Title
Three Essays on Household Adaptation

Author
Michuda, Aleksandr

Publication Date
2021

Peer reviewed|Thesis/dissertation

Three Essays on Household Adaptation

By

Aleksandr Michuda

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

in

Agricultural and Resource Economics

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Michael R. Carter, Chair

Travis Lybbert

Dalia Ghanem

Committee in Charge

2021

Dedicated to Iffat Chowdhury, my partner in life.

Contents

List of Figures

List of Tables

Abstract

Three Essays on Household Adaptation

The chapters in this dissertation provide a way of tracing out a history of sorts; of how households and people coped with negative shocks in the past: through cooperative land tenure arrangements and collectives, and now: through technological platforms that follow the "economic laws of the market." The questions I ask are: How do households adapt to new circumstances in developing countries? And how has the landscape of the labor market changed to incorporate new technology for these households to use? How has new technology and new data allowed us to say more about not just households' economic lives, but their political lives as well?

In the first chapter, I study how ride-share applications can be used to send money from cities to rural areas in times of bad rural shocks. Rural-urban linkages have long been a topic of study in the developing world. Remittances are often a key driver of these linkages and can act as insurance against rural weather shock risk, in the absence of availability and access to formal insurance products. The emergence of new technologies, such as ride-share and mobile money platforms, can be potentially transformative at allowing remittance flows to adjust more quickly to adverse shocks. I use a dataset of Uber driver labor supply and a rich dataset of weather indicators to estimate the effect of adverse weather shocks in rural areas on Uber drivers in Kampala, Uganda. Since I do not have explicit information on migrant status and rural connection, I leverage an external dataset of Ugandan voter registration and train a gradient boosting classifier on Ugandan surnames to predict drivers' regions of origin. I develop a switching regression estimator to address the misclassification bias from the predictions. I find that a one standard deviation increase in the intensity of agricultural drought is associated with an increase of 5.1 hours online in the month of the event (a 6% increase over average hours), providing suggestive evidence that Uber's flexibility is used to buffer against adverse weather shocks.

In the second chapter, I study ethnic voting in Uganda. A large literature on ethnic

voting has shown that ethnic identity is a major element in voting preferences for many developing countries, and that higher fractionalization in ethnicity leads to lower public goods provision. It is often difficult to disentangle ethnic identity and shared economic goals that stem from living in the same location. This paper differentiates between the effects of a voter's ethnic identity and their location, on voting behavior. We estimate these effects by pairing voter registration data with election outcomes from polling stations throughout Uganda. We overcome the challenge of measuring ethnic identity by using a machine learning algorithm that exploits variation in surnames across ethnic and linguistic groups on the polling station level. In particular, we use the letter sequences in each individual voter's surname to predict their ethnicity. We then use these predictions as the explanatory variable of interest in a regression model with the outcome being support for the incumbent president in the 2016 general election. We find that ethnic voting preferences are robust to location effects, but that location tends to amplify preferences, especially when in an area with similar views, and with a history of preferences disfavoring the candidate.

In the third chapter, I explore an institution that existed before technology began to shape developing economies: the cooperative farm. Even despite there being heterogeneity in ability in cooperative agriculture, these institutional forms have existed for many years, because of their ability to pool costs amongst its members. I look at how the success of recent policies aimed at privatizing these communal farms (where I take as a context, the *PROCEDE* reform in Mexico) are affected by market power in the supply chain. What I find is that a higher level of monopsony power in the supply chain leads to lower price passthrough, meaning that an increase in the world price of tortillas, for example, will not transmit completely to an increase in the price of corn. This implies less of an increase to privatization of the cooperative farm.

Acknowledgments

To my parents for coming to this country and showing me unconditional love.

To my brother and sister for always believing in me and for teaching me unconditional support.

To my Amu and Abu who have only shown me love since I've known them.

To my new family that has taught me what it feels like to have more sisters and brothers than I could ask for (thank you Iftiar, Erin, and Jetuji)

To my friends for making graduate school BEER-able (thank you Oscar, Cynthia, Juan, Cristina, Jesús, Matthieu, François and Karen and many more)

Chapter 1

Introduction

> ...the society which projects and undertakes the technological transformation of nature alters the base of domination by gradually replacing personal dependence (of the slave on the master, the serf on the lord of the manor, the lord on the donor of the fief, etc.) with dependence on the "objective order of things" (on economic laws, the market etc.).
>
> One Dimensional Man - 1964
>
> Herbert Marcuse

Herbert Marcuse's quote from the "One-Dimensional Man" was written before we had the internet, smartphones, Uber or computers capable of running large-scale machine learning algorithms, but it's more useful than ever to consider. New technologies have given people in developing countries a new way to make money, save money, reduce search frictions for labor opportunities, and it has given them access to the internet and all of its possibilities.

But in the field of economics, we generally take this increase in access and narrowing of the digital gap to be a positive change. And the chapters in this dissertation show that new platforms and technologies can be powerful. Uber can help a person send money much more quickly than before, machine learning can let us understand voting behavior in Uganda better than before. But it still remains to be seen, whether our trading of one master for another is actually a net-positive.

The chapters in this dissertation provide a way of tracing out a history of sorts; of how households and people coped with negative shocks in the past: through cooperative land tenure arrangements and collectives, and now: through technological platforms that follow the "economic laws of the market."

The questions I ask are: How do households adapt to new circumstances in developing countries? And how has the landscape of the labor market changed to incorporate new technology for these households to use? How has new technology and new data allowed us to say more about not just households' economic lives, but their political lives as well?

In the first chapter of my dissertation, I study the impact of new technologies on household adaptation. The topic itself finds its beginning in the study of urban to rural resource flows. This discussion has been around since Soviet industrialization, when the rural sector was used to industrialize urban centers. In developing economies later in history, however, the rural sector would actually be capitalized with resources from urban centers. This would include using urban labor income to insure against rural weather risk, or to use that income to invest in agricultural production.

Recently, with the introduction of the internet, smartphones, and GPS, new technologies have emerged aimed at making profit, but also in providing a service to people in developing economies. This comes at a time when many countries are suffering from large youth unemployment in the urban sector, and from a rural perspective, a time where there is a lack of formal insurance products or savings mechanisms for growers and subsistence farmers. Ride-share platforms such as Uber and SafeBoda provide a way to take part in

flexible work, quickly, and without interruption of employment.

A natural question to ask is whether these platforms actually create urban-to-rural resource flows and can act as a way for farmers to insure against drought or flooding risk.[1] There is potential for this to be the case, given that an Uber driver, for instance, can react quickly to a change in the circumstances of their family farm in the case of a bad yield realization. There is evidence to suggest that rural-urban resource flows are strong, with many Ugandans working in the city while living in the country, with strong ties to their places of origin.

The challenge in this context is to figure out how to have your cake and eat it too. Ride-share platform data is high-frequency, precise and objective, but without anything other than what is business critical. Uber certainly does not collect data on rural connection or rural landholdings, so there's no obvious way to connect drought shocks to Uber drivers. The solution comes in the form of anthropology and machine learning. As it turns out, Uber operates in a country where surnames can connect back to rural origins: Uganda. Since internal migration is usually intra-regional (apart from to Kampala), the ethno-linguistic content of surnames gives us geographic information. And once that link is made, weather shocks in rural area can be connected to urban Uber drivers.

With a large enough dataset of surnames (such as a publicly available voter registration dataset), we can use machine learning to create a probabilistic relationship between Uber drivers and their rural connections. What I also contribute to the nascent field of econometric causal identification with machine learning methods, is a switching regression estimator that is robust to the misclassification caused by this probabilistic relationship.

In the next chapter, I use the voter registration dataset that I used to connect Uber drivers to study a political economy topic important in Ugandan politics: ethnic voting. There is a large literature about ethnic voting behavior for many African countries. But two observations make these studies limited in scope: (1) the lack of external validity, and

[1] This is especially important in the face of increased (in frequency and severity) adverse weather events due to climate change.

(2) not being able to disentangle the effect of ethnic identity and location based effect that comes from living in a place with a particular ethnic group. For example, in Uganda, Yoweri Museveni has ruled as president since 1986. Over the course of his rule, he has fomented ethnic divisions and created an elite made up of those from the Central and Western parts of the country. It is well known that, due to historical and political reasons, the North of the country tends to resent and vote against Museveni in every election. But what would happen if a member of the North grew up or moved to the West? Is location, cultural values and economic goals deterministic of political beliefs, or is the ethnic effect so strong that it overpowers it?

In our paper, we use similar predictions to the ones made in Chapter 2, to see if there is some connection between predicted ethnic identity and voting patterns. Since we have access to the polling station of registration, we run an regression analysis that evaluates the effect of ethnic identity, along with being able to look at heterogeneous effects based on residence region. Since we have a nationally representative voter registration dataset, we can make these predictions for the whole country, and since we have data across the country, we also have regional variation in voting.

What we find is that the ethnic identity effect does not change based on location, but that location can attenuate or amplify an ethnic voting preference. If a voter from the East moves to the North, then they decrease their support even more than if they stayed in the East.

A unifying theme behind these chapters are the use of big data sets and how they can be used in the presence of data gaps. Data scarcity in the developing world is an ever-present problem and it inhibits the ability to make policy relevant research possible without large funds to collect data. In Chapter 2, I show how it is possible to connect two seemingly disconnected datasets by using publicly available big data to generate predictions that can infer their relationship. In Chapter 4, we go one step further and use this big data to make a novel methodological contribution for studying ethnic voting behavior. In this case the big data allows us to analyze voting data on a national level, not just for small

urban samples as has been done in the past.

Finally, I take a step back and explore an institution that existed before technology began to shape developing economies: the cooperative farm. Even despite there being heterogeneity in ability in cooperative agriculture, these institutional forms have existed for many years, because of their ability to pool costs amongst its members. I look at how the success of recent policies aimed at privatizing these cooperative farms (where I take as a context, the *PROCEDE* reform in Mexico) are affected by market power in the supply chain.

What I find is that a higher level of monopsony power in the supply chain leads to lower price passthrough, meaning that an increase in the world price of tortillas, for example, will not transmit completely to an increase in the price of corn. This implies less of an increase to privatization of the cooperative farm.

The last chapter of my dissertation stands in contrast to the previous two to show that modern technologies have changed the way that households cope with changes, and how the researcher has changed their approach to exploring problems. The institutional norms of the past necessitated that communities came together to pool and share in costs for the betterment of its members, whereas today, digital applications have taken over this task, for better or for worse. Similarly, the researcher must now rely more heavily on data-driven and computer-intensive methodologies to help them with conducting policy-relevant analysis (as in Chapter 2 and 3), and not through the development of a mathematical model to help them with gaining intuition on a subject (as in Chapter 4). This dissertation hopes to bring attention to these connections.

Chapter 2

Urban Labor Supply Responses to Rural Drought Shocks on Rideshare Platforms

2.1 Introduction

Rural-urban linkages have long been a topic of study in the developing world. Remittances are often a key driver of these linkages and can act as insurance against rural weather shock risk (Kazianga and Wahhaj (2020); Stobl and Valfort (2013); Mueller and Osgood (2009); McKay and Deshingkar (2014)), in the absence of availability and access to formal insurance products. In risk sharing environments, remittance money can act as both an additional source of income, but also a way to diversify income streams (Lucas (1997); Rosenzweig and Stark (1989); Stark and Lucas (1988); Lesetedi (2003)). This makes urban labor supply behavior in developing countries different from more developed countries, in that they are responsive not just to urban labor shocks, but also to incidences of adverse rural shocks as well. Remittance networks connect rural communities directly to the urban labor market and can have considerable positive welfare effects (Lagakos et al. (2018)). This is particularly true in countries where rural areas are vulnerable to weather shocks (Binswanger and Rosenzweig (1993); Townsend (1994); Dercon (2002)). Linkages are

buttressed by ethnic or clan relationships that persist through time (Gelan (2002)). In many developing countries with growing urban areas, a large part of the population still resides in the rural sector and is often employed in agriculture.

The emergence of new technologies, such as ride share and mobile money platforms, have relaxed the constraints for sending remittances in bad times. Ride-share platforms like Uber provide a way to flexibly adjust labor hours during times of adverse weather shocks. These new technologies can have far-reaching policy consequences in developing countries, both as a way to study the effects of shocks and as a way to see how flexibility in payment and choosing hours can aid in poverty alleviation. For instance, the advent of mobile money has made it easier to smooth consumption across time (Jack and Suri (2014); Riley (2018)), transfer remittances (Blumenstock et al. (2016)) and improve household resilience to shocks (Bharadwaj et al. (2019)). But although the ride-share platforms have been studied in the context of developed contexts such as the U.S. with Uber and Lyft (Berger et al. (2018); Shokoohyar et al. (2020)), it has been investigated less in the context of emerging economies. These platforms provide a way to flexibly generate income as well as provide an easy way to get paid instantly for the work that they do.[1]

Uber has had a positive impact on part-time workers in developed countries by providing an alternative way to creating income and as a way to flexibly supplement existing income on a part-time basis (Chen et al. (2017); Cook et al. (2019)). Among families that experience job loss, ride-share platforms provide a way to supplement income during times of unemployment or income uncertainty (Farrell et al. (2019); Koustas (2018); Abraham and Houseman (2019); Jackson (2019)). In terms of the developing world, however, there is little evidence on whether drivers exploit, or are able to exploit this flexibility. In a developing country context, Uber's flexibility allows drivers to adjust more sensitively to income shocks and provides a way to get around binding constraints, such as production technology frictions that limit workers' abilities to marginally adjust their hours in traditional work environments. Just as Uber drivers can increase hours on

[1] https://qz.com/641250/ubers-new-debit-card-will-help-drivers-get-paid-instantly/

platform when there is high demand, they may also increase their hours due to driver level idiosyncratic shocks. This paper presents a way to estimate labor supply response to adverse rural shocks, both as a way of measuring the strength of rural-urban linkages and as a way of investigating whether drivers utilize the flexibility of the Uber platform to buffer against such shocks.

Uganda is an agro-ecologically diverse country (Ogwang et al. (2012)) with a large amount of internal migration. Migration mostly took place to the district of Kalangala, Kampala and the peri-urban district of Wakiso (Mukwaya et al. (2012)). Most migration was intra-regional, but Kampala has a large amount of in-migration, with roughly the same percentage of the population that migrate from within the same region (Buganda) as outside of it (27% to 24%, respectively). There is sufficient variation in people from different areas of Uganda in Kampala, while having the rest of the country staying within the same macro-regions when migrating. Most internal migration is undertaken for economic reasons and is one of the major pathways out of poverty for Ugandans (no. 36996-UG Poverty Reduction and Economic Management (2006); Mukwaya et al. (2012)). In a follow-up study using the 2016/17 Uganda National Household Survey, the reasons for migration were still mainly for income (39.7%), with over 30.8% of Ugandans taking part in rural to urban migration to Kampala (UBOS (2017)). Subsistence agriculture made up 42.6% of income for Ugandans, and remittances made up an average of 6.5% of income in 2012/13, with several sub-regions having a larger proportion such as the Central sub-regions (8.1% for both) and the Kigezi sub-region (8.8%). In the 2016-2017 period, remittances rose to making up 6.9% percent of income and some sub-regions relied on remittances for 8-10% of their income (such as the Central regions and the West Nile region). Ugandan households receive their incomes seasonally (from subsistence farming, 32.3%) or daily, mostly through non-agricultural enterprises (26.6%). Migration and remittances make up a non-trivial part of life in Uganda. From 2012 to 2016, the percentage of Ugandans taking part in subsistence agriculture as the main source of their income, stayed stable, but Ugandans tended to go from non-agricultural enterprises to wage employment or take part in migration/remittance behavior.

