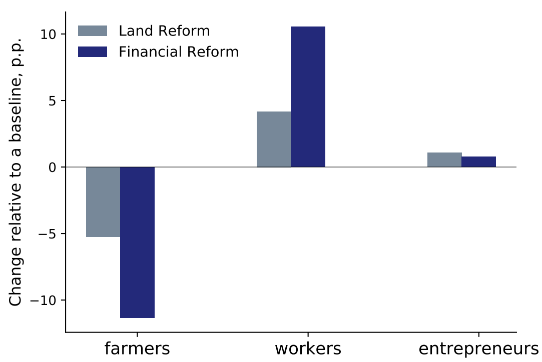
(a) Change in Prices




(b) Change in Output/Consumption




(c) Change in Occupation Shares




(d) Change in Average Productivity

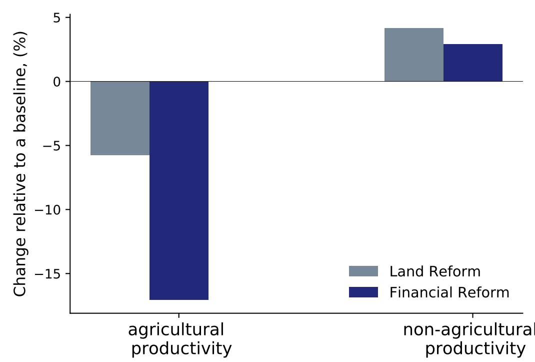
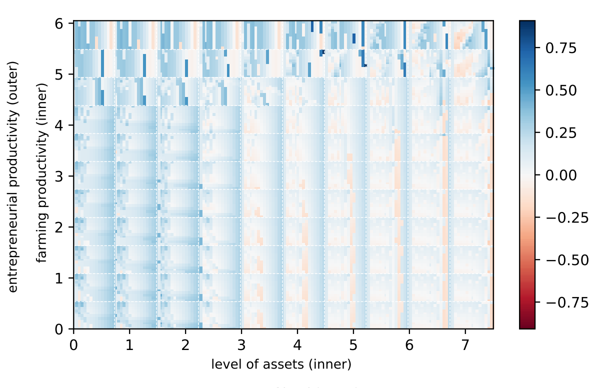



Figure A.9: Changes in Welfare Distribution: Financial Reform

(a) Communal Land Holders




(b) Private Land Holders

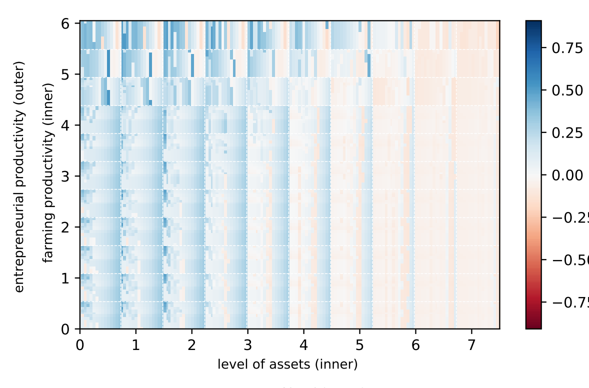
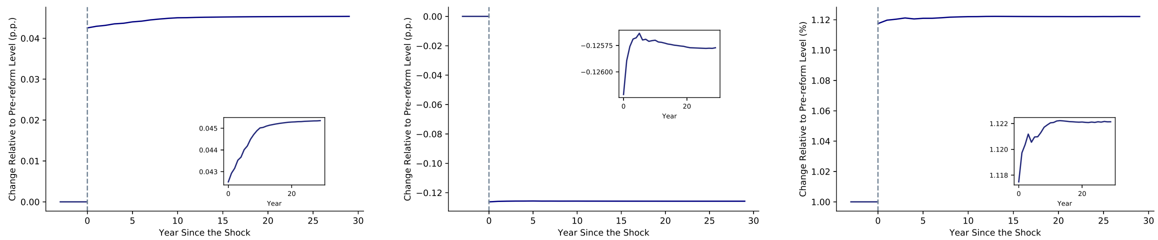


Figure A.10: Postreform Transition Dynamics for Prices

(a) Capital Interest Rate (b) Land Rental Rate (c) Wage




*Notes:* Prices are shown as deviations from their respective pre-reform values.

or through effective occupation. The main types of tenures established during the German era were: i) Freeholds granted mainly to European Settlers ii) Leaseholds iii) Crown Land – unowned land determined by the commissions, and iv) Customary Land Tenure for the land occupied by the natives.

Under the British rule, the first land tenure statute was the Land Ordinance of 1923, which declared all land, but freeholds acquired before, as being public land. Under 1928 extension, anyone holding land under customary tenure was declared a legitimate holder of the land. The main types of tenures established during the British era were: i) Freeholds ii) Granted Rights of occupancy (long-term for 33, 66 or 99 years; short term for less than 6 years; and from year to year) iii) Deemed rights of occupancy (in urban areas and rural areas, which was mostly held by native communities) iv) Public land.

### A.2.2 Land Tenure in the Post-Independence Era

The Land Ordinance 1923 continued to be the principal document on land tenure till 1999. In 1995 a National Land Policy was published and two pieces of legislation were introduced in 1999: Village Land Act No 5, which covered rural land, and Land Act No 4, which covered general land, including urban land.

Around 70 percent of land in the Mainland of Tanzania is considered to be Village Land

(80 percent of population), 28 percent is Reserved land (i.e. national parks), and 2 percent is general land (mainly urban, 20 percent of the population).

Village land is regulated by the Village Land Act, and divides land into three categories: communal land, occupied land and future (or reserved) land. The Village Land Act empowers village councils to maintain a register of village land. The Acts recognize two forms of tenure: i) the granted right of occupancy, and ii) customary right of occupancy.

As for now, and for the period of study in this paper, Tanzania presents a dynamic land tenure context. All land in Tanzania is owned by the state and held in trust by the president, but individuals residing on or using designated Village Land have the right to obtain formal documentation of their use rights in the form of a Certificate of Customary Right of Occupancy (CCRO). However, insufficient capacity of district land offices that issue CCROs, a lack of funds to pay associated fees, unfamiliarity with formal land laws and other factors have resulted in few villagers obtaining formal documentation for their plots. Furthermore, many villages have not yet completed the village land use management plans that are a prerequisite for CCRO issuance.

The Government of Tanzania and the donor community recognize that improving the security of land rights is essential to protecting the rights of smallholders, reducing disputes and tensions and maximizing the economic potential of the region. The Government, through various programs, often sponsored by the donor community, has made efforts to speed up village land demarcation, village land use planning and village land certification.

**Land Tenure Programs** A pilot Village Certification project was implemented in Mbozi District from 1999 as an effort to implement Village Land Act. By 2007 village boundaries of all 175 villages in Mbozi had been surveyed and 158 had been issued with Certificates of Village Land, and total of 1,117 CCROs have been issued. This experience was replicated in 10 Districts: Iringa (40 villages); Handeni (6 villages); Kilindi (10 villages); Babati (5 villages); Monduli (49 villages); Kiteto (6 villages); Kilolo (9 villages); Namtumbo; Ngorongoro (1 village); Muleba (2 villages). Countrywide, by 2016, around 400,000 CCROs have been issued in various villages and in the years 2014-15 around 49.2 billion shillings had been issued as loans by financial institutions, using CCROs as collateral URT (2016).

Another example of program that aims to improve situation with land property rights in Tanzania, is Feed the Future Tanzania Land Tenure Assistance (LTA) project. The U.S. Agency for International Development project works with 41 communities in central Tanzania to register land and issue Certificates of Customary Right of Occupancy to individual landholders, with a focus on increasing women's inclusion in property ownership. LTA has worked with villages to demarcate and digitally map and record almost 63,000 parcels. These previously undocumented parcels are now registered in the country's official land registry system, providing secure property tenure to 21,000 Tanzanians. The project is also working with local banks to encourage the acceptance of certificates as collateral and with villages to raise awareness of the new loan opportunities. Farmers have already begun using their land-backed loans to purchase fertilizer, high-quality seeds, tractors, and other agricultural inputs to raise their productivity and their incomes.

## A.3 Computational Algorithm

**Steady State** The solution algorithm starts with guessing steady state level of prices, $w$, $r^k$, $r^l$, $\eta$. Given the prices, solve the policy function for each set of state variables using value function iteration. The process yields the optimal occupational choice and policy functions for level of assets, consumption, capital, labor and land inputs. Obtain the stationary distribution of households by finding fixed point using forward iteration. Given the distribution and policy functions, obtain aggregate variables and use them to check whether market clearing conditions for the labor market, capital market, and land market are satisfied. Update the guess for prices and repeat until all market clears.

**Transition** First, compute the initial and final steady states. Then, choose a length $T$ for the transition, and guess a path for prices $\{w, r^k, r^l\}_{t=1}^T$. Solve the household problem along the transition path using backward induction: (a) taking value function in the final steady state, $V_{ssf}$, the market clearing prices as given, solve for household value functions and optimal occupational choice and policy functions for level of assets, consumption, capital, labor and land inputs; (b) repeat this process until solving back to the first period. Given the distribution and policy functions, obtain aggregate variables and use them to check whether market clearing conditions for the labor market, capital market, and land market are satisfied for each period along the transition path. Update the guess for prices and repeat until all market clears for all periods. Check whether $T$ is large enough by trying a larger $T$ and see if the equilibrium path is robust.

## A.4 Proofs of Propositions

**Proposition 1.** *Denote optimal choices of land used by farmers who owns land under communal and private property right regimes as $l_c^*$ and $l_p^*$, respectively. Then, if optimal land usage is larger than household land holding, $l_p^* > l_p$, and farmers' initial conditions in private and communal part of the economy are the same (i.e. same amount of land, skills and assets), we get:*

$$l_c^* \leq l_p^*$$

*Proof:* Let households living under communal and private property rights regime have the same amount of land holdings, have the same productive skills in each sector, and amount of assets. Conditional on farming, also assume that optimal land usage for household in private part of the economy be larger than household land holding, $l_p^* > l_p$. Let $\mu$ be the Lagrange multiplier on collateral constraint (with $\mu_c$ and $\mu_p$ for communal and private part of the economy, respectively). Then, optimal amount of capital used by the farmer is

$$k^* = \left(\exp\left(z_a\right)\left(\frac{\gamma_a}{r^l}\right)^{\gamma_a}\left(\frac{\alpha}{r^k+\mu}\right)^{1-\gamma_a}\right)^{\frac{1}{1-\alpha_a-\gamma_a}}$$

and

$$l^* = \left(\frac{\gamma_a \exp\left(z_a\right) k^{*\alpha_a}}{r^l}\right)^{\frac{1}{1-\gamma_a}}$$

then if $\mu_c = \mu_p = 0$, then $k_p^* = k_c^*$ and $l_p^* = l_c^*$.

If, $\mu_c > 0$ and $\mu_p > 0$, then $k_p^* \geq k_c^*$ and $l_p^* \geq l_c^*$ as $(\lambda_k - 1)q^l l \geq 0$. Moreover, for positive values of land holdings there would occur situation, when $\mu_c > 0$ and $\mu_p = 0$.

*and for assets holdings $a_{small} < a_{large}$, given everything else the same, the following true*

$$l_p^*(a_{small}) - l_c^*(a_{small}) \geq l_p^*(a_{large}) - l_c^*(a_{large}),$$

*Proof:* Fix $a_{small}$ and $a_{large}$, and let households with $a_{small}$ and $a_{large}$ differ only in the amount of assets while all other state variables being the same. Also, let $a_c^*$ and $a_p^*$ denote minimum levels of assets when collateral constraint binds, i.e. $\mu_c > 0$ and $\mu_p > 0$, in case of communal and private land holders, respectively. Then, $a_p^* \leq a_c^*$ as $(\lambda_k - 1)q^l l \geq 0$, and following cases are possible:

i) If $a_{small} \leq a_{large} \leq a_p^* \leq a_c^*$, then both when assets small or large collateral constraint binds. Therefore,

$$l_c^* = \left( \frac{\gamma_a \exp{(z_a)} (\lambda_k a)^{\alpha_a}}{r^l} \right)^{\frac{1}{1-\gamma_a}}$$

and

$$l_p^* = \left( \frac{\gamma_a \exp{(z_a)} (\lambda_k a + (\lambda_k - 1) q^l l_p)^{\alpha_a}}{r^l} \right)^{\frac{1}{1-\gamma_a}}$$

Then

$$l_p^*(a_{small}) - l_c^*(a_{small}) \geq l_p^*(a_{large}) - l_c^*(a_{large}) \Leftrightarrow$$

$$(\lambda_k a_{small} + (\lambda_k - 1) q^l l_p)^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{small})^{\frac{\alpha_a}{1-\gamma_a}} \geq (\lambda_k a_{large} + (\lambda_k - 1) q^l l_p)^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{large})^{\frac{\alpha_a}{1-\gamma_a}}$$

The inequality is true, given that function $f(x) = x^{\frac{\alpha_a}{1-\gamma_a}}$ is concave downward (as $f''(x) = \frac{\alpha_a(\alpha_a+\gamma_a-1)}{(1-\gamma_a)^2} x^{\frac{\alpha_a+2\gamma_a-2}{1-\gamma_a}} < 0$ for production function with decreasing return of scale), and $(\lambda_k - 1)q^l l \geq 0$

ii) If $a_{small} \leq a_p^* \leq a_{large} \leq a_c^*$, then both when assets small or large collateral constraint binds for household living in communal part, while for private part collateral constraint binds only for households with $a_{small}$. Then, the optimal level of capital for households with $a_{large}$ is

$$k_p^*(a) \leq \lambda_k a_{large} + (\lambda_k - 1) l_p$$

and, hence,

$$l_p^*(a_{small}) - l_c^*(a_{small}) \geq l_p^*(a_{large}) - l_c^*(a_{large}) \Leftrightarrow$$

$$(\lambda_k a_{small} + (\lambda_k - 1)q^l l_p)^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{small})^{\frac{\alpha_a}{1-\gamma_a}} \geq$$

$$\geq (\lambda_k a_{large} + (\lambda_k - 1)q^l l_p)^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{large})^{\frac{\alpha_a}{1-\gamma_a}} \geq$$

$$\geq (k_p^*(a))^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{large})^{\frac{\alpha_a}{1-\gamma_a}}$$

iii) If $a_{small} \leq a_p^* \leq a_c^* \leq a_{large}$ then when assets are small collateral constraint binds for all household, while for $a_{large}$ households using the optimal level of capital and land both in communal and private parts of the economy. Hence, $l_p^*(a_{large}) - l_c^*(a_{large}) = 0$ and we have that

$$(\lambda_k a_{small} + (\lambda_k - 1)q^l l_p)^{\frac{\alpha_a}{1-\gamma_a}} - (\lambda_k a_{small})^{\frac{\alpha_a}{1-\gamma_a}} \geq 0$$

iv) If $a_p^* \leq a_{small} \leq a_c^* \leq a_{large}$ is equivalent to iii) with $l_p^*(a_{large}) - l_c^*(a_{large}) = 0$.

v) If $a_p^* \leq a_{small} \leq a_{large} \leq a_c^*$ then households living in private part of the economy use the same amount of land – efficient, and, therefore,

$$l_p^*(a_{small}) - l_c^*(a_{small}) \geq l_p^*(a_{large}) - l_c^*(a_{large}) \Leftrightarrow$$

$$-(\lambda_k a_{small})^{\frac{\alpha_a}{1-\gamma_a}} \geq -(\lambda_k a_{large})^{\frac{\alpha_a}{1-\gamma_a}} \Leftrightarrow$$

$$a_{small} \leq a_{large}$$

vi) Finally, if $a_p^* \leq a_c^* \leq a_{small} \leq a_{large}$ none collateral constraint binding and all households use the same efficient amount of land, and

$$l_p^*(a_{small}) - l_c^*(a_{small}) = 0 \geq l_p^*(a_{large}) - l_c^*(a_{large}) = 0$$

*and for the levels of agricultural productivity $z_{small} < z_{large}$, given everything else the same*

$$l_p^*(z_{small}) - l_c^*(z_{small}) \leq l_p^*(z_{large}) - l_c^*(z_{large}),$$

*Proof:* Fix $z_{small}$ and $z_{large}$, and let households with $z_{small}$ and $z_{large}$ differ only in the level of their agricultural productivity while all other state variables being the same. Also, let $k_c^*$ and $k_p^*$ denote minimum levels of capital when collateral constraint binds, i.e. $\mu_c > 0$ and $\mu_p > 0$, in case of communal and private land holders, respectively. Also, denote $k_{small}^*$ and $k_{large}^*$ to be optimal level of capital used by households with agricultural productivity $z_{small}$ and $z_{large}$, respectively. Then, following the same six cases, but with level of capital as in previous part, analogous steps provide proof of proposition.