Although Uber's driver data provides high frequency logs of hours, earnings and the number of trips, it is limited to information that is easily logged and necessary for business practices. It is largely limited in terms of detailed demographic information that would be collected through traditional survey methods such as questions about remittance behavior, migrant status, or regional connection. I get around this problem by matching drivers to regions based on their surnames.[2] I use an external, nationally representative voter roll dataset for all of Uganda, which gives the names and locations of voter registration. This gives me the ability to link Uber drivers without the need for traditional survey methods. Once matched, I use a variety of agricultural drought indicators to act as shocks to the rural sector and see whether those shocks have any effect on driving behavior. Predictions of the model might be falsely matching drivers with weather events that they are not actually connected to. This would tend to attenuate the true effect. To account for this, I develop a switching regression estimator, inspired by Lee and Porter (1984) (Lee and Porter (1984)) that leverages the gradient boosting classification probabilities to account for misclassification.

My results show that Uber drivers in Kampala react to adverse weather shocks and use the flexibility of the platform to buffer against those shocks. I find that a one standard deviation increase in the intensity of agricultural drought leads to an increase of 5.1 hours online in the month of the event (a 6% increase over average hours). There are also heterogeneous responses to shock severity and income. Earnings increase by 92% if a severe shock occurred three months prior, but only 23% if it was mild. In order to give evidence for this being driven by agricultural shocks, I also aggregate and restrict the sample to the main and second growing seasons in the country, which is something the ASAP data allows me to do. When restricted to the main season, I find a large increase in hours as a result of higher drought intensity, with a smaller effect after the end of the growing season. There are similar, but smaller effects for the second growing season. This

[2]The ongoing COVID-19 pandemic has greatly limited the ability to conduct traditional in-person surveys. This may increase the reliance on mobile and internet based surveys, where there are issues with significant non-response bias. This paper shows how one can exploit already existing data to fill in the gaps using machine learning, that circumvents the issues that arise with such mobile and web surveys.

gives evidence to the idea that drivers use an increase in hours as a buffer after a bad harvest realization, not as anticipatory behavior at the end of the growing season.

This paper uses three distinct datasets for the analysis: (1) a dataset on Uber driver behavior, (2) a dataset of weather indicators used for constructing drought shocks and (3) a voter registration dataset with surname and locations. Figure 2.1 shows a visual representation of the data pipeline. I begin with the observation that the Uber driver data and the weather data cannot be merged. Weather data is at the pixel-time level and Uber driver data is at the driver-time level. If I assume that surnames in Uganda are representative of historical ethnic/tribal/linguistic ties and that internal migration is not so common as to completely dilute that signal, I can use surnames as a proxy for rural connection. In this case, I use an external nationally representative voter registration dataset that includes both surnames and locations of registration.[3] This provides the training data necessary to train a gradient boosting model that can generate predictions on regional connection, given a surname. Once I feed Uber driver surnames into the model, I can get a set of predicted regional memberships, as either a categorical variable or as a set of probabilities.[4]

The paper is organized as follows: Section 2 will give background on the empirical context about Uganda and Uber's presence there. Section 3 will explain how I capture adverse agricultural weather shocks by using a rich set of weather indicators using the European Commission's ASAP (Anomaly Hotspots of Agricultural Production) data. Section 4 will describe the classification of drivers into regions based on a nationally representative voter roll dataset. Section 5 will present the estimation strategy and how I plan to use the machine learning classification for estimation. This section will also present a switching regression

[3]The voter registration and weather data were at different levels of spatial aggregation, necessitating the need to aggregate to a level that was common between the two. This was done both for the surname predictions to align with the spatial levels in the weather dataset and because higher levels of specific spatial aggregation leads to better prediction accuracy.

[4]The surnames in the voter registration dataset were publicly available, but due to the data user agreement with Uber, it was not possible to observe the surnames of the drivers. As such, I created a pipeline where the trained model was transferred to an Uber data scientist and they would generate the predictions without me ever having access to the surnames.

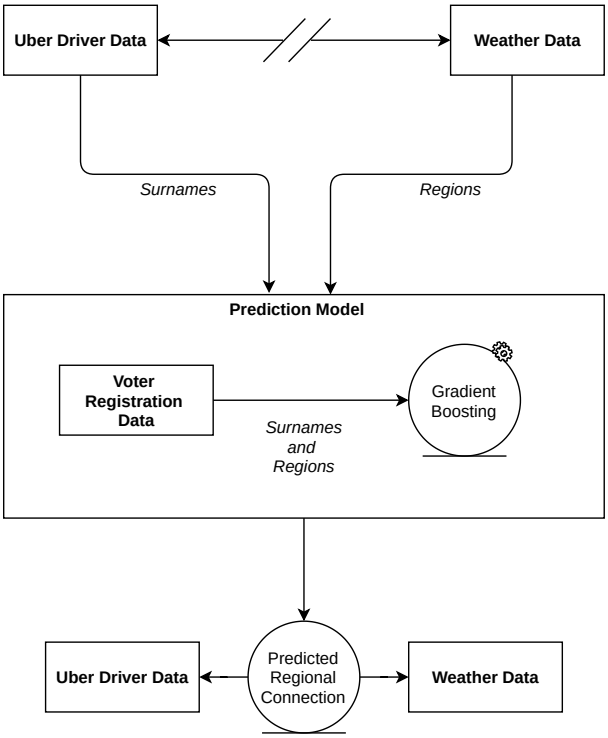

Figure 2.1. Data Structure Description

estimator that is robust to the misclassification in the machine learning predictions. Then I will present the results of both the misclassified OLS estimation alongside the switching regression estimation. Section 6 concludes.

2.2 Empirical Context

The data from Uber consists of a full dataset of Ugandan drivers in Kampala from 2016-2019. It is comprised of earnings and hours spent on the platform as well as some basic demographic characteristics on the drivers. It is an unbalanced panel of 14,455 drivers across 189 weeks. The data goes from May, 2016 to December, 2019 with drivers joining the platform over time. The week of a driver's first trip is given, but it is not possible to know if a driver has left the platform. Figure 2.2 shows hours online and the number of drivers on the platform to show how the platform has grown over time. Hours online for Uber can be defined as the time from turning on the app and becoming open to receiving ride requests to the time it is turned off. In rare cases, some drivers keep their phone on for long periods of time waiting for a ride request or they might forget to turn off the app

when they get home. This might create the illusion that a driver is online for 24 hours at a time. These cases are dropped from the sample. There is large growth from July 2016 to around July 2019 and there is a decrease in the number of drivers online. I define a driver as being on the platform if the date being considered is after the week of their first trip and before the last time they turned on the app. If there are gaps in between, I assume that their hours and earnings were 0 for that week. There is a clear drop in the number of drivers after April 2019, as I only fill in observations for drivers between their earliest and latest observations of positive hours. It may be that the drivers were still on the platform at that time, but I chose not to make assumptions about their status on the platform.

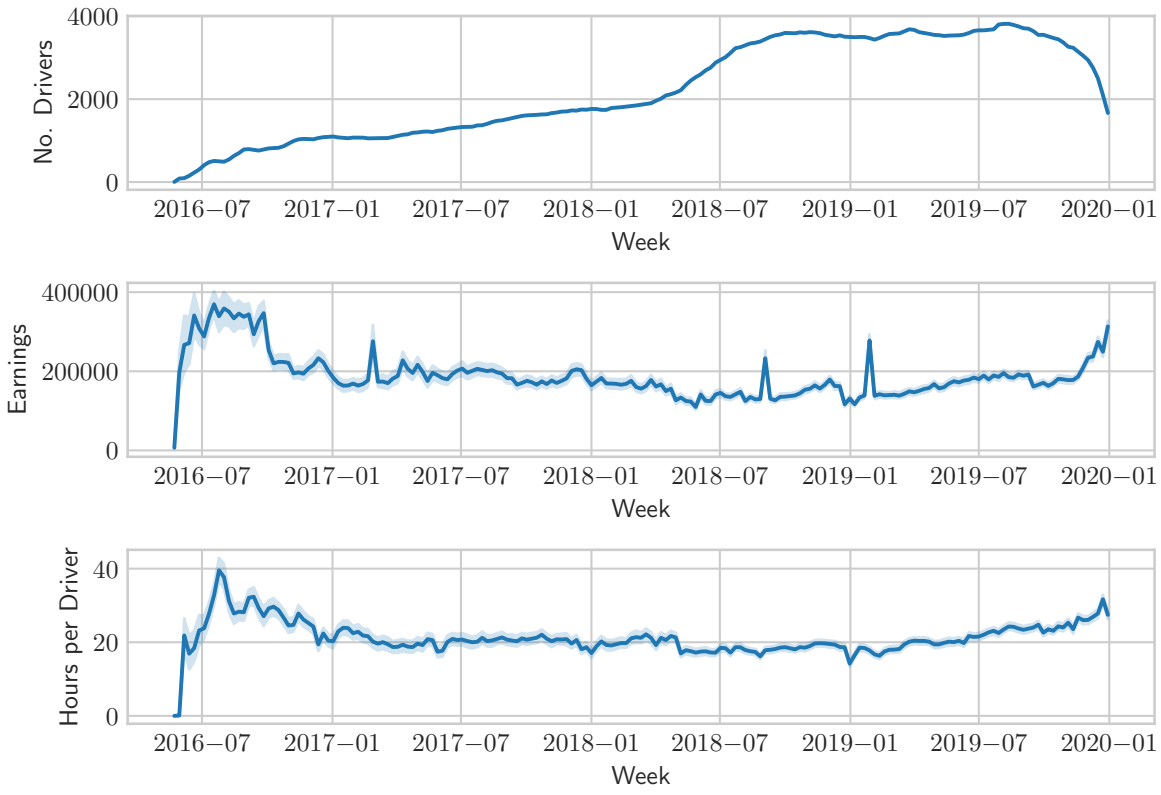

Figure 2.2. Uber Platform Growth in Uganda

Table 2.1 shows summary statistics for Ugandan drivers. Average hours online are around 21 hours. Most drivers are male (only 1 percent are female) and most vehicles are not shared by multiple drivers. Weekly earnings amount to around 45 USD a week.

Table 2.1. Summary Statistics of Ugandan Uber Drivers

	Mean	Std. Dev.
Weekly Earnings	156,067.91	140,444.26
Tips	47.84	143.97
Age Group	27.87	7.39
Lifetime Completed Trips	594.24	933.51
Hours Online	19.24	14.86
Female	0.02	0.13
Shared Vehicle	0.05	0.16

[a] Sample standard deviations in parentheses. Age, Lifetime completed trips, and Female are driver-specific. All other variables are driver mean across time. Earnings in local real 2017 Ugandan Shillings, with an exchange rate of ~3,600 UGX to 1 USD.

Figure 2.3 shows a density of weekly hours online, conditional on having nonzero hours. This shape is expected as most drivers use driving as a part-time job, driving for short periods of time. Although most drivers drive less than 20 hours a week, there are numerous drivers that can drive upwards of 80 hours a week.

2.3 Evaluating Extreme Weather Events

Uganda exhibits large variability in weather, both spatially and temporally, due to multiple geographic features such as inland water bodies, complex topography (Ogwang et al. (2012)), and El Nino (Kovats et al. (2003)). This makes Uganda a country with significant climatic variation given its small geographic size. In terms of agriculture, Uganda has to endure several types of adverse weather events that has the potential to hinder production and destroy yields, mainly, drought and flooding (Ogwang et al. (2012)). In this paper I will

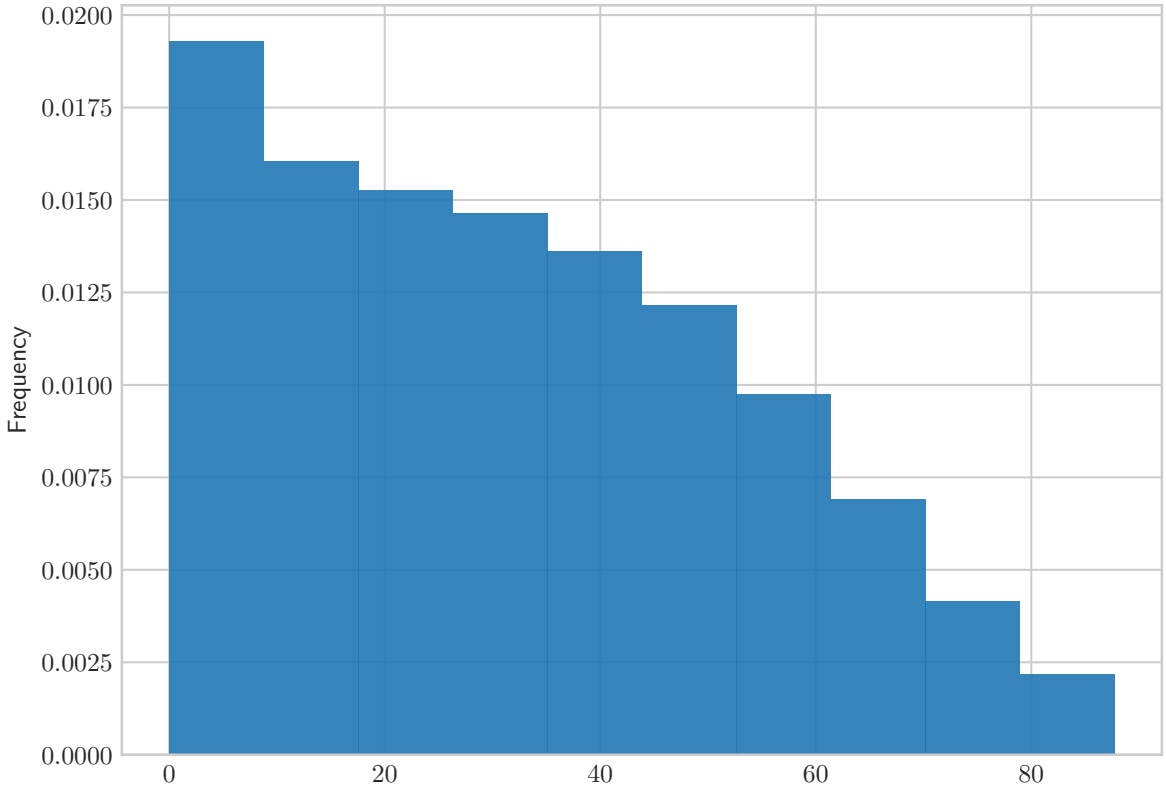

Figure 2.3. Histogram of Hours Online

use various weather indices to capture the intensity of biomass growth across the growing season which should capture both drought and flooding events.(Wilhite and Glantz (1985)).

I use a rich dataset of drought weather indicators used by the Anomaly hotSpots of Agricultural Production (ASAP) decision support system as part of the Monitoring Agricultural ResourceS (MARS) initiative at the Joint Research Centre (JRC) of the European Commission. ASAP provides a categorical warning system that tracks areas in the world that are at risk of agricultural drought events (Rembold et al. (2019)), but they make all constituent parts of their warning system available. The monitoring system uses various weather indices including temperature, precipitation and various NDVI (Normalized Difference Vegetation Index) metrics and utilizes remote sensing to ascertain the phenology of growing seasons and land cover masking of cropland versus rangeland. This phenology is then mapped to FAO crop calendars to ascertain the crop

being grown at that particular time. Understanding the timing of growing seasons is particularly important in our context, where I will need to do significant aggregation of spatial pixels to the regions I predict for drivers. ASAP data gives us the ability to filter out only the weather information that is pertinent for the question at hand. I will be able to restrict our analysis to particular growing and harvest seasons to get the best signal of agricultural drought shocks. The data that is provided by ASAP is at the 10-day or dekad level and uses the FAO Global Administrative Unit Layers (GAUL) dataset (GeoNetwork (2007)) for pixel aggregation.