*and for the levels of land holdings $l_{small} < l_{large}$, given everything else the same, we get*

$$l_p^*(l_{small}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_c^*(l_{large}).$$

*Proof:* Fix $l_{small}$ and $l_{large}$, and let households with $l_{small}$ and $l_{large}$ differ only in the level of their land holding while all other state variables being the same. Given that households only differ in the level of land holdings, then optimal levels of capital and land would be same for all households, $k^*$ and $l^*$:

$$k^* = \left( \exp\left(z_a\right) \left(\frac{\gamma_a}{r^l}\right)^{\gamma_a} \left(\frac{\alpha}{r^k + \mu}\right)^{1-\gamma_a} \right)^{\frac{1}{1-\alpha_a-\gamma_a}}$$

and

$$l^* = \left( \frac{\gamma_a \exp\left(z_a\right) k^{*\alpha_a}}{r^l} \right)^{\frac{1}{1-\gamma_a}}$$

Hence, household would deviate from optimal levels only when collateral constraint for some of them binds. This leads to the following cases:

i) If no constraints binds, then $l_p^*(l_{small}) - l_c^*(l_{small}) = 0 \leq l_p^*(l_{large}) - l_c^*(l_{large}) = 0$

ii) If collateral constraint binds only for those in the communal part of the economy, then

$l_c^*(l_{small}) = l_c^*(l_{large}) = \lambda_k a$ and $l_p^*(l_{small}) = l_p^*(l_{large}) = l^*$, hence

$$l_p^*(l_{small}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_c^*(l_{large}) \Leftrightarrow$$

$$l_c^*(l_{large}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_p^*(l_{small}) \Leftrightarrow 0 = 0$$

iii) If collateral constraint binds for households living in private part with $l_{small}$ and not $l_{large}$,[1] then it also binds for all households in communal part as $k^* \geq \lambda_k a + (\lambda_k - 1)l_{small} \geq \lambda_k a$. Then,

$$l_p^*(l_{small}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_c^*(l_{large}) \Leftrightarrow$$

$$l_c^*(l_{large}) - l_c^*(l_{small}) \leq l_p^*(l_{large}) - l_p^*(l_{small})$$

with $l_c^*(l_{small}) = l_c^*(l_{large}) = \lambda_k a$ we get

$$l_p^*(l_{large}) - l_p^*(l_{small}) = l_p^*(k^*) - l_p^*(k = \lambda_k a + (\lambda_k - 1)l_{small}) \geq 0$$

as $k^* > \lambda_k a + (\lambda_k - 1)l_{small}$ and land is strictly increasing in capital.

iv) If all constraints bind, then again $l_c^*(l_{small}) = l_c^*(l_{large}) = \lambda_k a$, and,

$$l_p^*(l_{large}) - l_p^*(l_{small}) = l_p^*(\lambda_k a + (\lambda_k - 1)l_{large}) - l_p^*(k = \lambda_k a + (\lambda_k - 1)l_{small}) \geq 0.$$

as $\lambda_k a + (\lambda_k - 1)l_{large} > \lambda_k a + (\lambda_k - 1)l_{small}$ and land is strictly increasing in capital.

**Proposition 2.** *Denote optimal choices of land used by farmers who owns land under communal and private property right regimes as $l_c^*$ and $l_p^*$, respectively. Then, if optimal*

[1] The opposite could not be true as $k^* \geq \lambda_k a + (\lambda_k - 1)l_{large}$ implies that $k^* \geq \lambda_k a + (\lambda_k - 1)l_{small}$

*land usage is lower than household land holding,* $l_p^* < l_p$*, and farmers' initial conditions in private and communal part of the economy are the same (i.e. same amount of land, skills and assets):*

$$l_c^* \geq l_p^*$$

*Proof:* Let households living under communal and private property rights regime have the same amount of land holdings, have the same productive skills in each sector, and amount of assets. Conditional on farming, also assume that optimal land usage for household in private part of the economy be smaller than household land holding, $l_p^* < l_p$. Then, given that households in communal part of the economy could not rent out their land and agricultural production function is increasing in land, households in communal part would use all their land for farming, $l_c^* = l_c$. Hence,

$$l_c^* = l_c = l_p > l_p^* \Leftrightarrow l_c^* \geq l_p^*$$

*and for the levels of agricultural productivity* $z_{small} < z_{large}$*, given everything else the same*

$$l_c^*(z_{small}) - l_p^*(z_{small}) \geq l_c^*(z_{large}) - l_p^*(z_{large})$$

*Proof:* Again, given that households in communal part are going to use all land holding, $l_c^*(z_{small}) = l_c^*(z_{large}) = l_c$, hence,

$$l_c^*(z_{small}) - l_p^*(z_{small}) \geq l_c^*(z_{large}) - l_p^*(z_{large}) \Leftrightarrow$$

$$l_p^*(z_{small}) \leq l_p^*(z_{large})$$

which holds, as $l^*$ is increasing in both $z_a$ and $k^*$, that is also is increasing in $z_a$.

*and for the levels of land holdings $l_{small} < l_{large}$, given everything else the same, we get*

$$l_c^*(l_{small}) - l_p^*(l_{small}) \leq l_c^*(l_{large}) - l_p^*(l_{large})$$

*Proof:* Following the above,

$$l_c^*(l_{small}) - l_p^*(l_{small}) \leq l_c^*(l_{large}) - l_p^*(l_{large}) \Leftrightarrow$$

$$l_p^*(l_{small}) \leq l_p^*(l_{large})$$

With $l^*$ increasing in $k^*$, when

i) collateral constraints not binding in neither cases, $l_p^*(l_{small}) = l_p^*(l_{large}) = l^*$.

ii) collateral constraint binding for $l_{small}$ and not for $l_{large}$,[2] we have

$$l_p^*(l_{large}) - l_p^*(l_{small}) = l_p^*(k^*) - l_p^*(k = \lambda_k a + (\lambda_k - 1)l_{small}) \geq 0.$$

as $k^* > \lambda_k a + (\lambda_k - 1)l_{small}$ and land is strictly increasing in capital.

iii) collateral constraint binds for both $l_{large}$ and $l_{small}$, then again $l_c^*(l_{small}) = l_c^*(l_{large}) = l * c$, and,

$$l_p^*(l_{large}) - l_p^*(l_{small}) = l_p^*(\lambda_k a + (\lambda_k - 1)l_{large}) - l_p^*(k = \lambda_k a + (\lambda_k - 1)l_{small}) \geq 0.$$

as $\lambda_k a + (\lambda_k - 1)l_{large} > \lambda_k a + (\lambda_k - 1)l_{small}$ and land is strictly increasing in capital.

[2]The opposite could not be true as $k^* \geq \lambda_k a + (\lambda_k - 1)l_{large}$ implies that $k^* \geq \lambda_k a + (\lambda_k - 1)l_{small}$

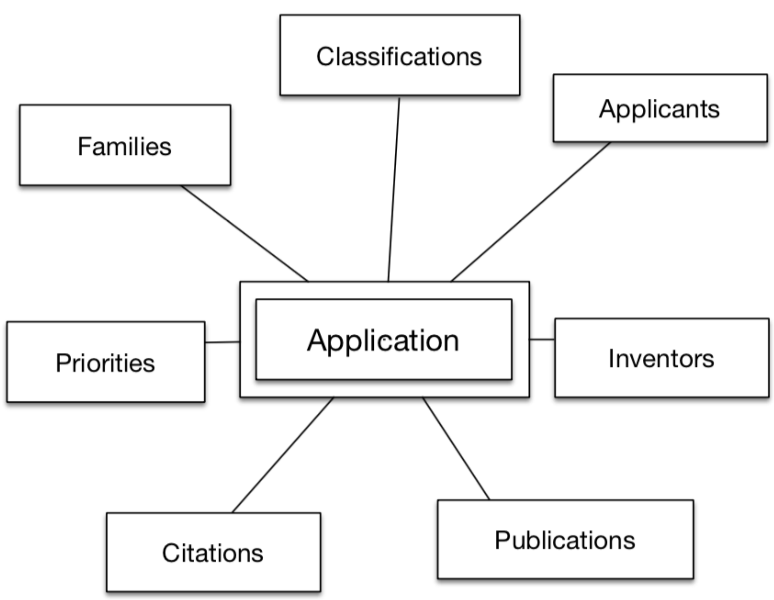
# Appendix B

# Appendix to Chapter Two

## B.1 Data and Construction of Fields of Knowledge

As shown in the Figure B.1, PATSTAT contains information regarding applications, publications, applicants, inventors, citations, patent families, technological categories, priorities, and so on. In this section we explain in details where and how all the data used in our analysis are obtained.

Figure B.1: Structure of the Patent Data in PATSTAT



### B.1.1 Further Details on PATSTAT and Patent Data Construction

**Country** As a default, the country or territory assigned to the patent is the country code of the respective inventor(s) of the patent. If the information about the inventor or the inventor's country is not available, then the country code of applicant(s) is used. Lastly, if both inventor(s) and applicant(s) information, or their country code are not available in the dataset, then the publication authority of the patent is assigned to the patent. In the first two cases, when multiple inventors or applicants are associated with a given patent and they are not from the same country or territory, then countries' weights are assigned to the patents. The weight is computed assuming that each inventor or applicant contributed equally. For example, when patent has four inventors, where one inventor is from the US and three are from the UK, then countries associated with this patent would be the US and the UK with weights 0.25 and 0.75, respectively.

Overall, the dataset contains over 110 millions patents for which country data is available. Around 40 percent of country data comes from information about inventors, less than 5 percent from information about applicants, and the rest comes from the publication authority of a given patent. Around half of the patents in the latter category have Japan as publication authority, since the data for inventors and applicants from Japan are not available in the dataset. The number of patents associated with each country or patent office is provided in the Table B.12.

In order to avoid having any breaks in our time series patent data due to geopolitical changes, the following modifications from the original data are made:

- German Democratic Republic (DD) was incorporated into the entry Germany (DE)
- The Democratic Yemen (YD) was incorporated into the entry Yemen (YE)

- For consistent analysis over time the following synthetic countries were created:
  - Czechoslovakia (CS) by merging data for Czech Republic (CZ) and Slovakia (SK) for the data starting on 1 January 1993.
  - Yugoslavia (YU) by merging data for Bosnia and Herzegovina (BA), Croatia (HR), Macedonia (MK), Montenegro (ME) , Serbia (RS), and Slovenia (SI) for the data since independence of each listed country (Kosovo is not a part of this merge due to absence of any patent data).
  - Union of Soviet Socialist Republics (SU) by merging data for Armenia (AM), Azerbaijan (AZ), Belarus (BY), Estonia (EE), Georgia (GE), Kazakhstan (KZ), Kyrgyzstan (KG), Latvia (LV), Lithuania (LT), Moldova (MD), Russia (RU), Tajikistan (TJ), Turkmenistan (TM), Ukraine (UA), Uzbekistan (UZ) for the data starting 26 December 1991.

We exclude China from our sample due substantial rise in number of Chinese patents since the 3rd revision of Patent law in China in 2008 (Figure B.2), which was not accompanied by similar increase in number of high quality patents. While we observe sharp increase in total number of patents, the same patterns are not observed in number of Triadic patents, which measures patents filed jointly in largest patent offices (Figure B.3).

Triadic patent families are patents filed jointly in the largest global technology markets: the Japan Patent Office (JPO), the U.S. Patent and Trademark Office (USPTO), and the European Patent Office (EPO)) is a standard measure of high-quality patents.

**Date** For the majority of patents, the date in which the application was physically received at the Patent Authority is assigned. When information about the application date is not available, the date on which the respective publication was made available to the public is used. Finally, if both of these dates are not available for a patent, then the date of the earliest

Figure B.2: Total Number of Chinese Patents

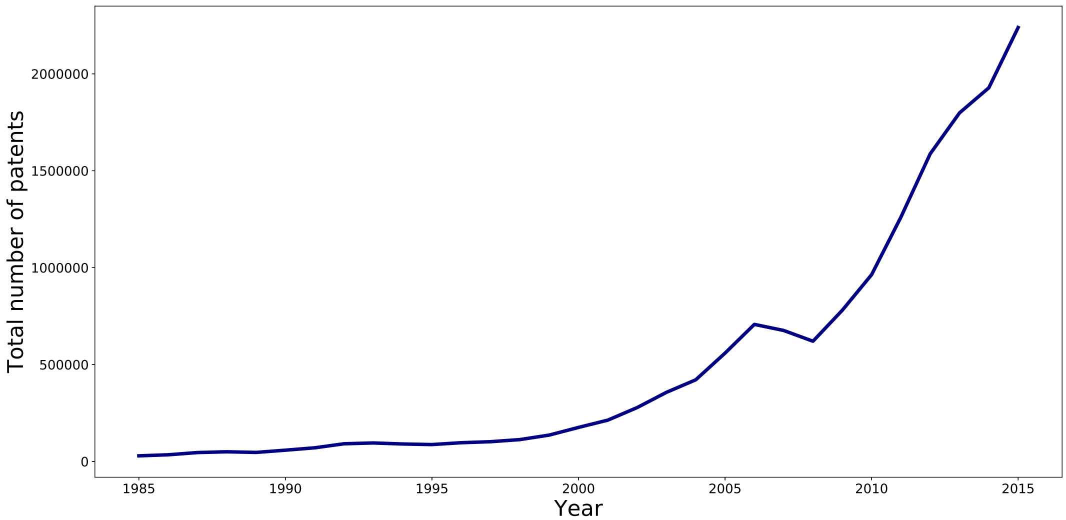
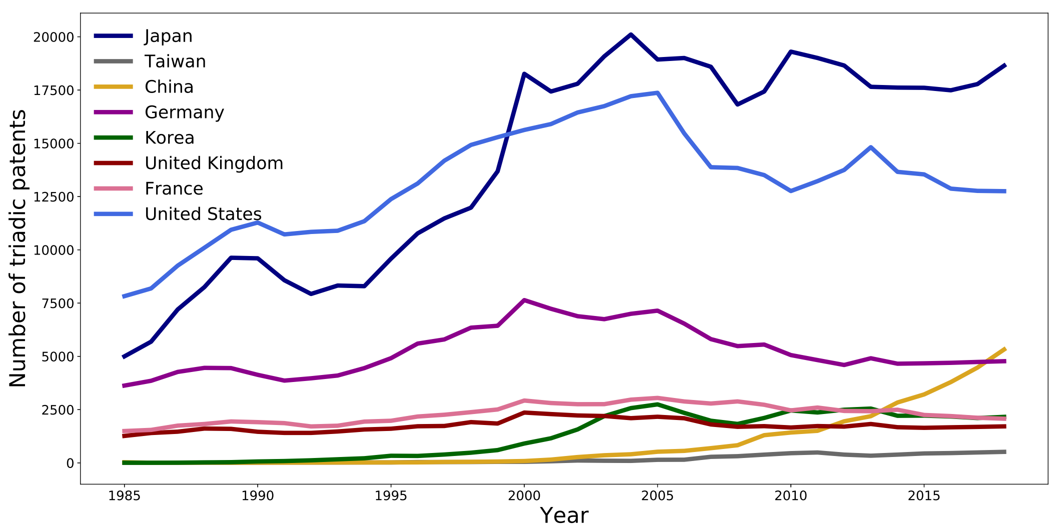



Figure B.3: Number of Triadic Patents per Country




*Data Source:* Organization for Economic Cooperation and Development (OECD).

publication of an application is used. Importantly, in this latter case, earlier applications of the same patent family are not considered.