GAUL is a unified set of geographic layers developed by the FAO to overcome any fragmentation across time when digitizing administrative unit layers globally. These administrative layers are particularly important because these are the delineations that I will use when predicting rural origins for our Uber drivers. They must be connected somehow to ethnic or regional boundaries that Ugandans find meaningful, whether that be through a shared language or a historical connection to a kingdom. In our particular case, I use two levels of classification, GAUL regions and a composite of GAUL and SAP Regions (an administrative delineation used by Uganda). GAUL and agro-ecological zones are often in line, making GAUL a good way of predicting climatic conditions (with minimal loss of surname predictive power compared to agro-ecological zones). SAP regions in Uganda roughly correspond (but not perfectly) to kingdoms in Uganda that have their own ethno-linguistic identity. Of course there are ethno-linguistic differences even within these kingdom or region boundaries, but in terms of estimation, there is an inherent tradeoff in the ability to find suitable climatic boundaries and ethnographic boundaries. Choosing one agro-climatic boundary gives higher precision in picking up weather events, with lower predictive power, and opposite for the other. I will use NDVI-based measures in our analysis with GAUL boundaries. Please see Appendix 2.7 for a more in depth information on the ASAP indicators and administrative boundaries used.

Figure 2.4 shows a figure of each of the weather indicators for the GAUL Composite regions. Where applicable, a dashed black line represents a level below which signifies the

start of a drought event. In Panel A, I see average air temperature, which is negatively correlated with the other three indicators. This is to be expected as higher temperatures are correlated with less precipitation and WSI. Temperature, by itself, however, is not a good indicator as different crops need a certain temperature to grow well. Across time, it seems that there are definite periods that would be candidates for agricultural drought shocks (areas where z-scores might dip below -1 and temperatures are high). But what's also important is to find an indicator that can discern across regions as well. The benefit of the current approach over, say, identifying on the time series alone, is that when there is enough variation *across* regions, I can identify the effects of weather on hours based on those that are and are not affected by the weather shock. Panel B for instance would not tend to be a good indicator to take advantage of this strategy as the SPI variables seem to be correlated across regions very strongly. Panels C and D exhibit such behavior: although Cumulative NDVI Z-score and WSI tends to be correlated across regions, there are periods of time where some regions are doing much better than others. I opt to use Cumulative NDVI Z-score in my analysis over WSI, however, because it is an indicator that maps onto agricultural practices more directly, including both precipitation and temperature effects together.

2.4 Classification of Regional Connection

Uganda is broken up into several kingdoms. These kingdoms are then further broken up into ethnic groups and then into clans. In some cases, Ugandans identify with their clan origins and their surnames connect back to those clans. Even siblings may have different surnames, but those surnames connect to a clan, not a family name.[5] In other cases, there are linguistic features that can identify a person from being from a particular ethnic group. Anecdotally, there are stories of how it is possible to pinpoint a person's regional origins (with differing levels of geographic accuracy) purely based on their surname.[6] In some ways, our classification procedure attempts to capture this ability, but in a data-driven

[5]See https://blog.caritas.us/blog-import/blog/ugandan-surnames-understanding-whats-in-a-name/ for an anecdotal account of this phenomenon.

[6]I only consider surnames in this cases since first names can often be biblical names that would muddle linguistic signal.

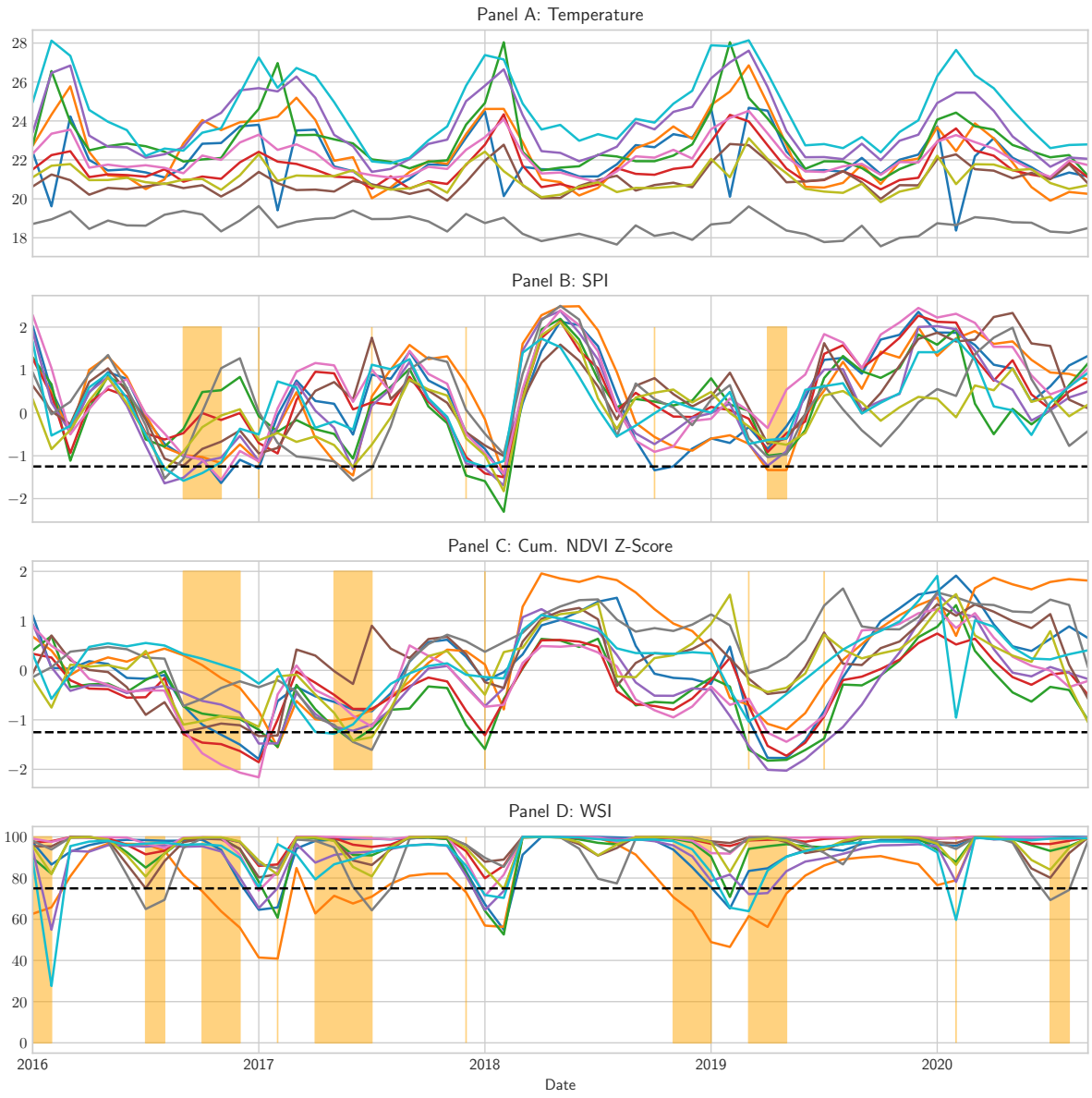

Figure 2.4. Weather Indicators with GAUL Regions

way.

The linguistic signal that I hope to capture is only as powerful as the amount of language diversity in Uganda. Although the official languages of Uganda are English and Swahili, any language can be taught in schools or used for legislative, administrative or legal purposes. Linguistic diversity in Uganda is a large topic of discussion in the areas of language planning (Namyalo and Nakayiza (2015); Altinyelken et al. (2014)).

One possible critique to this approach is that migration and the movement of refugees of the country might confound the classification. This is one drawback to the approach, as the signal of ethno-linguistic origin might be less confident and confound the classification. Another possible critique is the fact that given that Kampala is a melting pot of different ethnic groups from across Uganda, this will bias the classification towards putting more weight on Kampala. The classification, thus, does not include names from Kampala.

Finally, there is also the question of the strength of rural-urban ties and the extent to which drivers have strong ties to their places of rural origin and not Kampala or a region where their family might have moved to in the recent past. This is a fair critique, but anecdotal, as well as anthropological accounts suggest that Ugandans tend to have strong ties to their rural places of origin, whether that be through their family, community, landholdings, or other assets.

2.4.1 Voter Registration Data

The data used for classification is a nationally representative dataset of voter registration in Uganda for the year of 2016. This includes information on where they registered, and their surname. Based on the surnames, I constructed a gradient boosting model whose goal was to predict their region of registration based on the features of their surname. The way I break down their surname is also able to pick up on linguistic differences, so that I am not just predicting purely on the frequency of a name appearing in a particular region, but also on the *features* of the name.

3.1 shows summary statistics for the surnames and the labels used in the classification,

Table 2.2. Summary Table of Voter Registration Data

	District	Region	Surname	GAUL
Count	14,589,191	14,589,191	14,589,191	14,589,191
Categories	112	4	627,280	10
Most Frequent	Wakiso	Central	Akello	Lake Victoria Crescent
Occurrences of Most Frequent	898,940	4,195,548	103,412	3,442,242

SAP regions, agro-ecological zones, GAUL and the GAUL-SAP composite. There are 14,589,191 observations in the sample, coming from 112 districts. This accounts for almost all 134 districts in Uganda. There are 627,280 different surnames in the dataset, but with repetitions across districts. The GAUL-SAP composite and GAUL zones are the two main "target" variables that I classify for the purposes of the estimation strategy.

2.5 Empirical Strategy

In this section I will discuss the empirical strategy in regards to how the classification plays a role in the estimation. Suppose that a driver is connected to some region $r \in R$, for some set of regions R. If I had access to this information directly and I wanted to analyze the effect of some set of weather shocks W and controls C on a labor supply variable, y, I could run the following OLS specification:

$$y_{it} = W'_{irt}\beta + +C'_{it}\gamma + \varepsilon_{irt} \tag{2.1}$$

for driver i from region r at time t. I define $W_{irt} = \sum_{j \in R} 1\{r_i = j\} d_{jt}$ for some set of region specific weather shocks d. d can be a set of weather indicators, indicator variables for a shock or a set of distributed lag shocks. It is also possible to estimate an impact for each region, in which case I would estimate:

$$y_{it} = \sum_{j \in R} 1\{r_i = j\} d_{jt} \beta_j + + C'_{it}\gamma + \varepsilon_{irt} \tag{2.2}$$

Assuming that $d_{jt} \perp \varepsilon_{irt}\ \forall j \in R$ I can retrieve an unbiased estimate of the β parameters. However, I do not observe the true r, but rather a noisy signal $\hat{r}$, observed with measurement error. Since $\hat{r}$ is a categorical variable, I can think of the measurement error as $|R|$ dichotomous variables with an errors-in-variables problem, where $1\{\hat{r} = j\} = 1\{r = j\} + \nu_j\ \forall j \in R$. In this case, if I run either equation (2.1) or equation (2.2) with $\hat{r}$ in place of r, our estimates of β will be attenuated and the impact will be underestimated (Black et al. (2000); Aigner (1973)).

Since $\hat{r}$ is calculated using the voting registration dataset, I can use the training dataset to evaluate the accuracy of of the prediction. I can summarize the misclassification in a confusion matrix, where each cell describes the joint distribution between r and $\tilde{r}$. Denote this as $f_{\tilde{R},R}(\tilde{r}, r)$ for some probability density $f_{\tilde{R},R}$.

Table 2.3: The Confusion Matrix for the $r = 3$ Case.

	$\tilde{r} = 0$	$\tilde{r} = 1$	$\tilde{r} = 2$
$r = 0$	q_{00}	q_{01}	q_{02}
$r = 1$	q_{10}	q_{11}	q_{12}
$r = 2$	q_{20}	q_{21}	q_{22}

This is calculated post-fitting to evaluate the accuracy of the classifier. I can use this table to derive a conditional distribution, $f_{R|\tilde{R}}(r|\tilde{r})$. Let $\pi_{i,j} = f_{R|\tilde{R}}(r = i|\tilde{r} = j)$. Using the confusion matrix above, π_{ij} must be equal to:

$$\pi_{ij} = \frac{q_{ij}}{\sum_{k \in R} q_{kj}}$$
$$p_j = f_{\tilde{R}}(\tilde{r} = j)$$

Another feature of the machine learning model is that one of its outputs are a probability mass function for a driver's predicted probability of being in a given region. Let p_j as defined above, be this probability. Using these two features of the machine learning model, I can construct a new estimator that takes the prediction misclassification into account.

2.5.1 Misclassification as Switching Regression with Imperfect Sample Separation

In our data, I only observe (totally or partially) a sample from four random variables: Y_t, W_t, $\tilde{R}$. Y_t is a random variable for the outcome of interest, W_t is a vector of weather variables for each region of interest, $\tilde{R}$ is the predicted probability of an observation being classified from each region (Lee and Porter (1984)).

I then only observe the joint probability $f_{Y_t,\tilde{R}|W_t}(y_t, \tilde{r}|w_t)$ and I do not observe true membership r. I assume that r, $\tilde{r} \perp W_t$, which states that r and $\tilde{r}$ are independent of weather or drought, when conditioned on the names in our training dataset. This assumption is plausible in this setting, since there is probably no causal link between ethno-linguistic features in a name and weather. It could be that population dynamics are affected by weather and people with different places of origin will appear in different regions, but that would likely be from weather induced migration in the past and not within the time window as the study.

This condition implies that the prediction, $\tilde{r}$, is only a function of the name dataset, so $f_{\tilde{R}|W_t}(\tilde{r}|w_t) = f_{\tilde{R}}(\tilde{r})$. The same is true for $f_{R|\tilde{R},W_t}(r|\tilde{r}, w_t)$, so: $f_{R|\tilde{R},W_t}(r|\tilde{r}, w_t) = f_{R|\tilde{R}}(r|\tilde{r})$. Moreover, it is also important to note that, conditional on the true region of origin, Y_t is independent of the prediction, so that $f_{Y_t|\tilde{R},R,W_t}(y_t|\tilde{r}, r, w_t) = f_{Y_t|R,W_t}(y_t|r, w_t)$. I define $f_{Y_t|R,W_t}(y_t|r, w_t)$ as being normally distributed with mean $\varepsilon_t = y_t - \sum_{j \in R} 1\{r = j\} d_{jt}\beta_j$, and standard deviation σ.[7] d_{jt} is a Tx2 matrix with a column of 1's (for a constant) and a columns of drought statistics for region j, with β_j being a 2x1 vector of coefficients.

[7]Alternatively, you can also have a separate W_{rt} in the conditional distribution, but this was a clearer way to write the distribution and it only involves needing to estimate one parameter for the standard deviation than $|R|$ parameters.

I can then write the likelihood for one observation as in the following equation:

$$f_{Y_t,\tilde{R}|W_t}(y_t,\tilde{r}|w_t) = \sum_{j\in R} f_{Y_t|R,W_t}(y_t|r=j,w_t)\cdot f_{R|\tilde{R}}(r=j|\tilde{r})\cdot f_{\tilde{R}}(\tilde{r})$$

As a result, I have decomposed the joint distribution into three parts that I can observe:

1. $f_{Y_t|R,W_t}(y|r,w_t)$: The outcome as a function of the true region and covariates
2. $f_{R|\tilde{R}}(r|\tilde{r})$: The conditional distribution of the true region of origin, given the prediction. This can be derived from the so-called confusion matrix that gives the joint distribution, $f_{R,\tilde{R}}(r,\tilde{r},z)$
3. $f_{\tilde{R}}(\tilde{r})$: The predictions from the classifier.

Remembering the parameters as defined in (2.5), the likelihood for one observation for $R=3$ would be:

$$\begin{aligned}
&l(\beta,\sigma|y_t,y_t,\tilde{r}) = \\
&f_{Y_t|R,W_t}(y|r=0,w_t)\cdot \\
&[1\{\tilde{r}=0\}\pi_{00}p_0 + 1\{\tilde{r}=1\}\pi_{01}p_1 + 1\{\tilde{r}=2\}\pi_{02}p_2] + \\
&f_{Y_t|R,W_t}(y|r=1,w_t)\cdot \\
&[1\{\tilde{r}=0\}\pi_{10}p_0 + 1\{\tilde{r}=1\}\pi_{11}p_1 + 1\{\tilde{r}=2\}\pi_{12}p_2] + \\
&f_{Y_t|R,W_t}(y|r=2,w_t)\cdot \\
&[1\{\tilde{r}=0\}\pi_{20}p_0 + 1\{\tilde{r}=1\}\pi_{21}p_1 + 1\{\tilde{r}=2\}\pi_{22}p_2]
\end{aligned}$$

where $1\{\}$ is the indicator function. This likelihood function mirrors the imperfect separation switching regression estimator from Lee and Porter (1984) (Lee and Porter (1984)), unlike it, it does not simultaneously estimate the probability parameters. Those parameters are plugged in after the machine learning procedure is done. This likelihood function for one observation is then used to define the full likelihood, which is then maximized and estimates for β and σ are found.

$$L(\beta, \sigma | Y_{it}, W_{rt}, \tilde{r}_i, N) = \prod_{i}^{N} l(\beta, \sigma | Y_{it}, W_{rt}, \tilde{r}_i)$$

Simulation results using this estimator suggest that it is able to more precisely estimate the true effect of β in the presence of misclassification. This was tested using artificially created data with higher dependent variable variance, smaller differences between regimes, and higher correlation between weather shocks.

2.5.2 Econometric Results

The econometric results will illustrate two things: (1) that there is a significant increase in monthly hours as a result of an increase in drought intensity, and (2) that Uber drivers adjust their hours, sometimes within the same month of a drought event. Switching regression results will show that misclassification hides the true effect of the weather shock on labor supply, both in impact and *when* the impact takes place. I will first look at how hours are affected as a result of weather shocks, utilizing the variation across all drivers and calculating an average treatment effect across the whole sample.

The specification I will use is a distributed lag model:

$$\begin{aligned} Hours_{it} = \beta_0 + \beta_1 Drought_{r,t} + \beta_2 Drought_{r,t-1} + \beta_3 Drought_{r,t-3} \\ + Female_i + SharedVehicle_{it} + \alpha_i + m_t + \epsilon_{irt} \end{aligned}$$

where $Hours_{it}$ are hours online for driver i at time t. $Drought$ will be defined as the Cumulative NDVI Z-score, α_i are driver fixed effects and m_t are month fixed effects. $SharedVehicle_{it}$ is an indicator for whether the driver is sharing their vehicle with another driver. The reason for the lags are to capture changes in behavior through time and to understand the time it takes to adjust hours to weather shocks. This will be a way for us to evaluate whether the flexibility of Uber is able to allow drivers to adjust hours quickly enough. Note that since the outcome of interest is a z-score, a negative coefficient actually means that their drivers increase their hours as a result of drought. This is because higher drought intensity is coming from *decreases* in the z-score.

Figure 2.5. Hours Online and Weather Shocks

Figure 2.5 presents each GAUL region's cumulative NDVI z-score with the average hours online of each driver matched to that GAUL region. Orange bands represent times when the Z-score dips below -1.25, a cutoff used for severe drought. Cutoffs for the plot and the regressions were used so as to be sufficiently low, while also having enough support to estimate meaningful coefficients. Although these graphs include seasonal effects, the figures show that hours seem to drop when NDVI is higher and vegetation is greener, and begin to increase when NDVI dips. For instance, in the South Eastern region (fourth row, first column), the very large dip in NDVI (which goes beyond the cutoff value of -1.25) around the end of 2016, coincides with an increase in hours. Figure 2.5 also shows that different GAUL regions are affected at different times, adding strength to the overall identification. Generally, however, it's clear that the signal from these curves are noisy, which is what the switching regression framework is meant to account for.

Table 2.4. Effect of Drought Intensity on Hours Online

	OLS			Switching Regression		
	Hours Online	Hours Online	Hours Online	Hours Online	Hours Online	Hours Online
Intercept	75.762***	64.834***		81.338***	68.676***	
	(0.332)	(1.095)		(0.378)	(1.183)	
Drought Intensity t	0.712	-0.198	-4.069***	-0.690	-1.366	-5.134***
	(0.755)	(0.790)	(0.655)	(1.089)	(1.157)	(0.903)
Drought Intensity t-1	-0.918	-0.256	1.986**	-2.190	-2.144	1.182
	(1.138)	(1.211)	(0.916)	(1.773)	(1.926)	(1.565)
Drought Intensity t-2	2.187**	1.115	-0.490	5.512***	3.626**	0.152
	(1.110)	(1.194)	(0.896)	(1.672)	(1.848)	(1.615)
Drought Intensity t-3	-5.650***	-5.482***	-2.775***	-9.727***	-8.797***	-3.703***
	(0.723)	(0.757)	(0.628)	(0.971)	(1.037)	(0.916)
Shared Vehicle	37.707***	37.487***	26.650***	43.401***	44.583***	27.347***
	(1.464)	(1.462)	(1.649)	(1.766)	(1.756)	(1.806)
Female	-16.749***	-16.662***		-16.315***	-16.017***	
	(2.399)	(2.387)		(2.920)	(2.900)	
σ				92.052***	91.719***	63.609***
				(0.273)	(0.270)	(0.248)
N	69217	69217	69217	70067	70067	70067
Driver FE	No	No	Yes	No	No	Yes
Month FE	No	Yes	Yes	No	Yes	Yes

Note: Robust standard errors in parantheses. Data aggregated to the month level. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Month FE" refers to month fixed effects, "Driver FE" refers to driver fixed effects. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

2.5.3 Effect of Weather Shocks on Hours Online

First I will use the full sample, aggregated to the monthly level to see if there is an effect of weather on hours. Table 2.4 shows a table using GAUL-Composite predictions, with both OLS and switching regression estimates. Without month or driver fixed effects, I see that there are no significant effects from weather. But once I control for month-specific seasonality and then with driver fixed effects, there is a strongly significant effect in the same month and three months after a shock. When the switching regression approach is used, I see that these effects become even stronger. Drivers work five more hours in the month of an increase in drought intensity and then around almost 4 hours if that shock happened three months ago. This amounts to a 6 percentage point increase in hours driving for that month and speaks to Uber drivers using the platform to buffer against adverse weather shocks. Moreover, I can see that drivers change their behavior in the same month as the shock, signifying that the flexibility of Uber is being utilized, where alternative labor opportunities would not have let them change their behavior in this way.

There are also heterogeneous effects to the drought shock. Table 2.4 may be hiding heterogeneity to the response of a drought. In Table 2.5 I present a table with indicator variables for a mild and severe drought shock. I define a mild drought shock if a region has a Cumulative NDVI z-score of between -0.5 and -1.25. A severe drought is defined as a Cumulative NDVI z-score of less than -1.25. Note that in this case, an increase in hours in response to drought will be a *positive* effect. This gives us the ability to only consider drought events that are significantly severe.

Table 2.5. Heterogeneous Response to Drought

	OLS			Switching Regression		
	Hours Online	Hours Online	Hours Online	Hours Online	Hours Online	Hours Online
Intercept	75.248***	64.322***		77.039***	65.227***	
	(0.448)	(1.174)		(0.565)	(1.309)	
R-squared	nan	nan	nan	nan	nan	nan
	0.009	0.014	0.015	0.010	0.011	0.057
Mild Drought t	2.574***	0.317	5.200***	6.749***	1.410	8.564***
	(0.968)	(1.007)	(0.851)	(1.329)	(1.439)	(1.245)
Severe Drought t	4.530***	4.342***	6.474***	13.406***	10.891***	9.066***
	(1.528)	(1.632)	(1.358)	(2.165)	(2.530)	(1.975)
Mild Drought t-1	-4.082***	-1.155	-0.152	-10.792***	-5.034***	-2.506
	(1.171)	(1.225)	(0.937)	(1.646)	(1.833)	(1.741)
Severe Drought t-1	-3.830*	1.921	0.481	-9.238***	4.689	2.013
	(2.038)	(2.152)	(1.633)	(3.074)	(3.617)	(3.296)
Mild Drought t-2	-1.442	-2.339**	-0.421	-1.930	-3.042*	-0.090
	(1.131)	(1.181)	(0.905)	(1.547)	(1.727)	(2.345)
Severe Drought t-2	-0.357	-3.681*	-3.008*	-0.992	-6.045	-4.245
	(2.021)	(2.155)	(1.616)	(3.191)	(3.903)	(3.521)
Mild Drought t-3	3.441***	4.159***	3.427***	5.819***	6.518***	5.360***
	(0.907)	(0.941)	(0.797)	(1.217)	(1.349)	(1.243)
Severe Drought t-3	9.823***	9.912***	4.535***	19.300***	16.668***	5.693**
	(1.507)	(1.588)	(1.306)	(2.331)	(2.758)	(2.250)
Female	-16.551***	-16.422***		-17.863***	-17.335***	
	(2.397)	(2.387)		(2.495)	(2.479)	
Shared Vehicle	37.945***	37.699***	26.731***	37.456***	37.332***	25.661***
	(1.464)	(1.463)	(1.650)	(1.529)	(1.523)	(1.580)
σ				86.355***	85.996***	59.416***
				(0.225)	(0.222)	(0.206)
N	69217	69217	69217	69955	69955	69955
Driver FE	No	No	Yes	No	No	Yes
Month FE	No	Yes	Yes	No	Yes	Yes

Note: Robust standard errors in parantheses. Data aggregated to the month level. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Month FE" refers to month fixed effects, "Driver FE" refers to driver fixed effects. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP.Mild drought assumed to be any z-score below -0.5. Severe drought is any z-score that is below -1.25. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

Table 2.5 goes through the same set of additions of fixed effects as the previous tables. Once driver fixed effects and month fixed effects have been accounted for, it might seem that some of the significant effect sizes are negative, suggesting that drivers actually reduce hours after a severe drought. However, the switching regression results suggest otherwise. Misclassification hid the true effect, increasing effect sizes from small and positive or even negative, to large, positive and significant effects. Drivers increase driving by 8.5 hours in the same month if the drought is mild and a little over 9 hours if the drought is severe. If the shock was three months prior, drivers increase their hours by around 5.3-5.6 hours. What is peculiar is that there does not seem to be a difference in response. The difference in impact between mild and severe droughts are not significantly different. This may be due to the fact that drivers are constrained by the amount of demand they are able to serve and so their hours are capped.

For earnings, I use the inverse-hyperbolic sine of earnings in order to take into account periods of time when there are 0 earnings. Earnings are a more complicated variable to consider as they also depend on demand side shifters, not just driver-supply incentives. This makes it more difficult to interpret, but since I already have results from hours, we can use those insights to understand the effects of earnings. Earnings no longer change in the same period as a drought shock. There is a 25% increase in earnings with a two-month lag, and then a 45% decrease in earnings as a result of the three month lag. This may be due to demand-side effects that are constraining drivers and decreasing earnings.

Table 2.6. Effect of Drought Intensity on Earnings

	OLS			Switching Regression		
	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)
Intercept	9.898***	9.243***		9.892***	9.169***	
	(0.025)	(0.092)		(0.027)	(0.094)	
Drought Intensity t	0.238***	0.117**	0.043	0.280***	0.087	0.049
	(0.056)	(0.059)	(0.052)	(0.071)	(0.076)	(0.061)
Drought Intensity t-1	-0.003	-0.005	0.105	-0.107	-0.089	0.056
	(0.084)	(0.089)	(0.072)	(0.112)	(0.121)	(0.090)
Drought Intensity t-2	0.156*	0.193**	0.112	0.340***	0.395***	0.255***
	(0.082)	(0.087)	(0.070)	(0.109)	(0.117)	(0.087)
Drought Intensity t-3	-0.382***	-0.460***	-0.299***	-0.583***	-0.713***	-0.452***
	(0.054)	(0.056)	(0.051)	(0.067)	(0.071)	(0.058)
Shared Vehicle	4.148***	4.118***	3.653***	4.156***	4.121***	3.598***
	(0.034)	(0.036)	(0.093)	(0.034)	(0.036)	(0.087)
Female	-0.114	-0.100		-0.122	-0.124	
	(0.201)	(0.201)		(0.205)	(0.203)	
σ				6.215***	6.184***	4.611***
				(0.012)	(0.012)	(0.014)
N	69217	69217	69217	69955	69955	69955
Driver FE	No	No	Yes	No	No	Yes
Month FE	No	Yes	Yes	No	Yes	Yes

Note: Robust standard errors in parantheses. Data aggregated to the month level. Dependent variable inverse-hyperbolic sine transformed. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Month FE" refers to month fixed effects, "Driver FE" refers to driver fixed effects. Earnings in real 2017 Ugandan Shillings, with an exchange rate of 3,600 UGX to 1 USD. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

Now I look at the heterogeneous effects as I did with hours online. Table 2.7 uses the same definitions of mild and severe drought shocks. I find that there is considerable heterogeneity in response, but that these effects are likely confounded by demand side issues as well. The contemporaneous and three-month lag terms are significant, but earnings actually drop by 22% in the case of a mild drought, and increase by 44% in the case of a severe drought. Earnings increase by around 92% if a severe drought event occurred three months prior. These results are muddied, however, by strong negative effects both one and two months after, which do not occur when I use hours as a dependent variable. These are likely due to demand-side general equilibrium changes, which might stem from a decrease in rider income as a result of an increase in food prices, lower tourism after a shock or some other negative effect to rider income as a result of the rural-urban linkage.

2.8 shows regression results when the data is restricted to the main harvest season. The data is aggregated to the three month level and two lags of the Cumulated NDVI Z-score are constructed, one corresponding to the vegetation index three months prior (after harvest) and six months prior (in the beginning of the growing season). In this table, I can see how each region reacts to drought differently. In this case I see that there is a very large effect in the Central region (the same region as Kampala), with both 3-month and 6-month lags leading higher hours. The same is true for the East and West regions (although the West's six-month lag is not significant). Drivers are more sensitive to post-harvest season, than to growing season effects, suggesting that drivers only react once adverse harvests have been realized.

2.5.4 Second Season Restriction

When I restrict the data to the second season harvest, I only aggregate to the two-month level, since the second growing season is shorter than the main one. This can be seen in 2.9. From this table, I can see that although some regions have positive effects, both the impacts and the significance of the results are weaker, suggesting that drivers react much more strongly to the main season harvest. This is to be expected as most agricultural income is made after the main season harvest.