**Technology Classification** The technology information is provided by the allocation of all patents by the PATSTAT into different groups according to International Patent Classification (IPC). The IPC is a four level hierarchical classification system. and the first version of the IPC system entered into force in 1975, after the Strasbourg agreement (1971). It comprises eight sections (indicated by a letter), followed by two digits indicating the class and a letter for subclass. The subclass is followed by one to three digits (group number) and two more digits separated by a backslash (subgroup). The most recent version of this classification identifies 131 classes, 646 subclasses, 7523 main groups, and 68,899 subgroups. The main sections include (A) Human Necessities, (B) Performing Operations, Transporting, (C) Chemistry, Metallurgy, (D) Textiles, Paper, (E) Fixed Constructions, (F) Mechanical Engineering, Lighting, Heating, Weapons, Blasting (G) Physics, and (H) Electricity.

Each patent is assigned one or more IPC subclasses, and in the latter case respective weights are given according to the relevance of each IPC subclass for a particular patent with the sum of weights being equal to one for each patent. The relevance of each IPC subclass is being determined assuming that each IPC main group at the lowest fourth level of the classification assigned to the patent in the database, is equally important (i.e., we use a uniform distribution).

**Industry Classification** In order to analyze the relationship between patenting and economic activity, a mapping of the IPC subclasses to the NACE/ISIC codes is needed.

Part of the economic data that are used in the paper is classified according to NACE2 (Statistical Classification of Economic Activities in the European Community, Rev. 2) or to ISIC4 (International Standard Industrial Classification of All Economic Activities, Rev. 4). Since NACE is a derived classification of ISIC, correspondence between the two is straightforward: categories at all levels of NACE are defined either to be identical to, or form subsets of, single ISIC categories.

Patents are assigned industry codes (NACE2) using the correspondence table that is provided by EUROSTAT in co-operation with KU Leven / Belgium.[1]. Additionally, in order to obtain industry codes in accordance with earlier version of NACE (Rev. 1.1) or ISIC (Rev. 3), EUROSTAT correspondence table is used.[2]

**Family** Every patent application belongs to exactly one DOCDB family. In the trivial case, an application belongs to a DOCDB family which consists of no other family members except this application itself. Generally speaking, if two applications claim exactly the same prior applications as priorities, then they are defined by the EPO as belonging to the same DOCDB simple family. DOCDB family members generally refer to the same invention. Overall, in our dataset 64.2 percent of all patent applications are those which have only one member in the family.

In order to obtain the sequence for each application, and then patent, within the same family we use the date the application was filed as an indicator. The applications that were filed earlier are considered to have lower sequence number. In case the application filing date is the same for multiple applications within the family, the earliest publication date was used as a second criterion. In order to obtain the first patent within the first application in the family, we use the patent with earliest publication date within application.

### B.1.2 World Input-Output Database

The World Input-Output Tables and underlying data associated to them cover 43 countries and the Rest of the World for the period 2000-2014. Data for 56 sectors are classified according to the International Standard Classification revision 4 (ISIC). We use an aggregation

[1]The data and the methodology to create them is described in `https://ec.europa.eu/eurostat/ramon/documents/IPC_NACE2_Version2_0_20150630.pdf`

[2]`https://ec.europa.eu/eurostat/ramon/relations/index.cfm?TargetUrl=LST_REL&StrLanguageCode=EN&IntCurrentPage=10`

scheme to have consistence with our patent data. This aggregation is characterized by 40 regions of which 36 are countries, then analogously to our patent data we have a synthetic Soviet Union, Yugoslavia, and Czechoslovakia, and the Rest of the World throughout the period.

The World Input-Output Database covers 56 sectors. Using World Input-Output Tables (WIOT) for each pair of countries, sector, and year, we construct trade flows, gross output, intermediate purchases, and value added expressed in US dollars as follows. WIOT contains information about intermediate consumption,$X_{ij,ks,t}$, of goods from sector $k$ and country $i$ by sector $s$ in country $j$ in period $t$. Final purchases split into final consumption expenditure by households, by non-profit organizations serving households, and by government, the sum of which we denote as $X_{ij,kC,t}$, as well as gross fixed capital formation and changes in inventories and valuables, denoted as $X_{ij,kI,t}$.

The Socio-economic accounts (SEA) are a part of WIOD and contain industry-level data on employment, capital stocks, gross output and value added at current and constant prices. The industry classification is the same as for the WIOT. Nominal values in the SEA are denoted in millions of national currency. In order to be consistent with other output and trade data nominal values are converted into the US dollar using exchange rates that has been used to create the WIOT (and is provided in the WIOTs). To obtain output data in constant prices for synthetic countries in a base year, the sum of dollar value for each member of synthetic country for year 2010 is used. Output values in constant prices for other years are obtained by summing together all countries included in a given synthetic country dollar value in a base year multiplied by a respective volume index.

A set of additional statistics that is used in our analysis is computed from those tables. First, for each pair of countries, year and sector, we calculate trade flows. Specifically, sector $k$ exports from country $i$ to country $j$ in a year $t$ as $X_{ij,k,t} = \sum_s X_{ij,ks,t} + X_{ij,kC,t} + X_{ij,kI,t}$. Intermediate consumption and intermediate imports for a country $j$ and sector $s$ in a year

$t$ as $X_{js,IC,t} = \sum_k \sum_i X_{ij,ks,t}$ and $X_{js,II,t} = \sum_{k \neq j} \sum_i X_{ij,ks,t}$, respectively; gross output of sector $k$ in country $i$ in a year $t$ as $X_{ik,GO,t} = \sum_s \sum_j X_{ij,ks,t} + \sum_j X_{ij,kC,t} + \sum_j X_{ij,kI,t}$; and value added as a difference between gross output and intermediate consumption for a given sector and country.

Second, to obtain measures of TFP growth at a country-sector level we use "dual" and "primal" approaches as in Hsieh (1999) and Hsieh (2002). "Dual" TFP growth measure for sector $k$ in country $i$ in a year $t$ is computed as

$$\Delta \ln TFP = s_K \hat{r} + s_L \hat{w}$$

where $s_K$ and $s_L$ are factor shares of capital and labor inputs, respectively, and $\hat{w}$ and $\hat{r}$ are growth rates of wage and capital rental rate, respectively. All the variables are computed for a given period of time and at the country-sector level. To smooth fluctuations in factor shares, we use average in periods $t+1$ and $t$.

"Primal" TFP growth is computed as

$$\Delta \ln TFP = \hat{Y} - s_K \hat{K} - s_L \hat{L}$$

where $\hat{Y}$, $\hat{L}$, $\hat{K}$ are the growth rates of output, labor and capital, respectively.

### B.1.3 Industrial Statistics Database

The UNIDO Industrial Statistics Database (INDSTAT2) contains time series data on the manufacturing sector for the period 1963 onwards for 170 countries. Data are arranged at the 2-digit level of the International Standard Industrial Classification of All Economic Activities (ISIC) Revision 3 pertaining to the manufacturing sector, which comprises 24 industries. However, it should be noted that time period covered by the database, as well as item coverage, differ from country to country.

The database contains eight indicators of industrial statistics, including the index numbers of industrial production, which show the real growth of the volume of production.[3] To be consistent with the rest of data used in the paper, data for members of three synthetic countries were merged together. The exception is index data, which were not obtained due to different time coverage for each member country. Additionally, data for German Democratic Republic and Federal Republic of Germany are merged together for the period prior 1990, and together with the data for Germany starting in 1991 form single time series. Unfortunately, these data does not report capital stock data.

### B.1.4 Construction of Fields of Knowledge and Clustering Algorithm

In this section we describe the clustering algorithm in addition to the details on the distance measure used in this procedure.

To obtain our technology similarity (distance) measure, we first compute the number of times the same inventor has patent activity in each pair of IPC subclasses summed over all inventors. For every inventor $s$, a vector $x_s$ of patent activity for each pair of IPC codes $i, j$ contains elements

$$x_{ij,s} = \begin{cases} 1, \text{ if inventor has patent with both IPC codes } i \text{ and } j \\ 0, \text{ otherwise} \end{cases}$$

We use the dot product of vectors to compute a matrix of the number of times two codes appear in the patent activity of the same inventor for all inventors. Of course, simply comparing the amount of patent activity for each pair of IPC subclasses ignores the fact

[3]The earliest year for which data on the index of industrial production are available is 1981.

that some IPC subclasses are simply appear more frequently. Dividing by the number of different inventors that have a patent in particular technology we get $650 \times 650$ proximity matrix $T$, where element $T_{ij}$ is the probability of the inventor to have a patent with IPC subclass $i$ conditional on having a patent with IPC subclass $j$. The diagonal elements of the matrix, $i = j$, are equal to one. The final matrix does not need to be symmetric either in terms of absolute numbers, or in terms of the strength of obtained links. For example, manufacture of dairy products (A01J) is found to be the closest to dairy product treatment (A23C), while dairy product treatment is the closest to foods, foodstuffs, or non-alcoholic beverages (A23L).

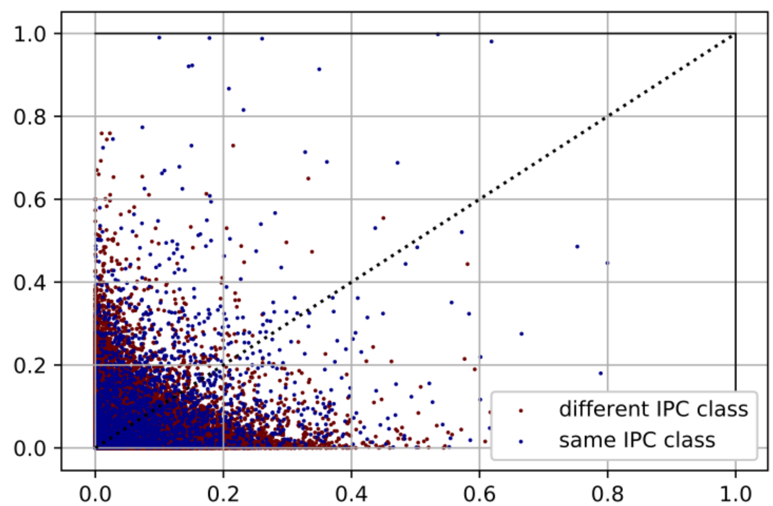
Formally, the inverse measure of distance between IPC codes $i$ and $j$ equals

$$\phi_{ij} = P(x_{ij}|x_{jj})$$

A graphical representation of the matrix is shown on Figure B.4, where each point represents a pair of IPC subclases $(i, j)$, and associated with them conditional probabilities, $\phi_{ij}$, on the axes. The more similar codes to each other the more north-east the point's location, with the most similar being at the point $(1, 1)$. There are several observations worth to be mentioned. First, the distance for most pairs to 45 degree line is not very big, suggesting quite tight relationship between $\phi_{ij}$ and $\phi_{ji}$. Large deviation from 45 degree line often arises when one of two subclasses appears in a small amount of patents, but often accompanied by the second subclasses. Second, codes at subclass level of IPC that are rarely belong to the patent(s) of the same inventor often being grouped into the same IPC classes (pairs of subclass codes having the same IPC class are blue points), suggesting that using IPC classes as definitions of fields of knowledge is not adequate for the analysis.

We use information about the strength of linkages between subclass level codes to perform the clustering analysis and group these subclasses into clusters, where each cluster represents

Figure B.4: Similarity matrix of IPC subclass-level codes




a separate field of knowledge. The aim of clustering analysis is to identify groups of similar objects (IPC subclasses) according to the selected criterion (conditional probability of being assigned to patent(s) which belongs to the same inventor). The clustering analysis is based on a proximity or distance matrix which includes the similarity evaluation for all pairs of objects. In order to obtain a symmetric matrix for the clustering analysis, we use the following dissimilarity measure

$$D_{ij} = 1 - (\phi_{ij} + \phi_{ji}) = D_{ji},$$

The basic approaches are hierarchical clustering and k-means clustering, which have many types. Given the nature of our proximity matrix and quite large set of IPC subclasses, the algorithm that is used in this paper is *k-medoids*. The goal of that algorithm is to minimize distance function by looking at all possible permutations of groupings, conditional on $k$ clusters.

The following steps provide brief description of the algorithm:

- Select $K$ points as the initial representative objects (i.e., as initial *k-medoids*)
- Repeat
  - Assigning each point to the cluster with the closest medoid
  - Randomly select a non-representative object $o_i$
  - Compute the total cost $S$ of swapping the medoid $m$ with $o_i$
    * $S = S_{o_i} - S_m$
    * $S_{o_i} = \sum_{j=1}^{k} s_{j,o_i}$, where $s_{j,o_i} = \sum I \in j D_{I,o_i}$
    * $S_m = \sum_{j=1}^{k} s_{j,m}$, where $s_{j,m} = \sum I \in j D_{I,m}$
    * In words, $S_x$ is the sum of distances to the center within each cluster
  - if $S < 0$, then swap $m$ with $o_i$ to form the new set of medoids
- Until convergence criterion is satisfied

One feature of the *k-medoids* clustering is that number of clusters $k$ should be specified in advance. Hence, the analyst needs to have some tools for determining the number of clusters. In order to decide on the optimal amount of clusters for *k-medoids* algorithm, Silhouette coefficient is used. The idea behind this criterion is to minimize distance between elements within cluster, and to maximize distance across clusters, while punishing for singletons.

Formally, for each point $x_i$, its silhouette coefficient $s_i$ is:

$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

where

- $\mu_{in}(x_i)$ is the mean distance from $x_i$ to points in its own cluster
- $\mu_{out}^{min}(x_i)$ is the mean distance from $x_i$ to points in its closest cluster

And, $s_i = 0$ if $x_i$ belongs to a singleton cluster. Then, the Silhouette coefficient (SC) is the mean values of $s_i$ across all the points: $SC = \frac{1}{n}\sum_{i=1}^{n} s_i$. SC close to +1 implies good clustering: points are close to their own clusters but far from other clusters.