Table 2.7. Effect of Drought Shock Severity on Earnings

	OLS			Switching Regression		
	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)	$sinh^{-1}$(Earnings)
Intercept	9.990***	9.282***		10.043***	9.232***	
	(0.034)	(0.098)		(0.041)	(0.105)	
R-squared	nan	nan	nan	nan	nan	nan
	0.018	0.027	0.025	0.002	0.003	0.058
Mild Drought t	-0.188***	-0.258***	-0.170**	-0.210**	-0.444***	-0.288***
	(0.073)	(0.076)	(0.068)	(0.095)	(0.106)	(0.084)
Severe Drought t	-0.023	0.381***	0.257**	0.072	0.726***	0.443***
	(0.115)	(0.123)	(0.108)	(0.158)	(0.176)	(0.144)
Mild Drought t-1	-0.246***	-0.240***	-0.211***	-0.410***	-0.278**	-0.275***
	(0.087)	(0.092)	(0.075)	(0.118)	(0.135)	(0.097)
Severe Drought t-1	-0.122	0.033	-0.071	-0.228	0.243	0.021
	(0.152)	(0.162)	(0.130)	(0.220)	(0.249)	(0.185)
Mild Drought t-2	-0.308***	-0.351***	-0.273***	-0.466***	-0.683***	-0.520***
	(0.085)	(0.090)	(0.073)	(0.113)	(0.129)	(0.095)
Severe Drought t-2	-0.175	-0.254	-0.208	-0.258	-0.740***	-0.522***
	(0.149)	(0.159)	(0.127)	(0.208)	(0.238)	(0.176)
Mild Drought t-3	0.130*	0.258***	0.154**	0.163*	0.443***	0.236***
	(0.068)	(0.071)	(0.065)	(0.088)	(0.098)	(0.081)
Severe Drought t-3	0.798***	0.862***	0.493***	1.233***	1.556***	0.926***
	(0.108)	(0.114)	(0.103)	(0.135)	(0.149)	(0.129)
Female	-0.106	-0.085		-0.116	-0.108	
	(0.201)	(0.201)		(0.204)	(0.204)	
Shared Vehicle	4.156***	4.125***	3.652***	4.157***	4.123***	3.598***
	(0.034)	(0.037)	(0.093)	(0.035)	(0.037)	(0.087)
σ				6.209***	6.171***	4.606***
				(0.012)	(0.012)	(0.014)
N	69217	69217	69217	69955	69955	69955
Driver FE	No	No	Yes	No	No	Yes
Month FE	No	Yes	Yes	No	Yes	Yes

Note: Robust standard errors in parantheses. Data aggregated to the month level. Dependent variable inverse-hyperbolic sine transformed. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Month FE" refers to month fixed effects, "Driver FE" refers to driver fixed effects. Earnings in real 2017 Ugandan Shillings, with an exchange rate of 3,600 UGX to 1 USD. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP. Mild drought assumed to be any z-score below -0.5. Severe drought is any z-score that is below -1.25. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

Table 2.8. Effect on Hours Online (Main Season Only)

	OLS		Switching Regression	
	Hours Online	Hours Online	Hours Online	Hours Online
Drought Intensity after Harvest	-31.92***	-106.2***	-15.78	-106.19***
	(6.366)	(14.03)	(20.08)	(28.44)
Drought Intensity after Growing	-5.268**	-2.391	1.98	-2.39
	(2.670)	(3.963)	(5.78)	(9.31)
Female	-25.63		59.41	
	(23.93)		(67.03)	
Shared Vehicle	62.30***	62.40**	101.45***	62.40*
	(13.75)	(26.86)	(24.82)	(32.90)
Constant	195.2***	203.8***	188.73***	
	(2.702)	(3.650)	(5.46)	
Observations	6107	2725	1557	1557
Driver	No	Yes	No	Yes
σ			194.42***	192.99**
			(4.44)	(92.25)

Note: Robust standard errors in parantheses. Data aggregated to the three month level and restricted to being 3 months after harvest and 6 months after the growing season. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Driver FE" refers to driver fixed effects. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

2.6 Conclusion

In this paper, I use a novel and rich dataset of labor supply behavior from Uber to say something about rural-urban linkages. Even in the absence survey data on migration behavior, I was able to connect drivers to regions by using a gradient boosting model utilizing their surnames as features. This allowed me to connect drivers to regions outside of Kampala in order to quantify the strength of weather shocks on labor supply. But as with any prediction, there is the risk of misclassification. This can be a potential hindrance to seeing a true effect in a regression context. I presented a novel maximum likelihood estimator that used the predictions from the machine learning mode that was robust to misclassification in the estimates, yielding unbiased estimates.

Despite the fact that I did not have in depth survey data on drivers, where they came from, and why they might have migrated to Kampala, I was still able to say much about their labor supply behavior. Given the fact that the error in our classification procedure

Table 2.9. Effect on Hours (Only Second Season)

	OLS		Switching Regression	
	Hours Online	Hours Online	Hours Online	Hours Online
Drought Intensity After Harvest	1.330	-58.50***	-9.97	-58.50 ***
	(5.301)	(10.78)	(10.211)	(14.045)
Drought Intensity After Growing	-6.410	-2.364	-3.14	-2.364
	(4.954)	(10.05)	(10.21)	(30.52)
Female	-54.78***		7.66	
	(13.25)		(55.21)	
Shared Vehicle	82.99***	50.64**	99.66***	50.64***
	(11.46)	(20.05)	(18.73)	(35.73)
Constant	146.3***	140.1***	142.31***	
	(2.867)	(4.588)	5.17	
Observations	5099	2258	1923	1923
Driver	No	Yes	No	Yes
Sigma			171.67	156.91
			(3.32)	(16.21)

Note: Robust standard errors in parantheses. Data aggregated to the two month level. Analysis done with FAO GAUL boundaries disaggregated to 10 regions. "Driver FE" refers to driver fixed effects. Drought intensity measured in terms of cumulative NDVI z-score over the growing period of a crop identified by ASAP. Unadjusted R^2 reported, with Mcfadden's pseudo-R^2 reported for switching regression models. No intercepts reported for fixed effects models. * p<0.10 ** p<0.05 ** p<0.01

was not correlated with the independent variables of interest, I was able to show that drivers were affected by adverse agricultural drought shocks and that this persisted for several months after such a drought.

This approach can be used in economic contexts outside labor supply and weather shocks. Machine learning provides the ability to use existing data at a policymakers disposal (such as names in a census or voter rolls) to make policy relevant statements about the economy, without undertaking household surveys. As with everything, it is a tradeoff; it gives the researcher the ability to carry out analysis at the cost of having prediction error; but as I have presented, methods exist that can help the researcher deal with the fuzziness of machine learning in causal estimation.

Utilizing a rich dataset of weather indicators and varying sets of spatial aggregation, I illustrated the tradeoff between more aggregation, but with better prediction, and more

precise weather indicators, but at a loss of predictive power. This tradeoff is a fundamental one to this paper: if you choose a target variable that is too disaggregated or is motivated more by bureaucratic/administrative boundaries, then you risk losing the signal that helps you predict correctly. On the other hand, using regions that are too aggregated leads to imprecise weather effects; you might be losing your chance at observing important weather events or missing the timing of the reaction to those events.

Future work will focus on improving the method by which drought indicators are constructed. There is also much work to be done in regards to improving the machine learning model. Most of this paper focused on the misclassification caused across regimes, but more work needs to be done to understand how the misclassification can have a dynamic effect as well. This is interesting given how the effects of the lags changed from OLS to maximum likelihood. Although this paper focused on a substitute for household surveys, future work will include investigating the mechanisms behind the results in this paper by surveying people in Kampala directly. This can help to elucidate what drives the effects found in the paper.

Given that most migration into Kampala is from different regions of Uganda, the population of study in this case is particularly important for the study of development and poverty alleviation. In some parts of Uganda, remittance income makes up more than 10% of a household's income, making a policy targeted at a city potentially important to rural areas as well. If these results are even partly driven by remittance behavior, then Uber can be a good tool for vulnerable rural households to insure against adverse weather shocks.

2.7 Appendix: Weather Indicators

Broadly speaking, there are three types of weather indicators in the ASAP dataset:

2.7.1 Precipitation-Based Measures

There are two main precipitation based indicators that we will use: a 3-month Standardized Precipitation Index (SPI) and the water satisfaction index (WSI).

Cumulative rainfall and SPI are both derived by ASAP from either CHIRPS (Climate Hazards group Infrared Precipitation with Stations) or ECMWF (European Centre for Medium-Range Weather Forecasts). The 3-month SPI is a standard drought index that was developed in McKee et al. (1993) (McKee et al. (1993)) and is a z-score of a structurally fitted time series of precipitation (usually with a gamma distribution). This fitting is done to correct for that fact that rainfall does not exhibit normality at shorter time scales. The SPI, along with the Palmer Drought Severity Index are two standard ways of measuring drought before the advent of remote sensing techniques like NDVI. NDVI (Normalized Difference Vegetation Index) and SPI are better for different contexts. For instance, according to (Nakalembe (2018)), NDVI based measures perform better than precipitation based measures in the Karamoja region of Uganda for characterizing agricultural drought.

The WSI is a measure of availability of water to crops during the growing season. It uses both precipitation and evapotranspiration in order to estimate water available to the plant (Boogaard et al. (2018)). It ranges from a value 0 (no rain, dry soil and complete inability to sustain crops), to 100 (no deficit, and the crop always meets its water requirements).

2.7.2 Cumulative NDVI Z-score

The NDVI is an index using satellite data that measures the level of vegetation in a region, based on land reflectance from satellite data. It measures the level of "greenness" of the earth and the more it turns to yellow, the more there is a chance that a drought has occurred, as the plant matter on the earth is decreasing. ASAP's NDVI data is derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). For all NDVI based measures, first the cumulative NDVI is calculated which measures the increase of

vegetation from the start of the season of interest. Then a z-score is calculated using historical cumulative NDVI to get an idea of how typical that pixel's NDVI that year.

2.7.3 Harvest Indicator

In order to be more precise about when we posit that a driver will be able to react to a drought event, we also use FAO crop calendars to understand when harvest and growing seasons are. This way we can better test our theories of when a driver would be more likely to react as well as when rural households would be more likely to be in need of assistance.

The three main crops grown Uganda are maize, millet and beans. There are two growing seasons, as depicted in Table 2.7.3.

Season	Crops Grown	End of Planting	End of Harvest
Main	Beans-Millet-Maize	March	Mid June-July
Second	Maize/Millet	Oct.	Feb. (of following year)

Table : Harvest Seasons for GAUL-SAP Composite Regions.

Although there is variation in when growing seasons start, the harvest seasons are usually at the same times. This might be due to market conditions dictating when to harvest, but the end of the growing season and the end of harvest are lined up.

There are various theories of information transmission that could be applicable in our context. Since our data is noisy by design, we cannot necessarily say what a significant effect is telling us. But if our main hypothesis is to be believed, using the crop calendar in conjunction with the ASAP data can give us a more precise look into why that effect is significant. If a change in hours occurs directly after a bad harvest season, this might stronger evidence for the linkage being driven by agricultural drought.

2.7.4 Comparison of Different Indicators

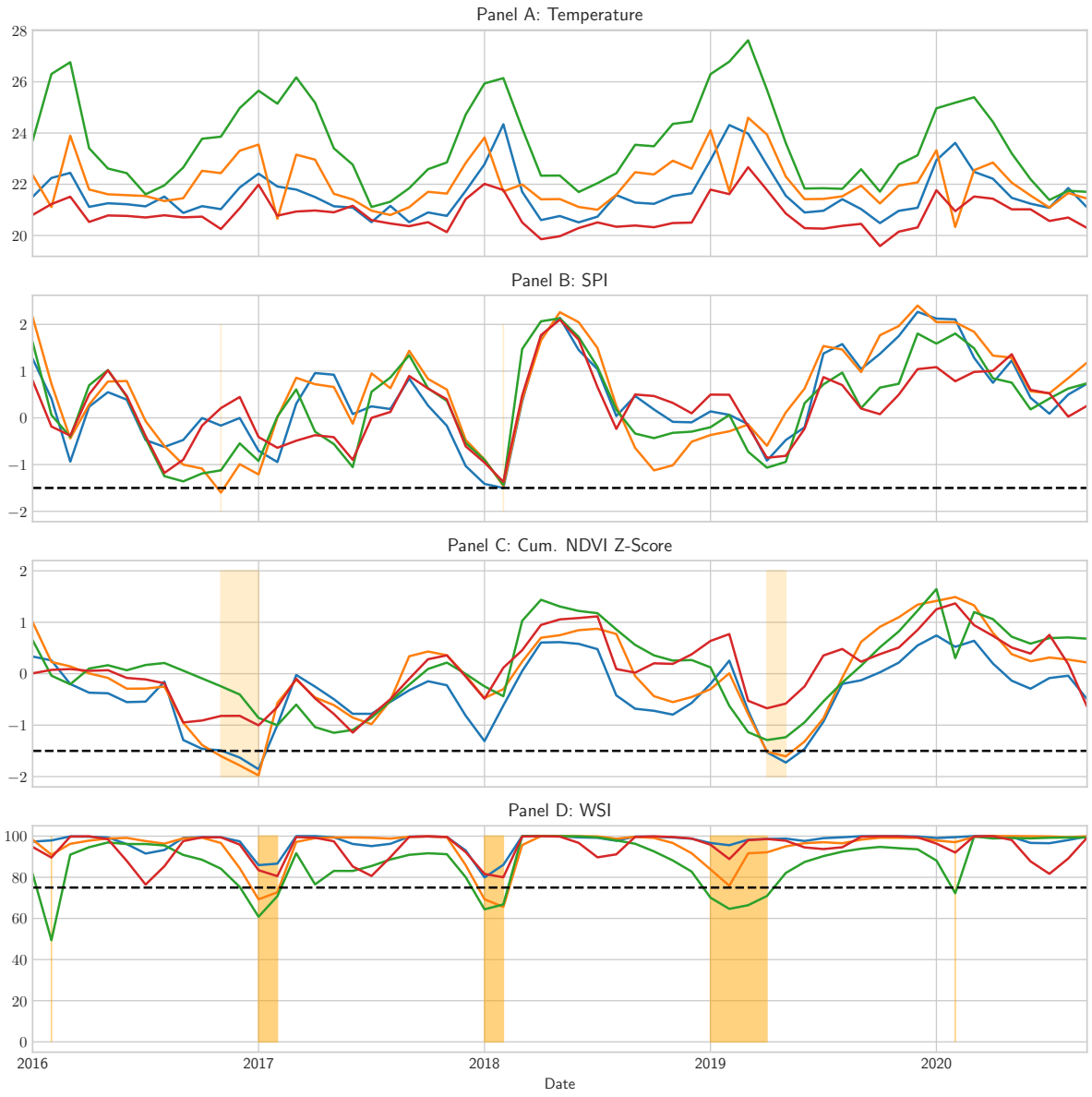

Figure 2.6. Time Series of Weather Indicators for GAUL-SAP Composite

2.7.4.1 GAUL-SAP Composite Regions

Figure 2.6 shows a figure of each of the weather indicators for the GAUL-SAP Composite regions. Where applicable, a dashed black line represents a level below which signifies the start of a drought event. In Panel A, we see average air temperature, which is negatively correlated with the other three indicators. This is to be expected as higher temperatures are correlated with less precipitation and WSI. Temperature, by itself, however, is not a good indicator as different crops need a certain temperature to grow well. Across time, it seems that there are definite periods that would be candidates for agricultural drought shocks (areas where z-scores might dip below -1 and temperatures are high). But what's also important is to find an indicator that can discern across regions as well. The benefit of the current approach over, say, identifying on the time series alone, is that when there is enough variation *across* regions, we can identify the effects of weather on hours based on those that are and are not affected by the weather shock. Based on this, however, Panel B for instance would not tend to be a good indicator to take advantage of this strategy as the SPI variables seem to be correlated across regions very strongly. Panels C and D exhibit such behavior: although Cumulative NDVI Z-score and WSI tends to be correlated across regions, there are periods of time where some regions are doing much better than others.

This can be illustrated by the shaded bands in Panels B-D. These represent periods of time where there is a drought in at least one region, with at least one area not being under drought. Compared to Panel B, Panel C has more such periods and those periods are long. Panel D has less such events, and while they start at the same times as Panel C, they have different durations.

2.8 Appendix: Details on Classification Methodology and Performance

Figure 3.1 shows an ethnographic map of Uganda. Even if classification is successful, our estimation strategy would not be successful, if Kampala did not also have a fair amount of ethnic heterogeneity. According to CIA World Factbook, Kampala is a city that includes

Figure 2.7. An Ethnolinguistic Map of Uganda (from: Ethnologue: Languages of Uganda)

people of many kingdom memberships, including the Buganda (the kingdom in which Kampala resides), Banyankole, Basoga and others, with significant variation, making the sorting of drivers a potentially powerful identification mechanism.[8]

In order to use the ASAP data, however, it was necessary to use FAO GAUL administrative layers, and this led to a need to aggregate data correctly for the classification so that the weather datasets could be used. This led me to create a SAP-GAUL composite which is an attempt to resolve differences between SAP regions and GAUL layers. Figure 2.8 shows the differences in these maps, which shows that there is only a small a discrepancy in the

[8]"Uganda - Central Intelligence Agency." https://www.cia.gov/library/publications/the-world-factbook/geos/ug.html. Accessed 8 Jan. 2020.

southwest of the country. A comparison between GAUL and agro-ecological zones is also presented in fig. 2.9. Although there is a marked difference between these delineations, they are broadly similar.

2.8.1 Tradeoffs in Classification

These different spatial delineations speak to a tradeoff when it comes to prediction precision and weather precision. Predicting into smaller delineations that are created based on climatic crop suitability (as agro-ecological zones and to a lesser extent, GAUL), would produce worse classification performance, while creating more precise weather shocks. On the other hand, predicting on the basis of ethno-linguistic boundaries (such as the GAUL-SAP Composite regions), would lead to higher predictive power but less precision for weather shocks, since those regions contain multiple agro-ecological zones. Any averaging of weather indicators would contain data from multiple zones that have different climatic conditions. For instance, one climactic zone might be experiencing a drought, while the others in a region would not, and we would fail to pick up that weather event.

Figure 2.10 Panel A shows a number of surnames in each region of Uganda. Since we'd like to identify rural origins outside the largest cities of Kampala, we drop the districts of Kampala and Wakiso.[9] The red bar shows the number of names that are dropped as a result of dropping these districts. The same is also shown for Panel B and agro-ecological zones. Although procedures exist to equalize the frequencies in each label of the target, we keep the frequencies as they are since the frequency of a surname in a target variable provides important information, especially in the case a nationally representative sample as this one.

2.8.2 Classification Method

Matching or predicting a person's ethnic or historical origins is not a new topic of study, albeit it is a controversial one. Surnames have been used since the 19th century to understand relationships between populations subgroups. But there are several issues involved in classifying a person into an "ethnicity" or group, some of which pose a

[9]Wakiso contains the second largest city in Uganda, Kira as well as being a popular peri-urban district.

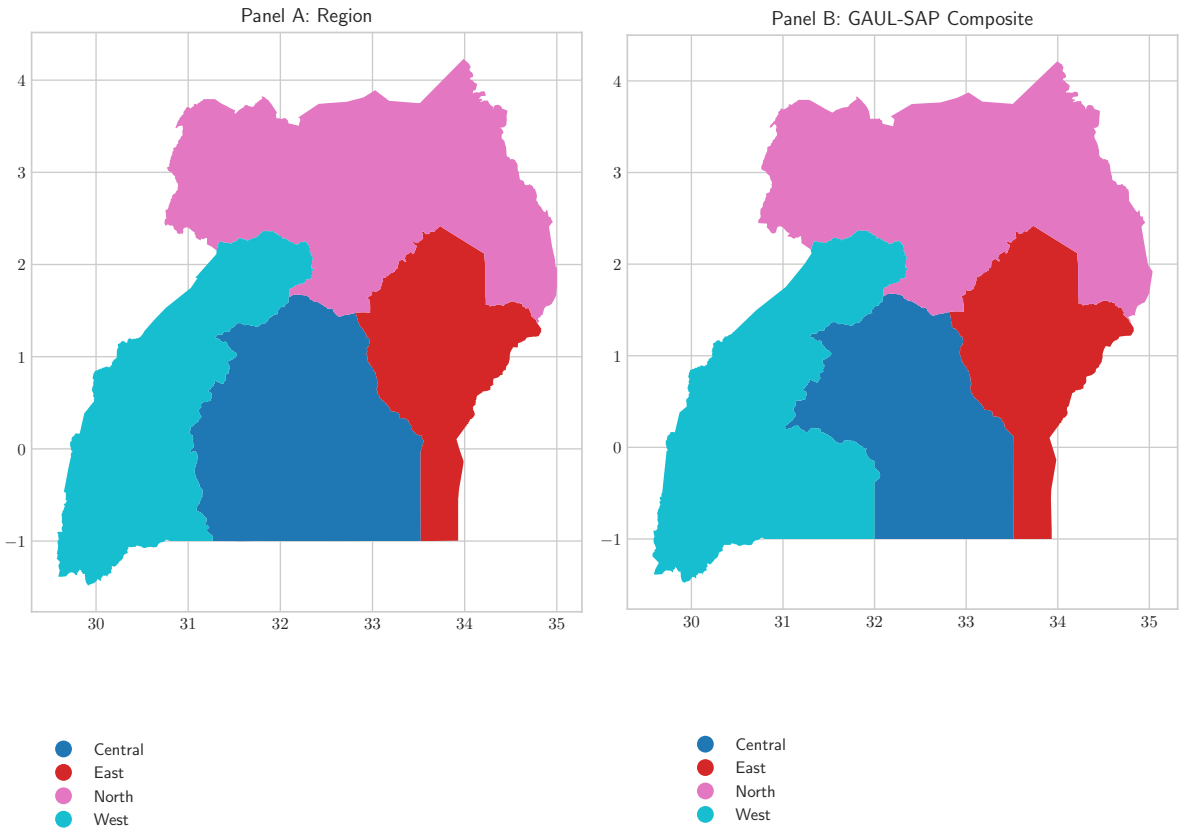

Figure 2.8. A Spatial Breakdown of Region and GAUL-Region Composite

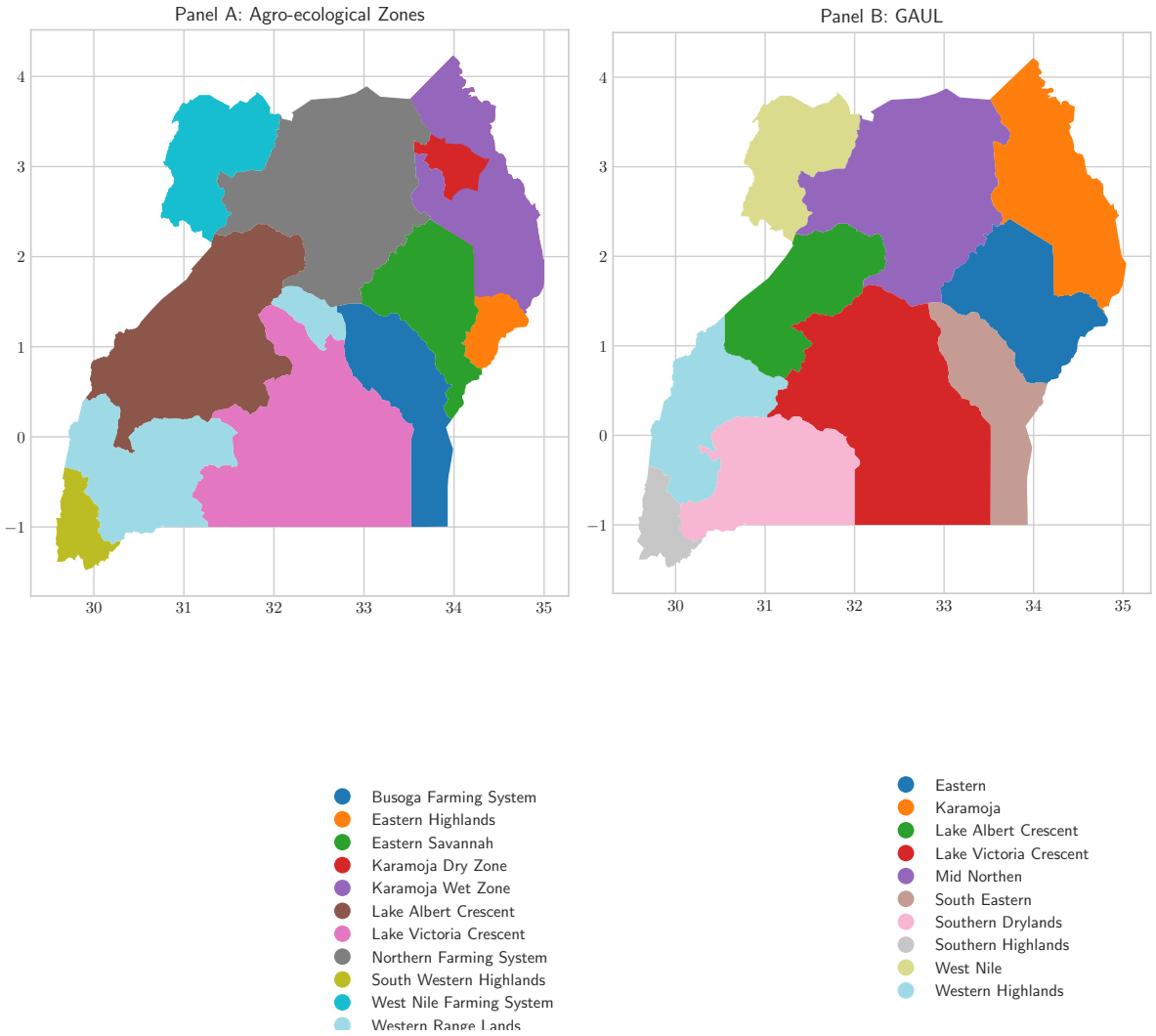

Figure 2.9. A Spatial Breakdown of GAUL and Agro-ecological Zones

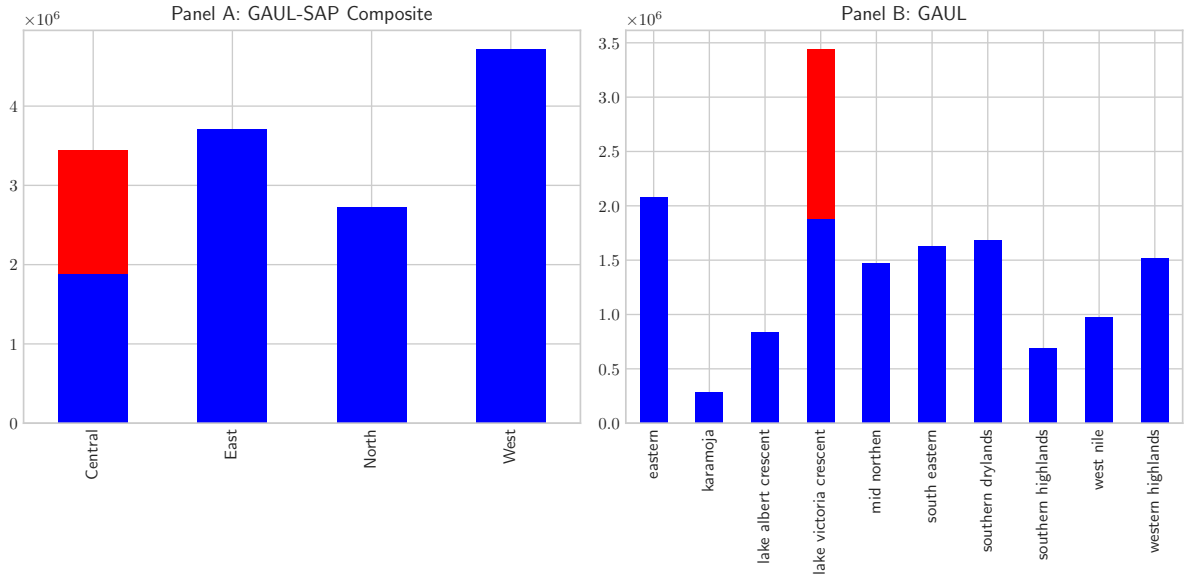

Figure 2.10. Frequency of Surnames by Region

methodological problem and some of which are more philosophical. For instance, the idea of race/ethnicity is often reduced to one variable, but it is actually part of a multi-faceted set of self-identifications. Moreover, ethnic categories are pre-determined based on institutional, historical or governmental criteria and not based on an open question that is then grouped based on response commonality. "Ethnicity" or any group membership is also something that can change through time based on an individual's self-assessment and it is problematic for it to be determined by bureaucracy or by a computer (Mateos, 2007b).[10] However, there is evidence to suggest that there is a correlational chain that connects surnames to rural places origin. In many places (including Uganda) surnames are passed down patrilinearly and are correlated with genetic variation in Y Chromosomes. Moreover, there is also correlation between variation in genetics and linguistic variation across the globe. In many cases, this can also lead to significant correlation with geography, although that connection is a weaker one. There are several factors that can confound the attribution of a surname to a particular geographic location, such as marriage, or

[10] As our analysis is not centered around finding the *impact* of ethnicity on our variable of interest, but only using it as a proxy for spatial connection, these issues are not as troubling. It is, however, important to note that the impacts that are found throughout this paper are not meant to equate labor market activity to ethnic origins, or make any genetic or cultural implications.

migration, but the hope is that there is enough correlation to proxy the rural-urban link.

There are various techniques that have been used for connecting surnames to ethnic/caste/geography that have been used in social science literature, such as using administrative data and leveraging its representativeness (Algan et al. (2016)), fuzzy matching search strategies (Mateos, 2007a), name classification software using public names data (Smith et al., 2017), probability estimation using frequencies (Bhusal et al., 2020), as well as hybrid approach using fuzzy matching and a Naive Bayes machine learning method (Monasterio, 2017).[11] However, these approaches, are often cumbersome to implement well if they were not designed for a particular language, or make assumptions that do not apply to text data (Naive Bayes techniques).[12]

In terms of using a similar exercise but with machine learning, there is a robust literature in computer science and computation linguistics around text analysis and language acquisition (Cavnar et al., 1994), as well as one for using similar and more cutting edge methods on names (Llitjos and Black, 2001; Qu and Grefenstette, 2004; Pervouchine et al., 2010). Some early methods included using so-called Naive Bayes techniques or SVM (Support Vector Machines).

We will use two principle methods to classify drivers into different regions or agro-ecological zones: what we call a table classifier and a machine learning classifier. The table classifier creates a frequency table of surnames and the percentage of occurrences in the target variable that they belong to. For instance, if the surname "Akello" appeared in the North, 3/10 times of all appearances of "Akello," then that surname would be given a 30% chance of appearing in the North.

While the table classifier is straightforward, classification can only occur with names that

[11] Onomap is a software that maps to ethnicity based on surname data, but is not available for any African country.

[12] Fuzzy matching algorithms are often built around a particular language, so that it is easier to tell when two words are similar. Since different languages have different common typos or letter substitutions, these fuzzy matching algorithms are very accurate and good at what they do. However, the current paper classifies names from languages for which such fuzzy matching algorithms do no exist, and basic fuzzy string matching is not very performative.

appear in the nationally representative data set, or that appear in a form exactly like the one in the dataset.[13] The other method used was to train a machine learning classifier on the name data. In this case, we used a gradient boosting classifier with tf-idf weighted n-grams. Gradient boosting uses many regression trees to find the best "rule" of how to discriminate between one group and another.

N-grams are a method used in computational linguistics in order to break up a word into constituent parts for analysis. It is a technique ubiquitous across linguistics, machine learning and natural language processing. For example, the name "Akello" as a 1-gram would be the one-letter variables of "A," "k," "e," "l," "l," "o."" As a 2-gram, we would have "Ak," "ke," "el," "ll," "lo," and then for 3-grams, 4-grams, etc... From then it is possible to count the number of occurrences of an n-gram in a particular word or to weight each gram in a specified way. One way is to use tf-idf (term-frequency, inverse document frequency) weighting of the n-grams to get a more informative set of variables to train on. Tf-idf weighting weights the frequency of a gram by how often that gram occurs in the whole dataset. This weights grams that occur very often lower, and weighs rarer grams more. This has the advantage of having the machine learning algorithm focus more on the part of the word that would give the most meaningful signal of coming from one label over another. The tf-idf transformed n-grams then become features or variables used for fitting a machine learning algorithm. In our case we ran the model with 1,2,3 gram features. Using higher order n-grams did not yield better results, but increased computation time considerably.