## B.2 Stylized Facts on World Innovation

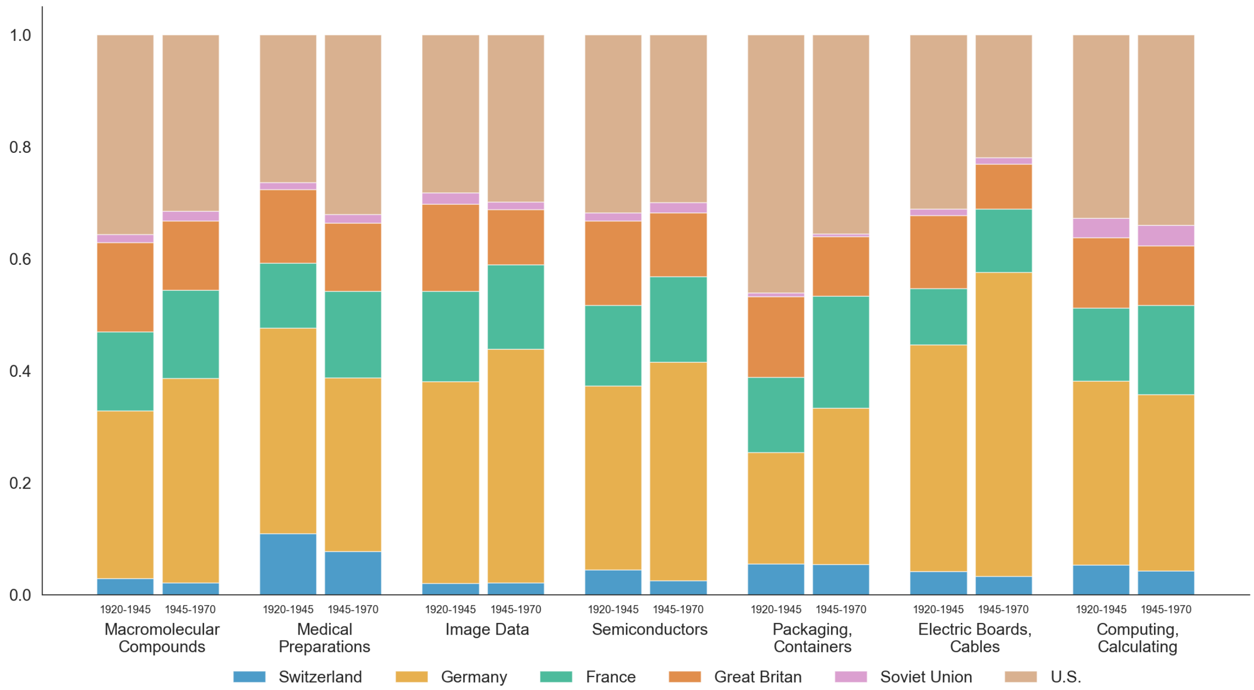
This section contains supplementary tables and figures to Section 2.3 in the main text of the paper, as well as some additional facts on the evolution of world innovation across time and space.

### B.2.1 Supplementary Tables and Figures

Table B.1: Evolution of Top Fields of Knowledge

| | **< 1920** | **20-25** | **25-30** | **30-35** | **35-40** | **40-45** | **45-50** | **50-55** | **55-60** | **60-65** | **65-70** | **70-75** | **75-80** | **80-85** | **85-90** | **90-95** | **95-00** | **2000-05** | **05-10** | **10-15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Diagnosis and Surgery | 3.4 | 2.8 | 2.6 | 2.7 | 2.4 | 2.1 | 3.2 | 3.1 | 3.0 | 3.9 | 5.0 | 5.7 | 6.3 | 7.6 | 9.8 | 10.3 | 7.9 | 7.2 | 6.4 | 6.0 |
| Medical Preparations | 0.8 | 0.6 | 0.8 | 1.0 | 1.3 | 1.6 | 1.9 | 2.5 | 2.9 | 2.9 | 3.1 | 3.3 | 3.7 | 4.2 | 4.7 | 4.9 | 3.7 | 3.6 | 4.0 | 3.9 |
| Liquid and Gaseous Products | 3.0 | 2.7 | 2.8 | 2.8 | 2.4 | 2.3 | 2.2 | 2.4 | 2.5 | 2.6 | 2.9 | 3.2 | 3.0 | 2.6 | 2.4 | 2.0 | 1.6 | 1.5 | 1.5 | 1.3 |
| Packaging, Containers | 4.2 | 4.6 | 4.5 | 5.5 | 4.8 | 3.7 | 4.7 | 4.2 | 4.5 | 5.0 | 4.4 | 3.4 | 2.8 | 2.3 | 2.4 | 1.8 | 1.3 | 1.2 | 1.1 | 1.1 |
| Organic Chemistry | 0.4 | 0.5 | 1.0 | 1.2 | 1.8 | 2.6 | 2.2 | 2.3 | 2.3 | 2.2 | 2.0 | 2.0 | 1.8 | 1.6 | 1.3 | 1.1 | 0.8 | 0.7 | 0.8 | 0.7 |
| Macromolecular Compounds | 1.3 | 1.3 | 1.5 | 2.0 | 2.8 | 3.4 | 3.2 | 4.0 | 4.6 | 4.9 | 4.6 | 4.9 | 4.5 | 4.5 | 3.9 | 3.3 | 2.7 | 2.5 | 2.7 | 2.5 |
| Building Constructions | 2.5 | 2.5 | 2.7 | 2.5 | 2.1 | 1.5 | 1.8 | 1.8 | 2.0 | 2.3 | 2.2 | 2.0 | 1.8 | 1.5 | 1.2 | 1.0 | 0.8 | 0.7 | 0.7 | 0.7 |
| Geothermal Systems | 3.9 | 4.0 | 4.5 | 4.8 | 5.6 | 5.4 | 4.8 | 5.2 | 5.5 | 4.9 | 4.5 | 3.9 | 3.9 | 3.5 | 2.8 | 2.3 | 1.8 | 1.8 | 2.1 | 2.2 |
| Computing, Calculating, Counting | 2.3 | 2.2 | 1.9 | 1.8 | 1.5 | 1.5 | 1.9 | 2.0 | 2.1 | 2.6 | 3.3 | 3.9 | 4.6 | 5.7 | 7.6 | 10.9 | 17.5 | 20.3 | 22.3 | 22.4 |
| Semiconductors and Solid State Devices | 1.1 | 1.6 | 1.8 | 1.8 | 1.8 | 1.9 | 1.9 | 2.3 | 2.7 | 2.9 | 3.4 | 4.0 | 4.9 | 5.8 | 6.6 | 6.9 | 7.8 | 8.5 | 7.9 | 7.4 |
| Electric Boards, Cables, Switches | 1.4 | 1.9 | 2.5 | 2.4 | 2.6 | 2.7 | 2.1 | 2.5 | 2.6 | 2.7 | 2.6 | 2.3 | 2.1 | 1.7 | 1.6 | 1.4 | 1.1 | 1.0 | 1.1 | 1.0 |
| Digital Information | 0.3 | 0.3 | 0.4 | 0.5 | 0.5 | 0.7 | 0.5 | 0.4 | 0.4 | 0.5 | 0.6 | 0.8 | 1.1 | 1.5 | 2.2 | 3.7 | 7.7 | 8.2 | 6.9 | 8.9 |
| Image Data | 0.8 | 0.9 | 1.3 | 1.3 | 1.6 | 1.5 | 1.6 | 1.5 | 1.3 | 1.3 | 1.6 | 2.0 | 2.2 | 3.1 | 3.7 | 4.7 | 5.5 | 4.8 | 4.1 | 4.3 |

Figure B.5: Countries Shares in Top Fields, 1920-1970



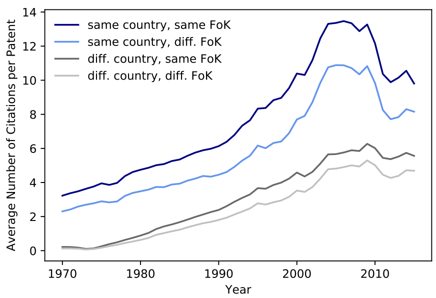

*Notes:* Country acronyms are: US United States, DE Germany, GB Great Britain, FR France, CH Switzerland and SU Soviet Union.

Table B.2: Correlation of Alternative Rankings

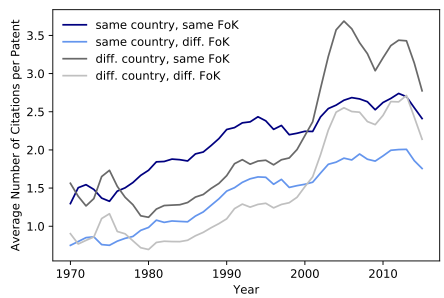
| | All pat. | First in the fam | Ever cit. | Ever cit. (1st fam.) | Weighted by citat. | Weight. cit. (1st fam.) | Average share |
|---|---|---|---|---|---|---|---|
| All patents | 1.00 | | | | | | |
| First in the fam. | 0.994 | 1.00 | | | | | |
| Ever cited | 0.988 | 0.987 | 1.00 | | | | |
| Ever cited (1st in the fam.) | 0.984 | 0.989 | 0.997 | 1.00 | | | |
| Weighted by cit. | 0.961 | 0.958 | 0.985 | 0.981 | 1.00 | | |
| Weighted by cit. (1st in the fam.) | 0.956 | 0.957 | 0.982 | 0.982 | 0.997 | 1.00 | |
| Average share across countries[4] | 0.794 | 0.799 | 0.806 | 0.809 | 0.810 | 0.813 | 1.00 |

### B.2.2 Measuring the Technology Frontier

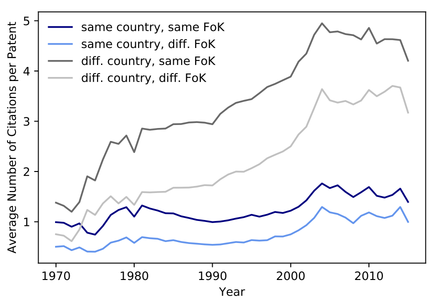
To measure the frontier for a given period of time we compute the rank of each field knowledge based on the its share in the total patent activity across the world. Then, we define country



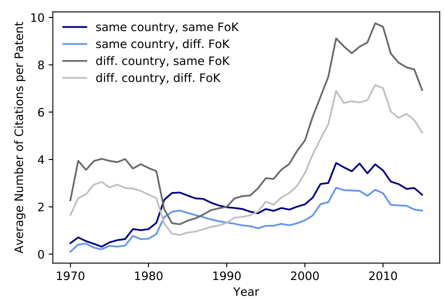

(a) US




(b) Germany




(c) France




(d) Great Britain

Figure B.6: Citation Dynamics, 1970-2015

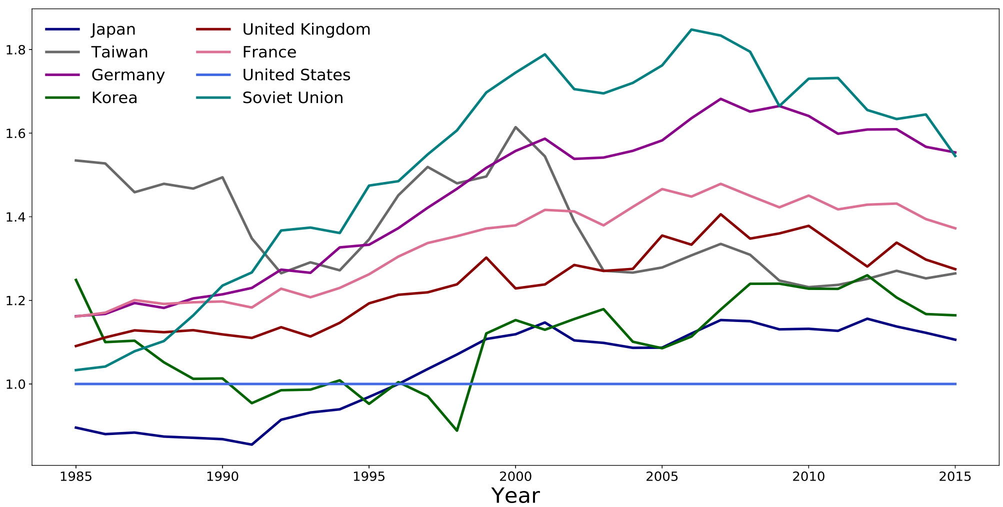
distance to frontier as a co-movement with expanding sectors:

$$d_{ct} \equiv \sum_{s} \chi_{cst} \times \text{Ranking}_{st}$$

where $\chi_{ist}$ is share of patents in country $c$, field of knowledge $s$, time $t$. We perform this analysis for the period of 1985 and onward to get capture the entire universe of patent activity across the world. Figure B.7 displays dynamics of the distance to frontier for "major" innovators for the last thirty years relative to the United States.

From the picture it is easy to notice steady decline of measure in a number of Asian countries closer to knowledge frontier, which means that larger and larger share of innovation activity in those countries happens in the top fields. This trend is very pronounced in

Figure B.7: Distance to the Technology Frontier (Relative to the U.S.)




countries like Korea and Taiwan, while Japan's transition happened earlier in 70s. Overall, Japan and the U.S. experienced similar dynamics: there were both top closest countries to the frontier among major inventors at the beginning of the sample period, and both got even closer for the last 30 years. There is very little change among countries like Great Britain, Germany, and France. Interesting pattern is observed for the USSR countries: it moved further and further from the frontier in the years before and after collapse of the USSR, while a reverse is observed starting in 2000's. The latter is primarily driven by innovation performance of Russia.

### B.2.3 Concentration of Innovation

Next, we look whether there is any observable concentration of innovation across countries within fields of knowledge and within countries but across fields of knowledge. The first one can be interpreted as world concentration of innovations and the latter as comparative advantage. For the purpose of the analysis we compute Theil index, which can be decomposed across partitions in an additive way. Formally Theil index over $(x_1, \ldots, x_N)$:

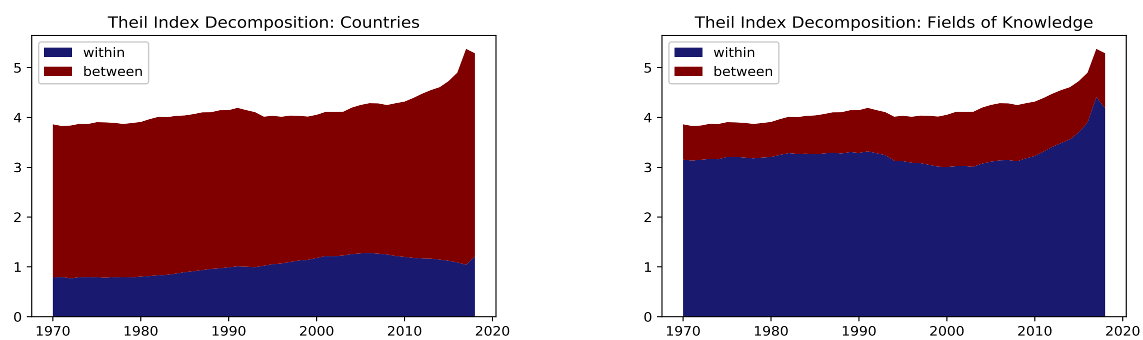
$$T = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\mu} \ln\left(\frac{x_i}{\mu}\right) = \sum_{i=1}^{N} \frac{x_i}{X} \ln\left(\frac{x_i N}{X}\right)$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$. And one can decompose Theil index in the following way:

$$T = \sum_j \left(\frac{X_j}{X}\right) T_j + \sum_j \left(\frac{X_j}{X}\right) \ln\left(\frac{X_j/X}{N_j/N}\right)$$

The first term represents the within-group concentration and the second term represents the between-group inequality, and $T_j$ are individual Theil indexes either computed for fields of knowledge, or for countries.

Figure B.8: Theil Index, 1970-2018, all countries

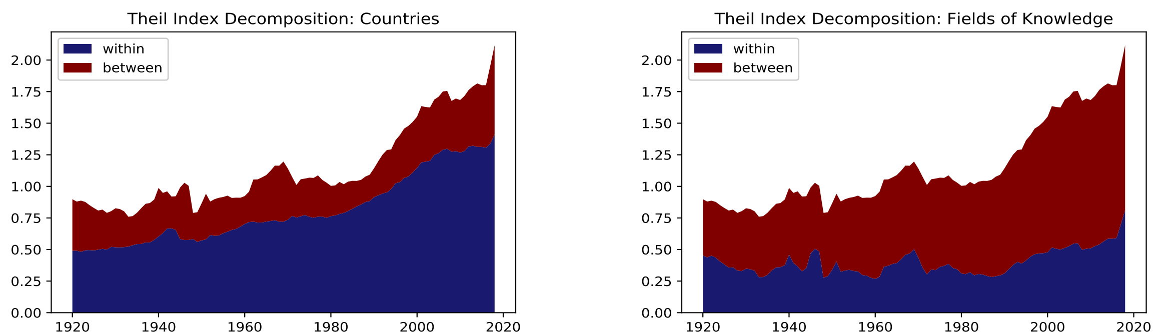



Individual Theil indexes for six countries for the period 1920-2017 are depicted in Figure B.10.

### B.2.4 Geography of Innovation

In this section we are trying to explore whether countries tend to move together towards innovation in certain fields, and if so, what factors are associated with such co-movement. Specifically, for each pair of countries we look at average value across each field of knowledge

Figure B.9: Theil Index, 1920-2018, six countries




for correlations between those two countries of changes in the share of given field across time. Mathematically, we explore the properties of the following expression:

$$\text{corr}_{cc'} = \mathbb{E}_{st}[\text{corr}(\Delta\chi_{cst}, \Delta\chi_{c'st})]$$

We impose some restriction on the sample for which we compute average correlation of changes in knowledge fields shares across time for each pair of countries to ensure that the results are not driven by a selection of certain time periods and/or fields. First, to remove idiosyncratic short term shocks we compute shares change for five years periods instead of for every year (we also perform as a robustness check computations using annual data). Second, for each pair of countries and each field of knowledge correlation is computed only if there is data for at least six common periods. Third, for each pair of countries average correlation is computed only if there is data for at least 50 fields of knowledge. Fourth, we use only patents that are first in the family to avoid co-movement driven by trade relationship between two countries. Finally, we look at period 1970-2015 to ensure broad coverage both of countries and fields of knowledge. Table B.3 displays average correlation between countries for "major" innovators.

Next, we explore what are underlying factors that make countries co-move in their inno-

Figure B.10: Theil Indexes for Individual Countries, 1920-2017

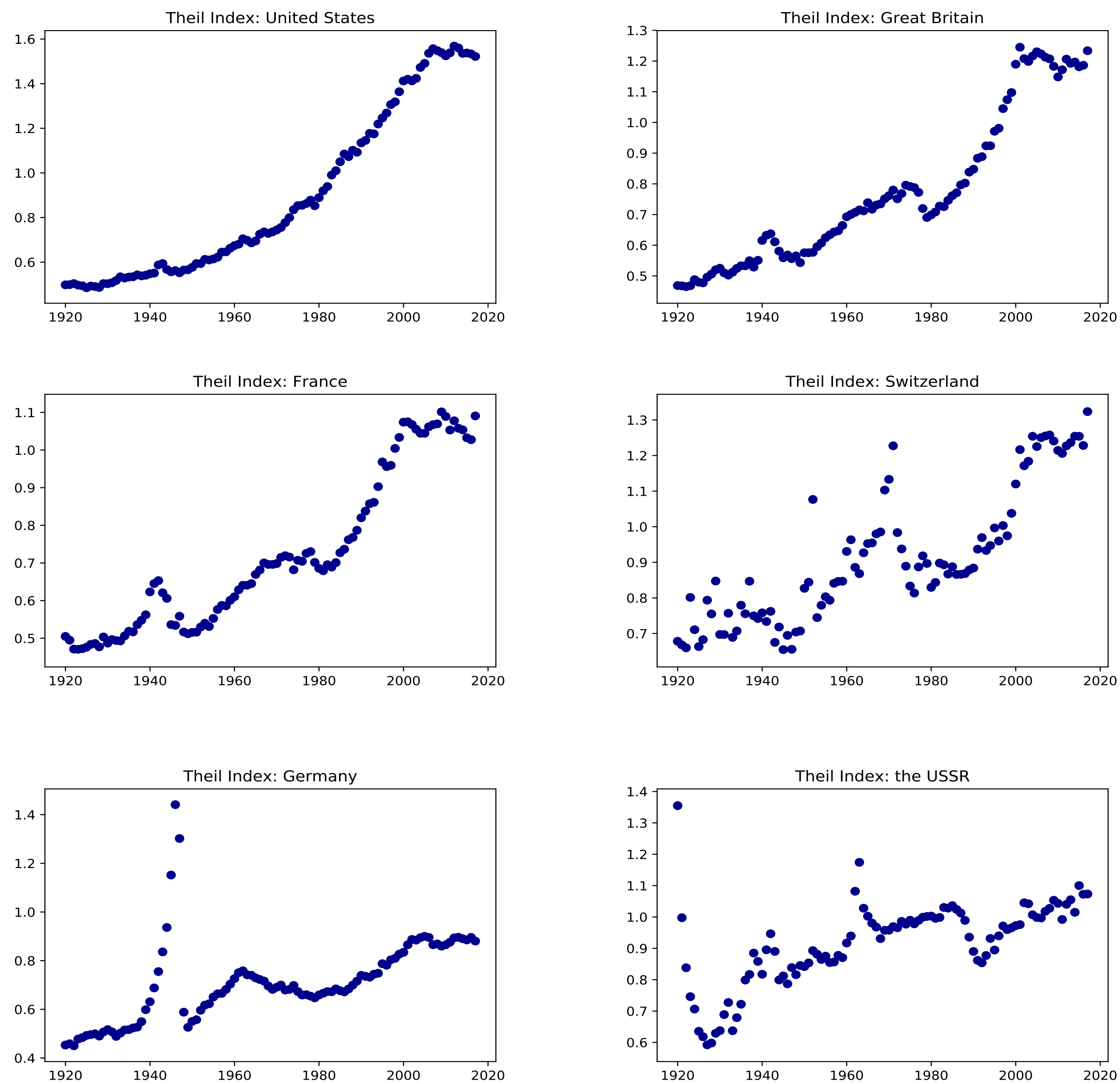


Table B.3: Average Correlation of Changes in Fields Shares Across Time

| | **AT** | **AU** | **BE** | **CA** | **CH** | **CN** | **DE** | **FR** | **GB** | **IT** | **JP** | **KR** | **SU** | **TW** | **US** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AT** | 1.00 | | | | | | | | | | | | | | |
| **AU** | 0.13 | 1.00 | | | | | | | | | | | | | |
| **BE** | -0.01 | 0.09 | 1.00 | | | | | | | | | | | | |
| **CA** | 0.16 | 0.27 | 0.14 | 1.00 | | | | | | | | | | | |
| **CH** | 0.10 | 0.13 | 0.12 | 0.08 | 1.00 | | | | | | | | | | |
| **CN** | 0.13 | 0.06 | -0.04 | 0.08 | -0.02 | 1.00 | | | | | | | | | |
| **DE** | 0.27 | 0.21 | 0.21 | 0.29 | 0.28 | 0.03 | 1.00 | | | | | | | | |
| **FR** | 0.09 | 0.19 | 0.23 | 0.21 | 0.21 | 0.02 | 0.28 | 1.00 | | | | | | | |
| **GB** | 0.08 | 0.23 | 0.16 | 0.25 | 0.14 | 0.09 | 0.31 | 0.28 | 1.00 | | | | | | |
| **IT** | 0.10 | 0.07 | 0.15 | 0.13 | 0.14 | 0.11 | 0.18 | 0.18 | 0.14 | 1.00 | | | | | |
| **JP** | -0.01 | 0.08 | 0.12 | 0.08 | 0.10 | 0.04 | 0.17 | 0.13 | 0.14 | 0.07 | 1.00 | | | | |
| **KR** | -0.03 | 0.00 | 0.08 | -0.05 | -0.04 | 0.05 | 0.07 | 0.04 | 0.08 | -0.07 | 0.08 | 1.00 | | | |
| **SU** | 0.03 | 0.11 | 0.09 | 0.05 | 0.07 | 0.04 | 0.07 | 0.09 | 0.03 | -0.02 | -0.06 | -0.04 | 1.00 | | |
| **TW** | -0.01 | 0.01 | 0.09 | 0.02 | 0.05 | 0.06 | -0.08 | 0.05 | 0.07 | 0.09 | 0.05 | 0.02 | -0.02 | 1.00 | |
| **US** | 0.16 | 0.28 | 0.13 | 0.35 | 0.16 | 0.09 | 0.31 | 0.28 | 0.31 | 0.18 | 0.13 | 0.03 | 0.00 | 0.06 | 1.00 |

vation specialization. Specifically, we look whether such factor as distance, common language or border, and belonging to the same colonial origin, are among the factors that can explain whether countries specialize in a common fields. We run the following regression:

$$\text{corr}_{cc'} = \alpha + \beta_1 \ln \text{Dist}_{cc'} + \beta_2 \mathbb{1}_{lang._{cc'}} + \beta_3 \mathbb{1}_{bord._{cc'}} + \beta_4 \mathbb{1}_{colony_{cc'}} + \varepsilon_{cc'}.$$

We also include country of origin and destination fixed effects. Table B.4 reports our results. As expected, we find distance plays a negative role, while common border and language are positively correlation with knowledge diffusion.

Table B.4: Co-movement in the Shares of Fields of Knowledge and Countries' Characteristics

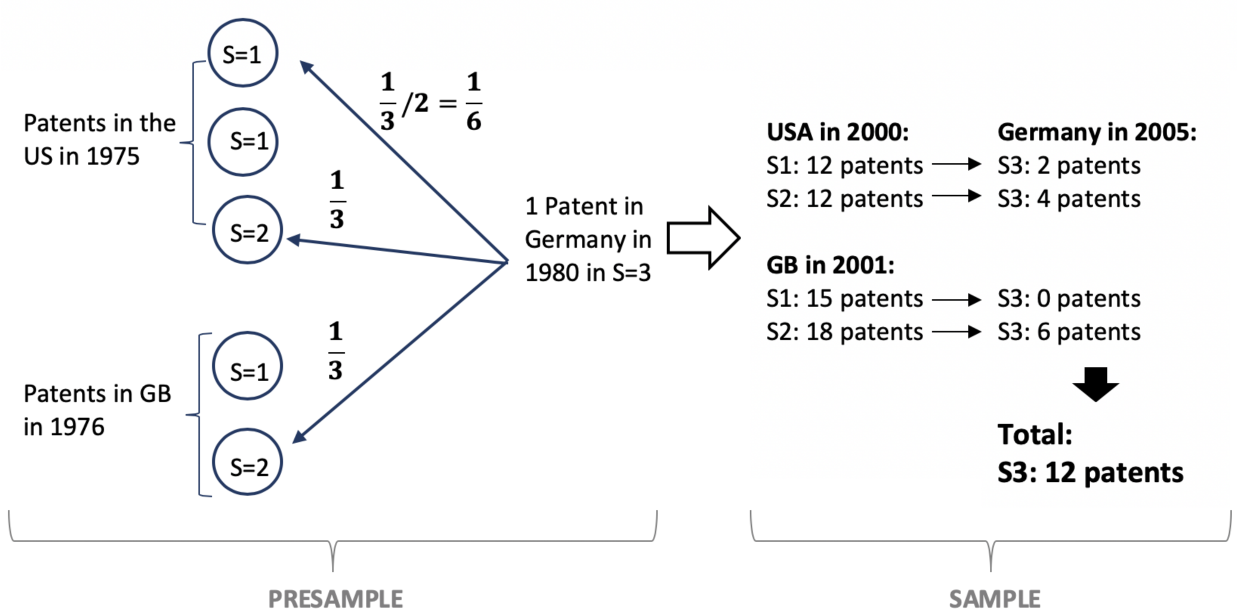
| | Average correlation | | | |
|---|---|---|---|---|
| log dist | -0.018 | -0.013 | -0.010 | -0.010 |
| | (0.003) | (0.004) | (0.004) | (0.004) |
| com border | | 0.033 | 0.026 | 0.025 |
| | | (0.008) | (0.008) | (0.008) |
| com lang | | | 0.035 | 0.036 |
| | | | (0.009) | (0.009) |
| com colony | | | | -0.040 |
| | | | | (0.033) |
| Constant | 0.215 | 0.169 | 0.144 | 0.147 |
| | (0.027) | (0.030) | (0.034) | (0.034) |
| $\text{Country}_1$ FE | Y | Y | Y | Y |
| $\text{Country}_2$ FE | Y | Y | Y | Y |
| # obs. | 737 | 737 | 737 | 737 |
| $R^2$ | 0.51 | 0.52 | 0.54 | 0.54 |

*Notes:* Changes in shares of fields of knowledge are computed for the period 1970-2015. Standard errors are in parentheses clustered at a country of origin level.