The results of the training yielded promising results. We use three metrics to quantify performance of each of the target variables: accuracy, precision and F-score. Accuracy can be simply defined as the number of correct times the trained model was able to correctly identify the class of the target in the testing data.[14] If we simplify to a binary classification framework, where there were only two classes to classify into (cat or dog, for instance), we

[13] Although fuzzy-matching was attempted with the table classifier, given that there is no matcher for any Ugandan language, the performance of this was too slow and often gave strange results.

[14] This means, for instance correctly identifying that the surname comes from the North, the South, the Center or the East in terms of region.

can formally write accuracy as:

$$Accuracy = \frac{TP}{TP + TN + FP + FN}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false-negative. This refers to how the model predicts the model into each class.

Precision and F-score were originally designed for binary classification models and are a standard metric for quantifying classification performance. But since there are four GAUL-SAP regions and ten GAUL layers, precision and f-score are redefined to include multiple classes. For binary classification, precision is:

$$Precision = \frac{TP}{TP + FP}$$

Precision tells us the probability correctly predicting the class, conditional on all the times it predicted an observation into that class. For multi-class classification, we calculate precision in a "one vs. all" fashion, where we treat the class of interest as one class and all other classes as another class. This reformulates the problem into C binary classification problems where C is the number of classes. Multi-class precision is then calculated as a weighted average of these C binary precision metrics.

The F1-score can be defined as the geometric mean of precision and recall, another metric used in classification.[15]

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The F1-score is a better metric than accuracy as it is robust to false negatives and false positives.

[15]Recall can be defined as $TP/TP + FN$ and tells us how robust predictions are to false negatives. In this case, we omit the recall score, but they were comparable to precision.

2.8.3 Training Results

For our target variable, we trained the model on 75% of the data, leaving out 25% for testing. The results can be seen below in table 2.8.3. 5-fold cross-validation was also run and yielded similar results.

Target	Accuracy	Precision	F-Score
SAP Region	0.79	0.79	0.79
GAUL-SAP	0.77	0.76	0.76
AEZ	0.68	0.67	0.67
GAUL	0.66	0.65	0.65

Table : Testing Score of XGBoost Classifier on Target against Ugandan Surnames.

Chapter 3

Disentangling the Effect of Ethnic Identity and Location on Voting Preferences in Uganda

3.1 Introduction

Economic development depends on political outcomes; in many developing countries ethnic identity has a significant association with voting behavior (Alesina and Zhuravskaya, 2011). There is a large body of literature that studies ethnic voting preferences and their impact on election outcomes (Wolfinger, 1965; Huber, 2012). These voting models posit that voters in places such as East Africa tend to vote for a candidate of the same ethnic background, even if that candidate takes part in corrupt politics or kickbacks (Carlson, 2015). In fact, many models of ethnic voting suggest that ethnic identity supersedes economic or other considerations (Berge et al., 2020; Eifert et al., 2010). A large literature also states that high ethnic fractionalization also leads to less public goods provision (Alesina et al., 2003). But many of these papers do not examine the extent to which location can drive voting preferences, despite ethnic identity and location being plausibly correlated due to shared political and economic beliefs (Ichino and Nathan, 2013; Kramon et al., 2019). Depending on where one lives, this might lead to voting preferences becoming more in line with the

majority group identity in a region, or create even more division, shedding light on the question of whether ethnic identity is the primary determinant of voting preferences in these contexts.

This paper uses a novel machine learning technique to derive a measurement of ethnic identity and then uses that to explore voting behavior. In our context, we take ethnic identity to mean being from a region based on surname features. Since our analysis uses a nationally representative sample, we can explore the role of location on voting behavior and estimate heterogeneous effects by location of residence. We estimate these effects by pairing voter registration data with election outcomes from polling stations across Uganda. We overcome the challenge of measuring ethnic identity by using a gradient boosting classifier that exploits surname variation in ethnic and linguistic groups across regions of Uganda. In particular, we use the letter sequences in each individual voter's surname to predict their ethnic identity, as in Michuda (2020). We then use this prediction of ethnic identity as the explanatory variable of interest in a regression model with the outcome that measures support for the incumbent president in the 2016 general election.

The innovation of this study is the use of data from across Uganda to associate each individual's surname with a region of the country and by leveraging machine learning to predict an ethnic identity for each individual in the study. This innovation overcomes the challenge of measuring ethnicity at a large scale, while also utilizing variation from a nationally representative data set. Other studies deal with measuring ethnicity by using observational or experimental data from relatively small geographic areas (Berge et al., 2020; Ichino and Nathan, 2013). Studies of small geographic areas have limited external validity. Our study allows us to shed new light on the relative roles of identity and location in explaining voting behavior using a nationally representative sample from a developing country.

3.2 Background

3.2.1 Ethnic Mixing and Internal Migration in Uganda

Uganda is considered one of the most ethnically diverse places in the world and is historically broken up into several kingdoms. These kingdoms are then further broken up into ethnic groups and then into clans. As an illustration, the official languages of Uganda are English and Swahili, but any language can be taught in schools or used for legislative, administrative or legal purposes. For instance, in the area of language planning, there is a vast literature on linguistic diversity in Uganda (Namyalo and Nakayiza, 2015; Altinyelken et al., 2014).

Migration and ethnic mixing in Uganda is driven by a complex set of geographic, economic and multi-lateral political factors. Uganda is made up of over forty ethnic groups and has been the center of many massive population movements. This makes ethnicity's and ethnic diversity's effect on voting an important factor when discussing Ugandan politics.

Internal migration in Uganda has historically been undertaken for various reasons, both economic, climatic and forced. Northern Uganda and Southwestern Uganda, for example, have been the center of some the worst humanitarian crises in history. Almost two million people from the Acholi region have been displaced to different parts of Uganda and abroad due to over twenty years of armed conflict (with the current president being an integral part of ending the conflict there). To the east, the Karamoja region, which has traditionally been an agro-pastoral population, experienced decades of both political insecurity and natural shocks that has led to migration into other parts of Uganda (OIM, 2015).

Internal migration in Uganda takes place mostly to the districts of Kalangala, Kampala and the peri-urban district of Wakiso (Mukwaya et al., 2012). Most migration is intra-regional, with the exception of Kampala, which has the same percentage of the population that migrate from within the same region (Buganda) as outside of it (27% to 24%, respectively). There is sufficient variation in people from different areas of Uganda in Kampala, while having the rest of the country staying within the same macro-regions when migrating.

A large amount of internal migration is undertaken for economic reasons and is one of the major pathways out of poverty for Ugandans (no. 36996-UG Poverty Reduction and Economic Management, 2006; Mukwaya et al., 2012). Other reasons for internal migration include marriage with inter-ethnic marriage becoming more commonplace through the years.

Figure 3.1 shows an ethno-linguistic map of Uganda. There is significant variation across the country, but in many cases, the dominant ethnic group coincides with the dominant kingdom present in that area. For the most part, kingdoms in Uganda are captured well using the Ugandan regions of North, East, West and Central. The regions in Uganda are made up of four macro-regions, roughly corresponding to larger ethno-linguistic boundaries, making it a promising target for classification, illustrated in Figure 3.2.

3.2.2 Presidential Politics in Uganda

Uganda has a presidential democracy with a unicameral parliament. Presidential and parliamentary elections are held every five years. The president is chosen by simple majority. Uganda's democracy is nascent, having its first election in 1962, but was followed by a period of dictatorial reign. The next presidential election was held in 1980. However, this regime was toppled by a military coup in 1985. Yoweri Museveni waged an armed rebellion in the Ugandan Bush War and took power along with his National Resistance Movement (NRM) in 1986. He has been president ever since.

In 2016, there were several candidates for the presidential election, including major opposition candidates, former military personnel, and academics, but Museveni still won by over 60% of the vote. The ability of Museveni to hold on to power has been a topic of discussion in political science and is often attributed to authoritarianism or neo-patrimonialism, but the 2011 elections signaled a change in approach for the Museveni regime in its use of "soft power" (Golooba-Mutebi and Hickey, 2016; Tangri and Mwenda, 2010). Museveni's ability to co-opt minority groups and listen to popular concerns has allowed him to stay in power and efficiently manage his rivals. The so-called "poverty tours" undertaken in the North of the country strategically allocated development funds

Figure 3.1. Ethnolinguistic Map of UgandaSource: Ethnologue

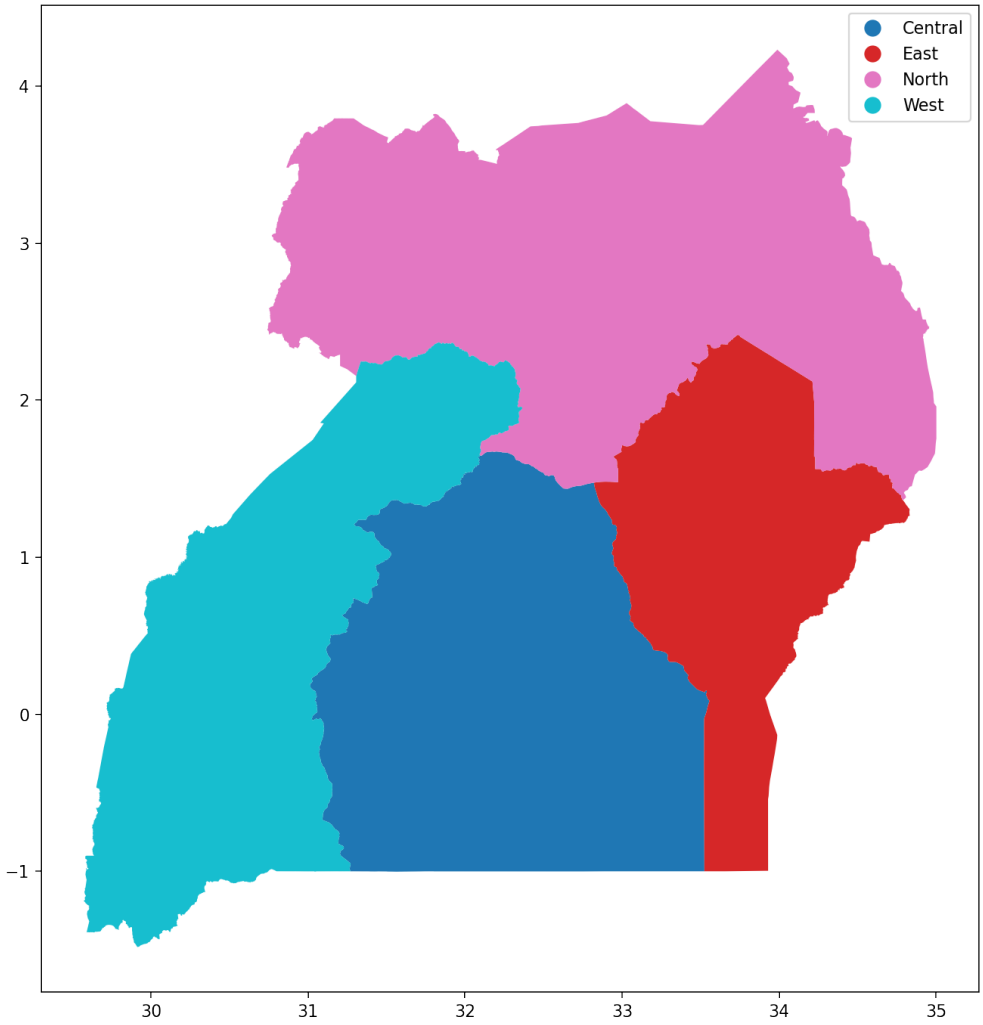

Figure 3.2. A Spatial Breakdown of SAP Regions in Uganda

to areas where his support was weakest. This strategy was so successful, that there was little need for actual repression. The same approach was also carried out again in 2016. Despite this, Museveni's tactics illustrate a rift that he has created across the country since he came to power; the Bush Wars brought support from the Central Region with the Baganda ethnic group, and being from the Southwest, he created an elite composed of the Central and West regions of the country. His rule over the years has led to rifts forming between the North-East parts of the country and the South-West elites.

Voting behavior along ethnic lines is a reality in Ugandan elections. In the early 2000s, 16% of Ugandans identified first and foremost with their ethnicity (Eifert et al., 2010). To study the effect of these ethnic tensions, there are two main issues: (1) finding data on ethnic membership with large scope, and (2) separating out the pure effect of ethnic identity that comes from feeling a sense of connection to those with the same ethnicity, and the effect that comes from having an ethnicity different from the majority ethnicity where they live. Our machine learning method allows us to overcome the first concern by being able to use a nationally representative data set, and by using those predictions. We can overcome the second concern by looking at how voting behavior of predicted identity changes across regions of residence.

3.3 Data

We study the 2016 presidential election in Uganda by constructing a data set from two nationally-representative data sets from the Ugandan Electoral Commission. Our outcome of interest is vote percentages for the incumbent, Yoweri Museveni. Our explanatory variables of interest are predicted ethnic composition, which we construct with individual-level voter registration data from polling stations in Uganda.

We use polling station-level data on the percentage of the vote that Museveni received and the amount of voter turnout among voters registered to vote at the polling station. The sample comprises around 18,000 polling stations. The average number of votes cast were 368 , of which 5 were counted as invalid.

Table 3.1. Summary Table of Voter Registration Data

	District	Region	Surname
Count	14,589,191	14,589,191	14,589,191
Categories	112	4	627,280
Most Frequent	Wakiso	Central	Akello
Occurences of Most Frequent	898,940	4,195,548	103,412

3.3.1 Voter Registration Data and Machine Learning

For each polling station in the country, we observe each individual registered to vote at that polling station. These individual-level data include each registered voter's location of registration and surname, as well as their age and gender. Our explanatory variable of interest is ethnic identity, which is closely tied to the surnames of individuals in Uganda; for an overview of why using surnames is a viable approach for connecting Ugandans to regions, please see Michuda (2020).

3.3.1.1 Voter Registration Data

Table 3.1 shows summary statistics for the surnames and the labels used in the classification, Ugandan Regions. There are 14,589,191 observations in the sample, coming from 112 districts. This accounts for almost all 134 districts in Uganda. There are 627,280 different surnames in the data set, but with repetitions across districts.

We constructed a gradient boosting classifier to predict each individual's region of registration based on the features of their surname. We break down the name into n-grams which is able to pick up on linguistic differences, so that we are not just predicting purely on the frequency of a name appearing in a particular region, but also on the *features* of the name. Due to Uganda's particular ethnic and migratory history, our machine learning classifier will be able to capture ethnic information, not just geographic information. Note that even despite this, our classifier does not capture the full ethnic diversity of Uganda, but there is enough of a signal for the purposes of the analysis.

The predictions from the machine learning algorithm allow us to create a set of probabilities

for each registrant's ethnic identity based on their surname, as in Michuda (2020). In contrast to that paper, we then merge those predictions back into the voter registration dataset, and average these probabilities across voters in a polling station. The data set for our analysis merges the election outcome data and the voter registration data averaged at the polling station-level. This sub-section presents the summary statistics and geographic variation for the main variables of interest in those data sets: voting behavior and ethnic identity.

Table 3.2 shows election results by candidate. The means in the table largely correspond to the official election results in the country. The current president, Museveni, won with over 60% of the vote, with the second being Kizza Besigye, who ran for the main opposition party, "Forum for Democratic Change" and lost with 30% of the vote. In general, politics for the last thirty years have centered around being pro- and anti- Museveni, with various opposition parties appearing in order to beat him.