## B.3 Empirical Analysis

### B.3.1 Identification

Figure B.11: Construction of Instrument: Example



### B.3.2 Robustness and Other Findings

Table B.5: 2SLS Estimates: $\phi_N = 1$

| | $\Delta \log(va\_em_{cst+n})$ | $n \in \{1, 2, 3\}$ |
|---|---|---|
| | OLS | 2SLS |
| | (1) | (2) |
| $\log(1 + pat_{cst})$ | 0.001 | 0.004 |
| | (0.001) | (0.001) |
| Controls | ✓ | ✓ |
| Country-Year FE | Y | Y |
| Sector-Year FE | Y | Y |
| # obs. | 8,169 | 8,169 |
| | First-stage estimates | |
| Predicted | | 0.470 |
| $\log(1 + pat_{cst})$ | | (0.081) |
| F-stat | | 33.8 |

*Notes:* All regressions include as control (log) values for capital and employment at period $t$ as well as averages for the period $t+n$ with $n \in \{1, ..., 3\}$ consistent with the equation specified in our theoretical framework. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.6: 2SLS Estimates: Robustness, Different Lags

| | $\log(va\_em_{cst+n})$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $\log(1 + pat_{cst})$ | 0.017 | 0.019 | 0.011 | 0.018 | 0.026 |
| | (0.008) | (0.008) | (0.004) | (0.008) | (0.010) |
| Controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| Country-Year FE | Y | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y | Y |
| # obs. | 8,357 | 8,357 | 9,744 | 9,053 | 8,358 |
| | First-stage estimates | | | | |
| Predicted | 0.461 | 0.392 | 0.457 | 0.460 | 0.463 |
| $\log(1 + pat_{cst})$ | (0.079) | (0.070) | (0.077) | (0.079) | (0.079) |
| F-stat | 33.9 | 31.1 | 34.7 | 34.3 | 34.0 |

*Notes:* Column (1) shows the results of our baseline regression, with average level of (log) value added per employment in the next 3 years as a dependent variable. Column (2) is analogous to Column (1) in terms of dependent variable, but uses inverse hyperbolic sine for the log transformation applied to a number of patents used both as an explanatory variable and as an instrument, i.e. $\ln(\sqrt{1 + pat^2} + pat)$. Columns (3), (4), and (5) use one, two, and three periods ahead value added per employment as dependent variable. All regressions include (log) values for value added per employment, capital, and employment as controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.7: 2SLS Estimates: 2000-2014

| | $\overline{\log(va\_em_{cst+n})}$ | $n \in \{1, 2, 3\}$ | | |
|---|---|---|---|---|
| | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) |
| $\log(1 + pat_{cst})$ | 0.004 | 0.017 | 0.006 | 0.017 |
| | (0.003) | (0.008) | (0.004) | (0.003) |
| $\log(va\_em_{cst})$ | 0.942 | 0.937 | 0.877 | 0.872 |
| | (0.016) | (0.016) | (0.016) | (0.014) |
| $\log(capital_{cst})$ | -0.016 | -0.014 | -0.220 | -0.219 |
| | (0.008) | (0.008) | (0.040) | (0.039) |
| $\log(employ_{cst})$ | 0.020 | 0.015 | 0.552 | 0.546 |
| | (0.010) | (0.010) | (0.044) | (0.043) |
| $\overline{\log(capital_{cst+n})}$ | | | 0.291 | 0.291 |
| | | | (0.047) | (0.047) |
| $\overline{\log(employ_{cst+n})}$ | | | -0.630 | -0.630 |
| | | | (0.048) | (0.047) |
| Country-Year FE | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y |
| # obs. | 8,357 | 8,357 | 8,169 | 8,169 |
| # countries | 36 | 36 | 36 | 36 |
| | First-stage estimates | | | |
| Predicted | | 0.461 | | 0.460 |
| $\ln(1 + patent_t)$ | | (0.079) | | (0.078) |
| F-stat | | 33.9 | | 34.9 |

*Notes:* Period of the analysis is 2000-14 using pre-determined matrix based on the data from 1970-90. First-stage estimates include all the controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Columns (1) and (3) report the results using OLS, and Columns (2) and (4) report the results obtained with 2SLS. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.8: 2SLS Estimates: Robustness, Different Lags, TFP growth (primal)

| | $\Delta \log(TFP_{cst+n})$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $\log(1 + pat_{cst})$ | 0.010 | 0.010 | 0.011 | 0.010 | 0.010 |
| | (0.006) | (0.006) | (0.006) | (0.006) | (0.005) |
| Controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| Country-Year FE | Y | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y | Y |
| # obs. | 8,336 | 8,336 | 9,710 | 9,025 | 8,336 |
| | First-stage estimates | | | | |
| Predicted | 0.472 | 0.405 | 0.468 | 0.471 | 0.472 |
| $\log(1 + pat_{cst})$ | (0.083) | (0.075) | (0.081) | (0.082) | (0.083) |
| F-stat | 32.5 | 29.5 | 33.3 | 32.7 | 32.5 |

*Notes:* Column (1) shows the results of our main regression for TFP growth estimated with the primal approach, with average level of TFP growth in the next 3 years as a dependent variable. Column (2) is analogous to Column (1) in terms of dependent variable, but uses inverse hyperbolic sine for the log transformation applied to a number of patents used both as an explanatory variable and as an instrument, i.e. $\ln(\sqrt{1 + pat^2} + pat)$. Columns (3), (4), and (5) use one, two, and three periods ahead TFP growth as dependent variable. All regressions include (log) values for the level of TFP, capital, and employment as controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.9: 2SLS Estimates: Robustness, Different Lags, TFP growth (dual)

| | $\Delta\log(TFP_{cst+n})$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $\log(1+pat_{cst})$ | 0.008 | 0.008 | 0.007 | 0.009 | 0.009 |
| | (0.003) | (0.003) | (0.003) | (0.004) | (0.004) |
| Controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| Country-Year FE | Y | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y | Y |
| # obs. | 7,391 | 7,391 | 9,304 | 8,613 | 7,951 |
| | First-stage estimates | | | | |
| Predicted | 0.470 | 0.401 | 0.466 | 0.468 | 0.470 |
| $\log(1+pat_{cst})$ | (0.080) | (0.072) | (0.078) | (0.079) | (0.080) |
| F-stat | 35.0 | 31.4 | 35.5 | 35.1 | 34.9 |

*Notes:* Column (1) shows the results of our main regression for TFP growth estimated with the dual approach, with average level of TFP growth in the next 3 years as a dependent variable. Column (2) is analogous to Column (1) in terms of dependent variable, but uses inverse hyperbolic sine for the log transformation applied to a number of patents used both as an explanatory variable and as an instrument, i.e. $\ln(\sqrt{1+pat^2}+pat)$. Columns (3), (4), and (5) use one, two, and three periods ahead TFP growth as dependent variable. All regressions include (log) values for the level of TFP, capital, and employment as controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.10: 2SLS Estimates: Robustness for TFP growth

| | $\overline{\Delta\log(TFP_{cst+n})}$ $n \in \{1,2,3\}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $\log(1+pat_{cst})$ | 0.010 | 0.017 | 0.008 | 0.011 |
| | (0.006) | (0.008) | (0.004) | (0.005) |
| $\log(\overline{1+pat_{cs1970-90}})$ | | -0.005 | | -0.005 |
| | | (0.004) | | (0.004) |
| $\log(1+\widehat{pat}_{cst-30})$ | | | 0.001 | 0.004 |
| | | | (0.003) | (0.003) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Country-Year FE | Y | Y | Y | Y |
| Sector-Year FE | Y | Y | Y | Y |
| # obs. | 8,336 | 8,336 | 8,336 | 8,336 |
| | First-stage estimates | | | |
| Predicted | 0.472 | 0.274 | 0.393 | 0.311 |
| $\ln(patent_t)$ | (0.083) | (0.062) | (0.068) | (0.060) |
| F-stat | 32.5 | 19.7 | 33.2 | 27.0 |

*Notes:* Column (1) shows the results of our main regression for TFP growth estimated using primal approach, Column (2) and (3) include separately to baseline regression historical levels of average patent activity and predicted number of patents driven by demand pull factors, respectively. And Column (4) includes both of them together. All regressions include (log) values for the level of TFP, capital, and employment as controls. Standard errors (in parentheses) are two-way clustered at the country and sector levels. Kleibergen-Paap Wald F-stat is reported for the first stage.

Table B.11: 2SLS Estimates: Innovation and Long-term Development

| | Dependent Variable is: $\overline{\log(gdp\_cap_{ct+n})}$ | | | | | |
|---|---|---|---|---|---|---|
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $\log(1 + pat_{ct})$ | 0.005 | 0.015 | 0.009 | 0.023 | 0.003 | 0.031 |
| | (0.003) | (0.005) | (0.004) | (0.008) | (0.002) | (0.021) |
| $\log(gdp\_cap_{ct})$ | 0.901 | 0.878 | 0.872 | 0.838 | 0.883 | 0.839 |
| | (0.027) | (0.026) | (0.035) | (0.040) | (0.019) | (0.040) |
| Country FE | Y | Y | Y | Y | Y | Y |
| Year FE | Y | Y | Y | Y | Y | Y |
| # obs. | 2,760 | 2,760 | 2,376 | 2,376 | 3,499 | 3,499 |
| # countries | 60 | 60 | 60 | 60 | 119 | 119 |
| F-stat | | 27.4 | | 20.7 | | 6.3 |

*Notes:* Period of the analysis is 1960-2016 for Columns (1)-(2), 1970-2016 for Columns (3)-(4) and 1980-2016 for Columns (5)-(6) using pre-determined matrix based on the data for the period pre-1950, pre-1960 and pre-1980, respectively. Standard errors (in parentheses) are clustered at the country level. Sample includes only high-income and upper-middle-income countries for Columns (1)-(4) and all countries for Columns (5)-(6). Columns (1), (3), and (5) present the results for OLS, and columns (2), (4), and (6) presents the results obtained with 2SLS. Kleibergen-Paap Wald F-stat is reported for the first stage.

## B.4 Distribution of Patents across Countries

Table B.12: Distribution of patents across countries

| Territory/ Institution ISO Code | Full Name of the Territory/Institution | Number of Patents in the Database | Year of the First Patent in the Database |
|---|---|---|---|
| AD | Andorra | 591 | 1948 |
| AE | United Arab Emirates | 5,256 | 1968 |
| AF | Afganistan | 70 | 1968 |
| AG | Antigua and Barbuda | 166 | 1903 |
| AI | Anguilla | 57 | 1982 |
| AL | Albania | 291 | 1917 |
| AM | Armenia [5] | 1,507 | 1991 |
| AN | Netherlands Antilles | 1,482 | 1906 |
| AO | Angola | 81 | 1905 |
| AP | African Regional Intellectual Property Organization (ARIPO) | 8,485 | 1964 |
| AQ | Antarctica | 2 | 2011 |
| AR | Argentina | 63,878 | 1879 |
| AT | Austria | 793,919 | 1873 |
| AU | Australia | 2,246,908 | 1876 |
| AW | Aruba | 61 | 1996 |

[5] Part of USSR

| | | | |
|---|---|---|---|
| AZ | Azerbaijan[5] | 2,022 | 1991 |
| BA | Bosnia and Herzegovina[6] | 806 | 1992 |
| BB | Barbados | 880 | 1970 |
| BD | Bangladesh | 1,584 | 1971 |
| BE | Belgium | 876,302 | 1862 |
| BF | Burkina Faso | 115 | 1993 |
| BG | Bulgaria | 49,371 | 1899 |
| BH | Bahrain | 323 | 1978 |
| BI | Burundi | 73 | 1975 |
| BJ | Benin | 71 | 1906 |
| BM | Bermuda | 1,286 | 1898 |
| BN | Brunei Darussalam | 257 | 1907 |
| BO | Bolivia | 413 | 1903 |
| BR | Brazil | 686,055 | 1891 |
| BS | Bahamas | 2,344 | 1903 |
| BT | Bhutan | 19 | 1999 |
| BV | Bouvet Island | 2 | 2004 |
| BW | Botswana | 44 | 1985 |
| BX | Benelux Office for Intellectual Property (BOIP) | 3 | 2016 |
| BY | Belarus[5] | 10,650 | 1991 |
| BZ | Belize | 163 | 1905 |
| CA | Canada | 1,942,715 | 1841 |

[6] Part of Yugoslavia

| | | | |
|---|---|---|---|
| CC | Cocos (Keeling) Islands | 4 | 1989 |
| CD | Congo, the Democratic Republic of the | 33 | 1958 |
| CF | Central African Republic | 81 | 1887 |
| CG | Congo | 164 | 1895 |
| CH | Switzerland | 1,188,116 | 1877 |
| CI | Cote D'Ivoire | 316 | 1973 |
| CK | Cook Islands | 15 | 1995 |
| CL | Chile | 17,207 | 1874 |
| CM | Cameroon | 595 | 1962 |
| CN | China | 18,680,839 | 1894 |
| CO | Colombia | 9,571 | 1894 |
| CR | Costa Rica | 5,760 | 1900 |
| CS | Czechoslovakia | 174,585 | 1882 |
| CT | Canton and Enderbury Islands | 10 | 2010 |
| CU | Cuba | 6,703 | 1876 |
| CV | Cape Verde | 17 | 2001 |
| CX | Christmas Island | 2 | 2017 |
| CY | Cyprus | 3,754 | 1921 |
| CZ | Czech Republic [7] | 95,641 | 1993 |
| DE | Germany | 10,180,051 | 1841 |
| DJ | Djibouti | 46 | 1996 |
| DK | Denmark | 518,066 | 1874 |
| DM | Dominika | 83 | 1993 |

[7] Part of Czechoslovakia

| | | | |
|---|---|---|---|
| DO | Dominikan Republic | 921 | 1913 |
| DZ | Algeria | 2,979 | 1890 |
| EA | Eurasian Patent Organization (EAPO) | 49,945 | 1996 |
| EC | Ecuador | 7,174 | 1881 |
| EE | Estonia [5] | 8,396 | 1991 |
| EG | Egypt | 17,817 | 1891 |
| EH | Western Sahara | 3 | 2015 |
| EM | Office for Harmonization in the Internal Market (Trade Marks and Designs) (OHIM) | 538 | 2005 |
| EP | European Patent Office (EPO) | 16,419 | 1979 |
| ER | Eritrea | 343 | 1955 |
| ES | Spain | 1,212,359 | 1827 |
| ET | Ethiopia | 402 | 1956 |
| FI | Finland | 511,666 | 1842 |
| FJ | Fiji | 83 | 1928 |
| FK | Falkland Islands (Malvinas) | 6 | 1998 |
| FM | Micronesia, Federated States of | 5 | 2008 |
| FO | Faroe Islands | 133 | 1895 |
| FR | France | 4,341,873 | 1844 |
| GA | Gabon | 214 | 1887 |
| GB | United Kingdom | 4,888,953 | 1782 |
| GC | Patent Office of the Cooperation Council for the Arab States of the Gulf (GCC) | 423 | 1976 |
| GD | Grenada | 65 | 1914 |

| | | | |
|---|---|---|---|
| GE | Georgia[5] | 6,420 | 1991 |
| GF | French Guiana | 17 | 1991 |
| GG | Guernsey | 8 | 2009 |
| GH | Ghana | 430 | 1977 |
| GI | Gibraltar | 349 | 1912 |
| GL | Greenland | 28 | 1950 |
| GM | Gambia | 96 | 1977 |
| GN | Guinea | 81 | 1886 |
| GP | Guadeloupe | 45 | 1994 |
| GQ | Equatorial Guinea | 25 | 1954 |
| GR | Greece | 99,905 | 1890 |
| GS | South Georgia and the South Sandwich Islands | 1 | 2017 |
| GT | Guatemala | 5,894 | 1877 |
| GW | Guinea-Bissau | 3 | 1963 |
| GY | Guyana | 158 | 1906 |
| HK | Hong Kong | 172,156 | 1892 |
| HN | Honduras | 369 | 1903 |
| HR | Croatia [6] | 13,155 | 1991 |
| HT | Haiti | 102 | 1900 |
| HU | Hungary | 198,884 | 1873 |
| IB | World Intellectual Property Organization (WIPO) (International Bureau of) | 675 | 1958 |
| ID | Indonesia | 5,719 | 1889 |

| | | | |
|---|---|---|---|
| IE | Ireland | 158,860 | 1874 |
| IL | Israel | 481,244 | 1904 |
| IM | Isle of Man | 96 | 2008 |
| IN | India | 310,660 | 1841 |
| IO | British Indian Ocean Territory | 8 | 1985 |
| IQ | Iraq | 366 | 1937 |
| IR | Iran, Islamic Republic of | 4,944 | 1890 |
| IS | Iceland | 8,238 | 1900 |
| IT | Italy | 1,714,990 | 1841 |
| JE | Jersey | 247 | 2010 |
| JM | Jamaica | 609 | 1895 |
| JO | Jordan | 2,218 | 1952 |
| JP | Japan | 31,402,515 | 1889 |
| KE | Kenya | 1,837 | 1921 |
| KG | Kyrgystan[5] | 685 | 1991 |
| KH | Cambodia | 176 | 1906 |
| KI | Kiribati | 13 | 1966 |
| KM | Comoros | 9 | 2001 |
| KN | Saint Kitts and Nevis | 132 | 1978 |
| KP | Korea, Democratic People's Republic of | 1,263 | 1963 |
| KR | Korea, Republic of | 5,438,867 | 1910 |
| KW | Kuwait | 1,396 | 1972 |
| KY | Cayman Islands | 1,273 | 1875 |
| KZ | Kazakstan [5] | 4,697 | 1991 |

| | | | |
|---|---|---|---|
| LA | Lao People's Democratic Republic | 242 | 1901 |
| LB | Lebanon | 2,199 | 1893 |
| LC | Saint Lucia | 20 | 1993 |
| LI | Liechtenstein | 19,729 | 1912 |
| LK | Sri Lanka | 1,860 | 1891 |
| LR | Liberia | 71 | 1964 |
| LS | Lesotho | 11 | 1993 |
| LT | Lithuania [5] | 7,040 | 1991 |
| LU | Luxemburg | 75,293 | 1880 |
| LV | Latvia [5] | 9,559 | 1991 |
| LY | Libyan Arab Jamahiriya | 72 | 1937 |
| MA | Morocco | 21,116 | 1898 |
| MC | Monaco | 6,845 | 1902 |
| MD | Moldova, Republic of [5] | 12,897 | 1991 |
| ME | Montenegro[7] | 180 | 2006 |
| MG | Madagascar | 298 | 1904 |
| MH | Marshall Islands | 35 | 1995 |
| MK | Macedonia, the Formet Yugoslav Republic of [7] | 360 | 1991 |
| ML | Mali | 149 | 1971 |
| MM | Myanmar | 267 | 1902 |
| MN | Mongolia | 605 | 1907 |
| MO | Macau | 630 | 1945 |
| MP | Northern Mariana Islands | 8 | 2003 |

| | | | |
|---|---|---|---|
| MQ | Martinique | 6 | 2000 |
| MR | Mauritania | 347 | 1968 |
| MS | Montserrat | 24 | 1994 |
| MT | Malta | 1,994 | 1902 |
| MU | Mauritius | 444 | 1908 |
| MV | Maldives | 12 | 1982 |
| MW | Malawi | 535 | 1973 |
| MX | Mexico | 58,718 | 1876 |
| MY | Malaysia | 43,645 | 1841 |
| MZ | Mozambique | 11 | 2005 |
| NC | New Caledonia | 153 | 1909 |
| NE | Niger | 222 | 1908 |
| NF | Norfolk Island | 4 | 2003 |
| NG | Nigeria | 803 | 1905 |
| NI | Nicaragua | 141 | 1894 |
| NL | Netherlands | 1,329,982 | 1874 |
| NO | Norway | 385,455 | 1874 |
| NP | Nepal | 704 | 1970 |
| NR | Nauru | 14 | 1982 |
| NU | Niue | 4 | 1996 |
| NZ | New Zealand | 183,876 | 1883 |
| OA | African Intellectual Property Organization (OAPI) | 9,335 | 1915 |
| OM | Oman | 406 | 1981 |

| | | | |
|---|---|---|---|
| PA | Panama | 5,066 | 1897 |
| PC | Pacific Islands (Trust Territory) | 1 | 1985 |
| PE | Peru | 4,058 | 1890 |
| PF | French Polynesia | 38 | 1996 |
| PG | Papua New Guinea | 55 | 1976 |
| PH | Philippines | 28,572 | 1901 |
| PK | Pakistan | 2,775 | 1956 |
| PL | Poland | 354,911 | 1887 |
| PM | Saint Pierre and Miquelon | 3 | 2015 |
| PN | Pitcairn | 12 | 1998 |
| PT | Portugal | 28,207 | 1889 |
| PW | Palau | 5 | 2003 |
| PY | Paraguay | 179 | 1910 |
| QA | Qatar | 1,251 | 1988 |
| RO | Romania | 93,879 | 1880 |
| RS | Serbia[7] | 5,978 | 1992 |
| RU | Russian Federation [5] | 1,137,189 | 1991 |
| RW | Rwanda | 42 | 1972 |
| SA | Saudi Arabia | 17,172 | 1887 |
| SB | Solomon Islands | 13 | 2003 |
| SC | Seychelles | 581 | 1901 |
| SD | Sudan | 246 | 1919 |
| SE | Sweden | 1,330,679 | 1847 |
| SG | Singapore | 130,498 | 1896 |

| | | | |
|---|---|---|---|
| SH | Saint Helena | 53 | 1989 |
| SI | Slovenia[7] | 21,005 | 1992 |
| SJ | Svalbard and Jan Mayen | 8 | 1895 |
| SK | Slovakia[6] | 22,355 | 1993 |
| SL | Sierra Leone | 343 | 1913 |
| SM | San Marino | 705 | 1947 |
| SN | Senegal | 275 | 1906 |
| SO | Somalia | 34 | 1961 |
| SR | Suriname | 56 | 1977 |
| ST | Sao Tome and Principe | 1,843 | 1975 |
| SU | Union of Soviet Socialist Republics | 1,361,507 | 1845 |
| SV | El Salvador | 792 | 1889 |
| SY | Syrian Arab Republic | 571 | 1903 |
| SZ | Swaziland | 192 | 1888 |
| TC | Turks and Caicos Islands | 123 | 1996 |
| TD | Chad | 78 | 1976 |
| TF | French Southern Territories | 5 | 2014 |
| TG | Togo | 56 | 1973 |
| TH | Thailand | 11,225 | 1896 |
| TJ | Tajikistan[5] | 1,368 | 1991 |
| TK | Tokelau | 62 | 2002 |
| TM | Turkmenistan[5] | 1,696 | 1991 |
| TN | Tunisia | 3,917 | 1889 |
| TO | Tonga | 45 | 1994 |

| | | | |
|---|---|---|---|
| TP | East Timor | 2 | 1998 |
| TR | Turkey | 68,292 | 1875 |
| TT | Trinidad and Tobago | 732 | 1898 |
| TV | Tuvalu | 44 | 1996 |
| TW | Taiwan | 1,492,230 | 1967 |
| TZ | Tanzania, United Republic of | 186 | 1972 |
| UA | Ukraine[5] | 185,014 | 1991 |
| UG | Uganda | 186 | 1911 |
| US | United States | 16,900,448 | 1790 |
| UY | Uruguay | 5,252 | 1897 |
| UZ | Uzbekistan[5] | 1,933 | 1991 |
| VA | Holy See (Vatican City State) | 62 | 1911 |
| VC | Saint Vincent and the Grenadines | 40 | 1993 |
| VE | Venezuela | 4,725 | 1891 |
| VG | Virgin Islands, British | 2,679 | 1984 |
| VN | Viet Nam | 4,098 | 1903 |
| VU | Vanuatu | 49 | 1993 |
| WF | Wallis and Futuna | 7 | 2012 |
| WO | World Intellectual Property Organization (WIPO) (International Bureau of) | 378,706 | 1980 |
| WS | Samoa | 129 | 1989 |
| XN | Nordic Patent Institute (NPI) | 1 | 2015 |
| YE | Yemen | 91 | 1985 |
| YU | Yugoslavia | 26,813 | 1891 |

| | | | |
|---|---|---|---|
| ZA | South Africa | 316,377 | 1889 |
| ZM | Zambia | 1,911 | 1920 |
| ZW | Zimbabwe | 2,962 | 1961 |

# Appendix C

# Appendix to Chapter Three

## C.1 Data Sources

We use data from the Nigeria COVID-19 National Longitudinal Phone Survey (Covid-19 NLPS) implemented by the National Bureau of Statistics to track the impact of the pandemic. The survey was conducted for one year on a monthly basis starting from the end of April, 2020, and included households interviewed face-to-face in 2018/2019 for Wave 4 of the General Household Survey Panel (GHS-Panel), which was designed to be representative at national and zonal levels. The extensive information collected in the GHS-Panel just over a year prior to the pandemic provides a rich set of background information on Covid-19 NLPS households. 1,950 households were successfully interviewed in Round 1, and the same households were contacted by phone in subsequent rounds.[1] There are total 12 phone surveys conducted on a monthly basis starting from the end of April 2020 (see Figure C.3).

We rely on data collected in Rounds 5 and 10 of Covid-19 NLPS and 2018/19 GHS-Panel. We choose these two surveys because they line up with the timing of the pre-pandemic information from GHS-Panel. Round 5 of Covid-19 NLPS was conducted in September 2020 and Round 10 in February, 2021. The post-planting part of the 2018/19 GHS-Panel was conducted in the period July–September 2018 and the post-harvest part in January–February, 2019. For the former, data on employment status and hours worked on a primary

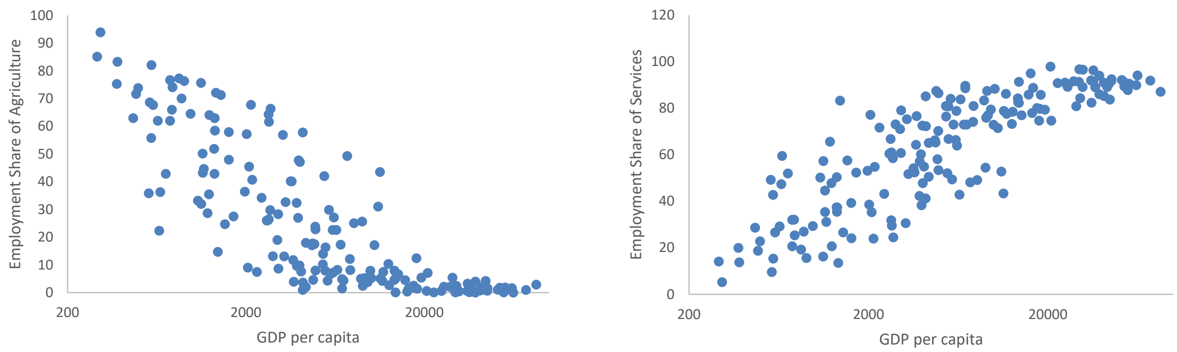
[1] Households that do not have access to a phone and could not be interviewed despite several call attempts were excluded from the sample, which may introduce potential selection bias. To overcome this bias, a balanced sampling approach was adopted, and phone survey weights are available.

job in the week before the interview is collected for up to six randomly selected members of households age 15-64 plus the primary respondent. In GHS-Panel, employment status and hours worked on each job a week before the interview were collected for each member of household age five and above. For consistency with the Covid-19 NLPS data, we use hours worked on the primary job, defined as job were individual spent the most time during the last week, rather than all jobs.

## C.2 Additional Tables and Figures

Figure C.1 plots the employment shares for women (out of all employed women) in agriculture and services in 2015 against GDP per capita for most countries in the world. The figure shows that in low-income countries, the majority of the female labor force is in agriculture, whereas services are relatively unimportant. The opposite pattern is observed in high-income economies, where the employment share of agriculture is negligible and most women work in services. The figure suggests that unlike in high-income countries, in low-income countries the specific impact of Covid-related shutdowns on contact intensive services does not play a substantial role for women's employment losses during the pandemic.

Figure C.2 depicts the cross-country relationship between income per capita and engagement of children in any learning activities during school closures. We use data from High Frequency Phone Surveys conducted by the World Bank to identify the share of households with children engaged in any learning activity after schools were closed due to Covid-19. Only households with children who attended school prior to the pandemic are considered when this share is calculated. The figure shows that in countries with higher income per capita, on average, children were more likely to continue their education during the pandemic. In a number of countries in Sub-Saharan Africa, children continued with learning activities in less than half of households.




(a) Women's Employment Share in Agriculture (b) Women's Employment Share in Services

Figure C.1: The Sectoral Composition of Women's Employment Across Countries in 2015

*Notes:* Women's employment in agriculture and services as a fraction of total women's employment in 2015. Each dot is a country. Source: World Bank Development Indicators; accessed online on 12/21/2021.

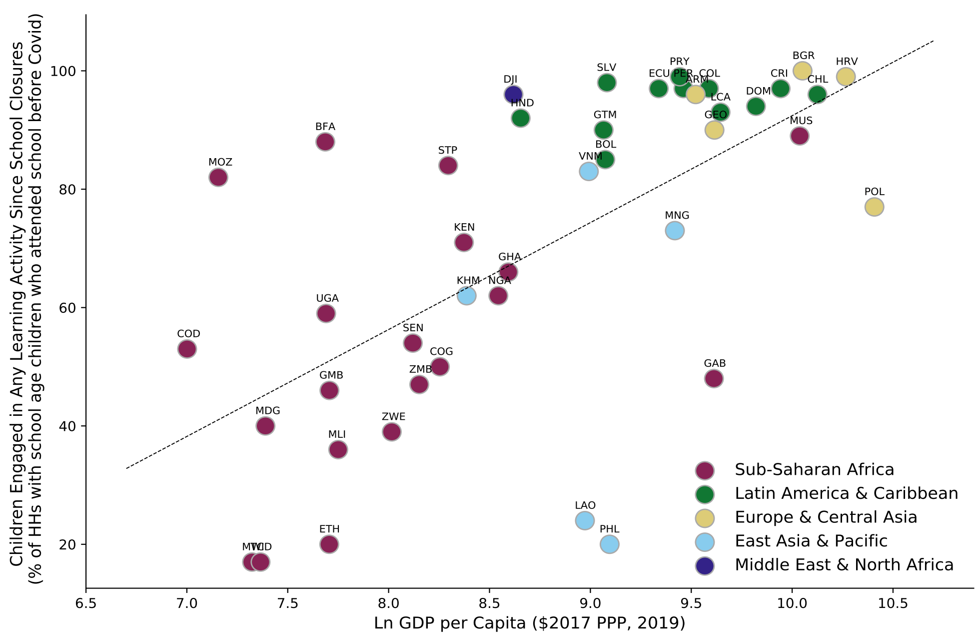



Figure C.2: Learning Activities during School Closures and Income

*Notes:* This figure is generated using data from High Frequency Phone Surveys (World Bank). Data collected during first rounds of phone surveys for each country is used for the share of HHs where children engaged in any learning activity. In most countries, first rounds were conducted in May-June 2020.

Figure C.3 provides a timeline of the stringency of government containment measures during the pandemic and of mobility data collected by Google. The figure also shows when each wave of the Covid-19 NLPS survey was conducted. The figure shows that restrictions were the most severe from April to July of 2020, and that by September (when the 5th wave that we use here was collected) restrictions were already more relaxed. There is little change overall between waves 5 and 10; however, most schools fully reopened in November of 2020, in between the data collection of these two waves.

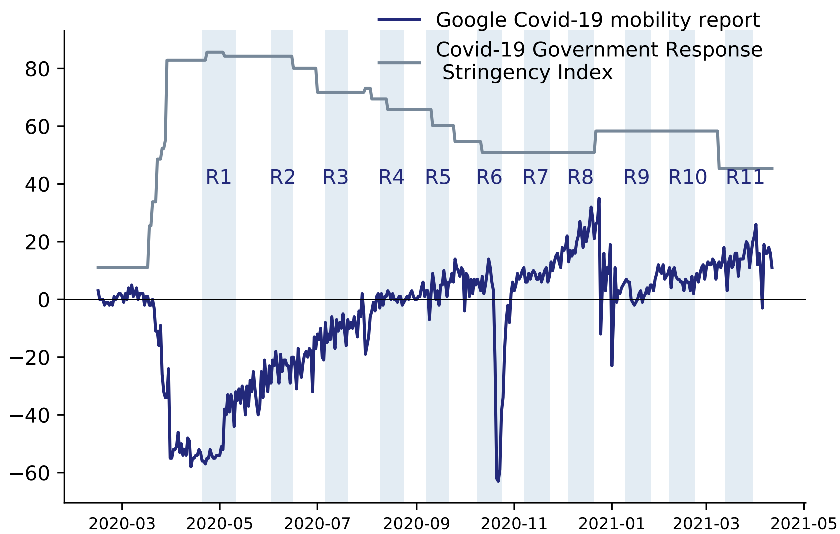



Figure C.3: Timeline of Government Restrictions and Population Mobility

*Notes:* Google Covid-19 mobility report shows mobility trends for public transport hubs (subway, bus, and train stations) relative to a baseline value – median value for the corresponding day of the week during the 5-week period Jan 3 - Feb 6, 2020. Covid-19 Government Response Stringency Index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, re-scaled to a value from 0 to 100 (100 = strictest).

Figure C.4 provides a timeline of school closures during the pandemic. The figure shows that schools were closed in March 2020 as a response to Covid-19 outbreak. School reopened partially for some students at the end of September 2020, and fully reopened for all students in November 2020.

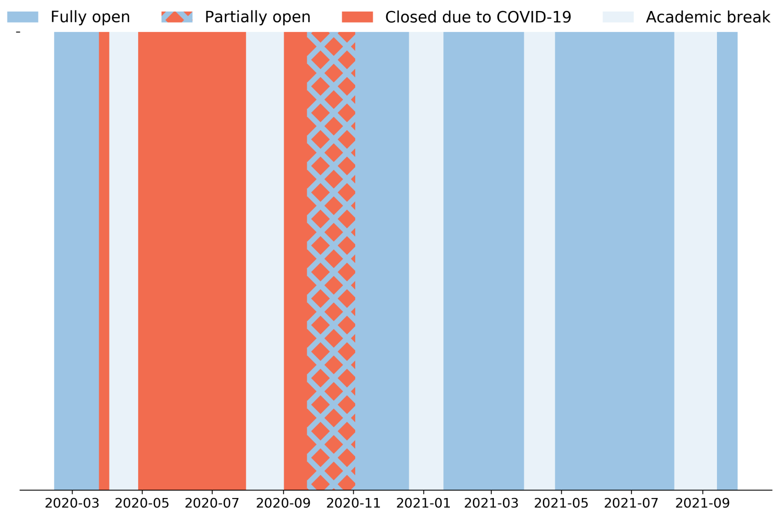
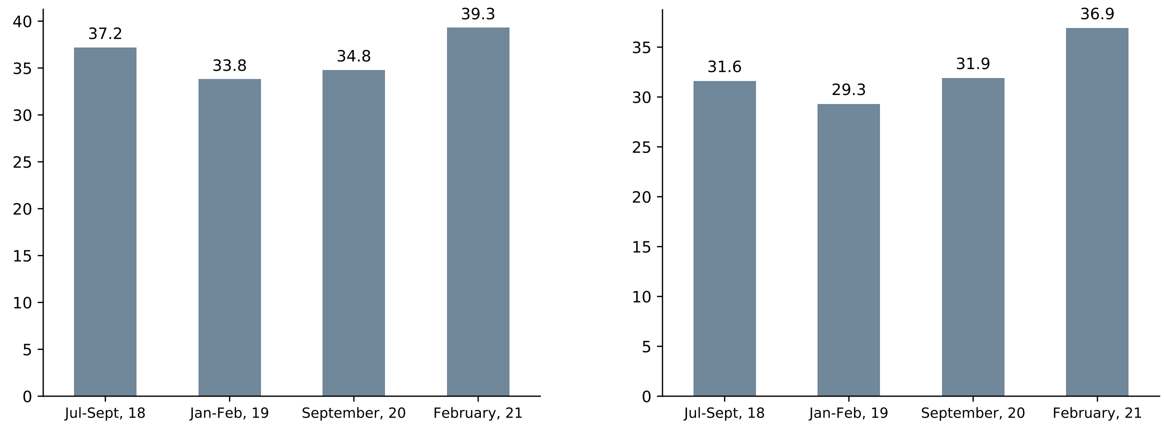



Figure C.4: Timeline of Schools Closures in Nigeria

*Notes:* This figure is generated using UNESCO "Global monitoring of school closures" data.

Figure C.5 provides an impression of the intensive margin of employment changes by plotting for each survey wave and each gender the weekly hours worked conditional on being employed. For wave 5 (September 2020), hours changes compared to the pre-pandemic period are moderate, but weekly working hours of both women and men are considerably higher than previously in the wave 10 data (February 2021). A caveat is that average weekly working hours are computed for the primary activity only. Therefore, increase in working hours might reflect that some individuals shift from multiple jobs to the single one, which can drive up average weekly hours for primary activity.

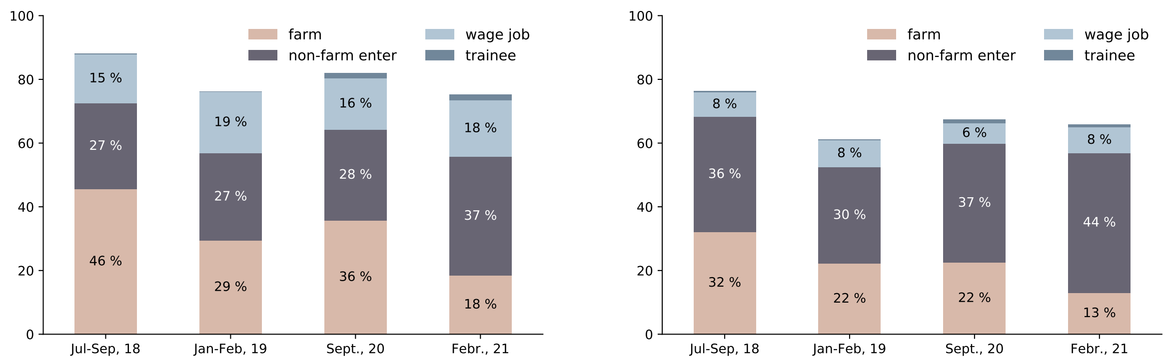
Figure C.6 depicts employment across different sectors for both women and men. The most notable change is a sharp rise in non-farm enterprise; for women, for example, we observe an increase from 30 percent in January-February 2019 to 44 percent in February 2021. The data is consistent with the view that households responded to income losses by increasing self-employment and small-scale entrepreneurship. We also observe a decline in agricultural employment; because only the sector of the primary job is reported, this may




(a) Male Adults (b) Female Adults

Figure C.5: Average Weekly Working Hours by Gender

*Notes:* Average weekly working hours are computed for the primary working activity and conditional on individual to have a job. Primary working activity is defined as the job in which the individual worked the most hours.




(a) Male Adults (b) Female Adults

Figure C.6: Share of Working Adults by Sector

*Notes:* The share of adults of age 21-55 that worked in the past week (at time when interview was conducted) at a given sector as a primary employment. Non-farm enterprise stands for the enterprise that belongs to a member of household. Sample includes $\approx 9,000$ and $\approx 4,000$ individuals for pre-Covid and Covid interviews, respectively.

reflect that some households members took on a new job as primary employment, leaving agriculture as a secondary activity.

Table C.1: Impact of Covid-19's Weekly Working Hours for Adults

| | Weekly Working Hours | | | | log (Weekly Working Hours) | | | |
|---|---|---|---|---|---|---|---|---|
| | Sept. | Sept. | Febr. | Febr. | Sept. | Sept. | Febr. | Febr. |
| Covid | -4.156 | -4.926 | 4.251 | 2.573 | -0.356 | -0.298 | 0.154 | 0.022 |
| | (0.980) | (1.297) | (1.739) | (1.935) | (0.067) | (0.082) | (0.135) | (0.152) |
| Covid × Female | | 1.414 | | 3.096 | | -0.107 | | 0.243 |
| | | (1.073) | | (1.252) | | (1.264) | | (0.010) |
| # Obs | 12,094 | 12,094 | 12,404 | 12,404 | 12,094 | 12,094 | 12,404 | 12,404 |
| R-squared | 0.21 | 0.21 | 0.23 | 0.23 | 0.21 | 0.21 | 0.24 | 0.24 |
| Mean Pre-Covid | 28.0 | 28.0 | 21.5 | 21.5 | | | | |
| Age FE | Y | Y | Y | Y | Y | Y | Y | Y |
| LGA FE | Y | Y | Y | Y | Y | Y | Y | Y |
| Control Variables | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are clustered at the state level. Controls include gender, urban, number of HH members, access to electricity & internet, ownership of different assets (radio, car, land, etc.), access to finance, consumption quantile before the pandemic, education and literacy of the individual, marriage status, whether individual is a head of household, and a dummy for pre-covid interview held in January. Results for weekly working hours that combine both intensive and extensive margins and we apply inverse-hyperbolic sine transform of hours worked last week for the logarithm.

Table C.1 displays individual-level regression results of the impact of the pandemic on both extensive and intensive margin of employment by gender that include individual and household controls and geographic fixed effects (LGA). The regressions confirm that individuals worked less in the early phase of the pandemic, but experienced an expansion of working hours later in the recovery, driven primarily by female working hours.

The combination of school closures and the socioeconomic impact of the pandemic might have induced some adolescents, especially from poor households, to stop their education and start working. Table C.2 displays regression results for the impact of the pandemic on the employment of individuals at ages 15 to 20. Panel A displays the results for all individuals aged 15-20 years old, while Panels B and C show the results for those who are supposed to be in secondary school or receive tertiary education, based on their age. We find that the pandemic led both to a higher probability for adolescents to work and more weekly working hours. While we observe an increase in the probability of performing some work for all age

Table C.2: Impact of Covid-19's on Employment and Hours of Work for Adolescents

| | Employment Status | | | Weekly Working Hours | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **Panel A: All individuals aged 15-20** | | | | | | |
| Covid | 0.075 | 0.078 | 0.056 | 4.515 | 4.087 | 4.830 |
| | (0.013) | (0.014) | (0.008) | (1.546) | (2.185) | (1.741) |
| Covid × Female | | -0.006 | | | 1.062 | |
| | | (0.016) | | | (2.652) | |
| Covid × Urban | | | 0.051 | | | -1.121 |
| | | | (0.016) | | | (3.095) |
| # Obs | 4,997 | 4,997 | 4,997 | 1,639 | 1,639 | 1,639 |
| R-squared | 0.93 | 0.93 | 0.93 | 0.44 | 0.45 | 0.44 |
| Mean Pre-Covid | 0.301 | | | 23.7 | | |
| **Panel B: All individuals aged 15-16** | | | | | | |
| Covid | 0.056 | 0.061 | 0.043 | -0.431 | 0.014 | -0.797 |
| | (0.011) | (0.015) | (0.008) | (2.977) | (3.305) | (3.009) |
| Covid × Female | | -0.008 | | | -1.094 | |
| | | (0.022) | | | (3.214) | |
| Covid × Urban | | | 0.042 | | | 1.870 |
| | | | (0.020) | | | (5.639) |
| # Obs | 1,828 | 1,828 | 1,828 | 457 | 457 | 457 |
| R-squared | 0.95 | 0.95 | 0.95 | 0.59 | 0.59 | 0.59 |
| Mean Pre-Covid | 0.250 | | | 20.2 | | |
| **Panel C: All individuals aged 17-20** | | | | | | |
| Covid | 0.087 | 0.091 | 0.067 | 5.763 | 5.143 | 6.714 |
| | (0.016) | (0.018) | (0.012) | (1.579) | (2.261) | (1.823) |
| Covid × Female | | -0.011 | | | 1.566 | |
| | | (0.017) | | | (3.592) | |
| Covid × Urban | | | 0.055 | | | -3.184 |
| | | | (0.019) | | | (3.406) |
| # Obs | 3,115 | 3,115 | 3,115 | 1,095 | 1,095 | 1,095 |
| R-squared | 0.93 | 0.93 | 0.93 | 0.47 | 0.47 | 0.47 |
| Mean Pre-Covid | 0.333 | | | 25.4 | | |
| Age FE | Y | Y | Y | Y | Y | Y |
| Occupation FE | Y | Y | Y | Y | Y | Y |
| LGA FE | Y | Y | Y | Y | Y | Y |
| Control Variables | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are clustered at the state level. Controls include gender, urban, number of HH members, access to electricity & internet, ownership of different assets (radio, car, land, etc.), consumption quantile before the pandemic, education and literacy of the HH's head, and dummy for pre-covid interview held in January. In regressions for weekly working hours only working adolescents are included.

groups, weekly hours are higher only for the older cohort. Additionally, we find that the probability of work increased more for those living in urban areas compared to rural. We find no significant differences in the effects of the pandemic between women and men.

To examine the possible role of the income channel for employment changes, we split the sample into the top 40% vs. the bottom 60% of households defined by consumption prior to the pandemic. Table C.3 displays regression results for the impact of the pandemic

Table C.3: Impact of Covid-19's on Employment in February for Different Income Groups

| | Employment Status | | | | Weekly Working Hours | | | |
|---|---|---|---|---|---|---|---|---|
| | Bottom 60% | | Top 40% | | Bottom 60% | | Top 40% | |
| Covid | 0.100 | 0.049 | -0.037 | -0.036 | 6.635 | 4.565 | 0.387 | -0.671 |
| | (0.043) | (0.044) | (0.041) | (0.042) | (2.331) | (2.710) | (2.211) | (2.207) |
| Covid × Female | | 0.086 | | -0.003 | | 3.462 | | 2.145 |
| | | (0.033) | | (0.024) | | (1.747) | | (1.375) |
| # Obs | 8,243 | 8,243 | 4,162 | 4,162 | 8,222 | 8,222 | 4,142 | 4,142 |
| R-squared | 0.23 | 0.21 | 0.30 | 0.30 | 0.24 | 0.24 | 0.28 | 0.28 |
| Mean Pre-Covid | 0.67 | 0.67 | 0.71 | 0.71 | 28.9 | 28.9 | 36.8 | 36.8 |
| Age FE | Y | Y | Y | Y | Y | Y | Y | Y |
| LGA FE | Y | Y | Y | Y | Y | Y | Y | Y |
| Control Variables | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* Robust standard errors (in parentheses) are clustered at the state level. Controls include gender, urban, number of HH members, access to electricity & internet, ownership of different assets (radio, car, land, etc.), access to finance, consumption quantile before the pandemic, education and literacy of the individual, marriage status, whether individual is a head of household, and a dummy for pre-covid interview held in January. Consumption quantiles are computed for pre-pandemic quantities.

on the employment by gender in February for the two groups. We find that the positive effect of the pandemic on women's labor supply in February 2021 is concentrated among poorer households. In fact, there is no effect for those households in the top 40% of the (pre-pandemic) consumption distribution. These findings provide suggestive evidence for the income channel.