Table 3.2 also shows summary statistics for average demographics of polling stations in the data set. Our main explanatory variable of interest is ethnic identity, which we summarize by the probabilistic ethnic-linguistic fractionalization (pELF) index averaged across polling stations. The ELF is a standard measure of ethnic fractionalization used in the literature on ethnic division (Alesina et al., 1999). In our case, rather than using the shares of each ethnicity, we use our predictions as those shares and calculate the probability that any two individuals have the same ethnic identity. The mean pELF is 0.40, which is lower than in other studies (Posner, 2004), but this may be due to our level of aggregation hiding further ethnic diversity. Table 3.2 shows demographic information for age and gender, which we use as control variables in our analysis; the average voter was around 37.0 years old and was slightly more likely to be a woman (52.2 %).

3.4 Regression Analysis on Voting Outcomes

We study the effect of predicted ethnic identity on support for the incumbent president using regression analysis. We observe voting behavior not for individual voters, but rather

Table 3.2. Summary statistics of Election Results

	Mean	Std. Dev.	Min	Max	N
Age	37.04	2.09	24.50	47.90	26814.00
Female	0.52	0.10	0.01	1.00	26814.00
Votes	368.27	129.55	0.00	946.00	26814.00
% Invalid	4.74	4.14	0.00	50.25	26814.00
Prob. ELF	0.40	0.12	0.12	0.75	26814.00
Voting Outcomes					
% Museveni	60.71	20.25	0.00	100.00	26814.00
% Besigye	34.58	18.88	0.00	95.50	26814.00
% Mbabazi	1.48	3.16	0.00	66.73	26814.00
% Bwanika	0.91	0.98	0.00	10.84	26814.00
% Baryamureeba	0.55	0.77	0.00	11.11	26814.00
% Kyalya	0.45	0.77	0.00	37.34	26814.00
% Biraaro	0.27	0.60	0.00	44.05	26814.00
% Mabirizi	0.25	0.50	0.00	37.17	26814.00

Note: National election data attained from Election Commission of Uganda for 2016 national election results. Ethnic fractionalization calculated using the Probabilistic ELF formula, which calculates the average probability that two voters in a polling station are from the same region.

for polling stations. Each polling station has data on vote shares for each presidential candidate in the 2016 general election, which our regression model pairs with data on predicted ethnic identity, age, gender, the percentage of voter turnout, and the percentage of invalid votes, aggregated across voters registered at the same polling station to give ethnic composition in a polling station.

There are challenges to identification in our regression approach. Our analysis hinges on increases to ethnic diversity or to an addition of a non-native ethnic identity to a particular polling station. This means that our identification is dependent on those that moved from one region to another. This introduces potential selection bias into our estimates, since these movers' voting preferences are likely correlated to some individual, economic or social factors that drove their decision to move. In order to deal with this, we add control variables such as gender, age and voter turnout, as well as sub-county fixed effects.

Figure 3.3 is a map of Uganda with each marker corresponding to a sub-county and the color corresponding to each the percentage of votes that Museveni received. Bluer hues correspond to sub-counties where Museveni won the majority and redder hues are where he lost. The most pro-Museveni areas in Uganda are those in the Northeast and Southwest, but it is clear that Museveni had support in more regions of the country. The largest opposition was in the North, with parts of the East and the Central region, mostly around Kampala. These are all unsurprising, due to the North and East's marginalization during the Museveni regime and the fact that Kampala is made up of a large number of urban dwellers and technocrats (where Museveni's post-2011 inflation could be felt and the IMF's rebuke of his tactics disheartened many technocrats).

Figure 3.3 shows that support for presidential candidates in Uganda is strongly correlated with the region of the polling station. But the question is whether this can be attributed to majority ethnic composition driving that behavior, or whether it is location driving those choices. We identify the effect of ethnic identity itself on voting outcomes using variation in the ethnic composition of polling stations within the same sub-county. This is a potentially important mechanism as it can help us also disentangle historical support

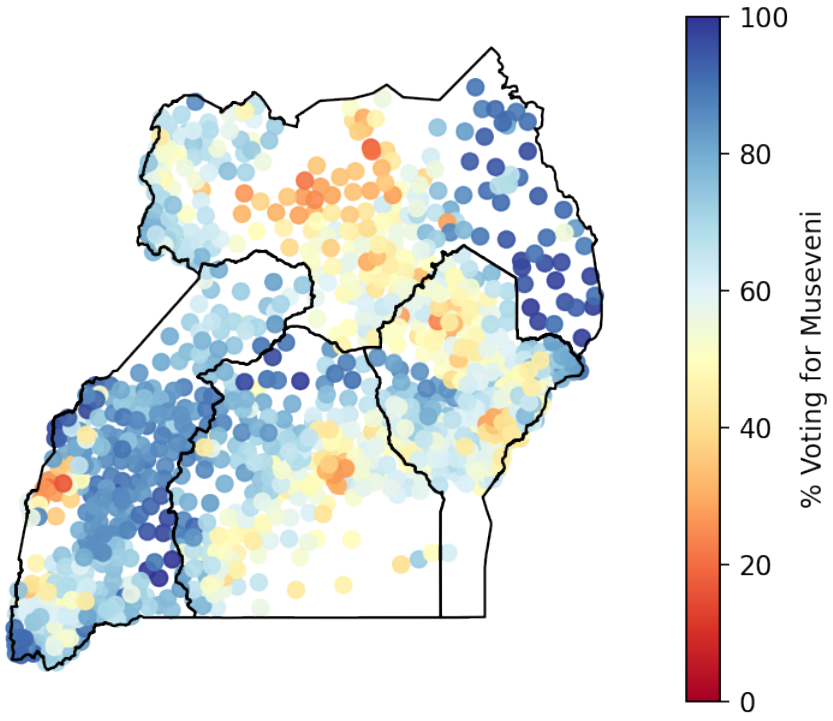

Figure 3.3. Museveni Voting Outcomes by Sub-county

Note: Each marker corresponds to a sub-county. Black lines demarcate Ugandan regions.

with election-specific attempts at Museveni trying to garner support in that election.

3.4.1 Fractionalization Effects

Figure 3.4 is a map of Uganda with colors corresponding to the mean pELF in each sub-county. It is not clear whether any systematic relationship exists between pELF and voting for Museveni, although there is some support for the argument that more ethnic fractionalization would tend to increase support for Museveni.

A cursory comparison of Figures 3.3 and 3.4 suggests that there might be some correlation between fractionalization and voting behavior. To explore this, we present a regression specification that uses our pELF measure. This variable of interest can tell us whether Museveni support is connected to general fractionalization in an area. We estimate the effect of ethnic fractionalization on voting outcomes for the following model of polling station i in sub-county g:

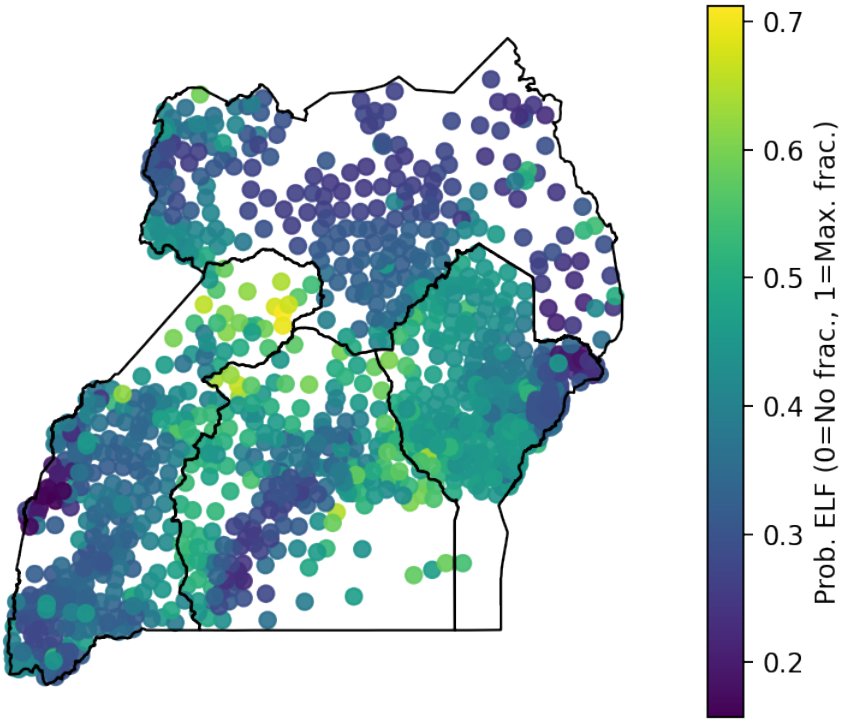

Figure 3.4. Probabilistic ELF by Subcounty

Note: Each marker corresponds to a sub-county. Black lines demarcate Ugandan regions.

$$\%Museveni_{ig} = \alpha + \beta pELF_{ig} \tag{3.1}$$
$$+\theta_1 age_{ig} + \theta_2 female_{ig} + \theta_3 turnout_{ig} + \theta_4 invalid_{ig} + \gamma_g + \varepsilon_{ig}$$

where $\%Museveni_{ig}$ is the vote percentage for the incumbent president, Yoweri Museveni. Control variables at the level of the polling station level are age, the average age of voters registered at the polling station, $female$, the proportion of female voters registered at the polling station, $turnout$ is the percentage of voter turnout, and $invalid$ is the percentage of invalid votes. γ_g are sub-county fixed effects and ε_{ig} is the error term. It is important to note that this measure, as well as predicted ethnic identity is a generated regressor. Simply using these as variables in a regression as if they were observed quantities might cause bias in inference. Future work will involve implementing a bootstrap procedure, such as in Wang et al. (2020) to derive correct standard errors.

To explore heterogeneous effects across residence regions, we also present a specification with pELF interacted with residence region:

$$\begin{aligned} \%Museveni_{ig} = \alpha + \sum_{r \in \{North, Central, East\}} \beta_r pELF_{ig} \cdot \gamma_r \\ + \gamma_r + \theta_1 age_{ig} + \theta_2 female_{ig} \\ + \theta_3 turnout_{ig} + \theta_4 invalid_{ig} + \gamma_g + \varepsilon_{ig} \end{aligned} \tag{3.2}$$

where γ_r are binary variables for residence region with γ_{West} being the omitted category.

Table 3.3 presents the results for the regression in Eq. 3.4.1. A one-point increase in fractionalization leads to a 0.32pp increase in support for Museveni.The increase in support for Museveni due to pELF may be due to several causes. One might be that, despite wanting Museveni out, it is difficult to create a united coalition of opposition or shared values with higher ethnic heterogeneity (Alesina et al., 1999). Or perhaps this effect is being driven by higher fractionalization leading to a higher incidence of a particular ethnic identity in a polling station, such as higher numbers of people with ties to the Center and West.

Given that our sample is nationally representative, we can also investigate the heterogeneous effects of pELF across residence region. Table 3.6 shows the estimated effects of interactions between pELF and residence region. What we find is that in contrast to Table 3.3, the region-specific effects to fractionalization are all mostly statistically insignificant, apart from the effect in the East. This suggests that higher fractionalization does not have a regional component apart from the Eastern region of the country, which is associated with less support for Museveni. These results seem puzzling, but they elucidate how fractionalization measures are implicitly functions of region. Fractionalization measures are essentially measures of variance of ethnic identity, and implicitly depend on the ethnic group in a location. For instance, high fractionalization in the East, essentially reduces to meaning "how many more ethnicities are there than Eastern ones?" When considering

Table 3.3. Effect of Prob. ELF on Museveni Support

	Museveni %	Museveni %	Museveni %
Intercept	59.934***	-89.869***	
	(2.678)	(7.979)	
Prob. ELF	1.954	39.282***	32.597***
	(7.358)	(4.934)	(2.979)
Age		1.516***	0.561***
		(0.192)	(0.098)
Female		30.810***	25.252***
		(2.071)	(1.328)
Voter Turnout		0.897***	0.543***
		(0.023)	(0.023)
Percent Invalid Votes		0.119*	0.004
		(0.064)	(0.022)
N	26814	26814	26814
R^2	0.00	0.33	0.23
SC FE	No	No	Yes

Note: Clustered standard errors at the sub-county level in parantheses. Ethnic fractionalization calculated using the Probabilistic ELF formula, which calculates the average probability that two voters in a polling station are from the same region. "SC FE" refers to subcounty fixed effects. Unadjusted R^2 reported. Location Effects calculated using the delta method. * p<0.10 ** p<0.05 *** p<0.01

fractionalization measures like this, the effects are not necessarily in terms of the frame of reference. As such, although fractionalization is a standard in the literature, it cannot get us closer to answering our main question.

3.4.2 Heterogeneity in Ethnic Identity by Residence Region

Since the nature of support for Museveni is most likely region-dependent, we also run a specification that uses each ethnic identity prediction as a separate variable. We estimate the effect of ethnic identity on voting outcomes for the following model of polling station i in sub-county g:

$$\begin{aligned} \%Museveni_{ig} = \alpha + \beta_1 p_{North} + \beta_2 p_{East} + \beta_3 p_{Central} \\ +\theta_1 age_{ig} + \theta_2 female_{ig} + \theta_3 turnout_{ig} + \theta_4 invalid_{ig} + \gamma_g + \varepsilon_{ig} \end{aligned} \tag{3.3}$$

The explanatory variables of interest are $p_{Central}$, p_{North}, and p_{East}, the ethnic compositions. The omitted variable, p_{West}, is the ethnic identity of the president. Relative to this omitted category, we might expect that the North and East ethnic identities will decrease support for the president. We once again include our control variables, age, $female$, $turnout$, $invalid$, as well as sub-county fixed effects. We also consider a specification where we interact these predictions with residence region to capture their heterogeneous effects.

$$\%Museveni_{ig} = \alpha + \sum_{r \in \{North,East,Central\}} \sum_{q \in \{North,East,Central\}} \psi_{r,q} \cdot p_r \cdot \gamma_q \tag{3.4}$$

$$\begin{aligned} +\theta_1 age_{ig} + \theta_2 female_{ig} + \theta_3 turnout_{ig} + \theta_4 invalid_{ig} + \gamma_g + \\ \gamma_r + \beta_1 p_{North} + \beta_2 p_{East} + \beta_3 p_{Central} + \varepsilon_{ig} \end{aligned} \tag{3.5}$$

Figure 3.5 shows the variation in each ethnic identity. Most variation takes place at the borders of each region of the country. For instance, in the map for the West, there is significantly more variation at its borders, shown by the "greener" dots. This is to be

Table 3.4. Region-specific Prob. ELF Effect on Museveni Support

	Museveni %	Museveni %	Museveni %
Intercept	53.787***	-100.438***	
	(3.696)	(7.198)	
Prob. ELF	48.436***	77.094***	43.627***
	(9.282)	(7.894)	(6.875)
InRegion==North	-1.159	1.530	
	(6.150)	(4.837)	
InRegion==East	13.439***	4.483	
	(4.783)	(4.148)	
InRegion==Central	-2.462	0.575	
	(5.413)	(4.195)	
InRegion==North x Prob. ELF	-34.065**	-30.296**	-2.128
	(16.220)	(12.466)	(10.890)
InRegion==East x Prob. ELF	-69.062***	-48.531***	-26.450**
	(11.926)	(10.094)	(12.369)
InRegion==Central x Prob. ELF	-42.287***	-35.645***	-12.742*
	(13.857)	(10.578)	(7.475)
Age		1.848***	0.555***
		(0.153)	(0.097)
Female		30.746***	24.987***
		(2.508)	(1.253)
Voter Turnout		0.794***	0.545***
		(0.022)	(0.023)
Percent Invalid Votes		0.199***	0.002
		(0.055)	(0.022)
N	26814	26814	26814
R^2	0.14	0.42	0.24
SC FE	No	No	Yes
Region FE	Yes	Yes	Yes

Note: Clustered standard errors at the sub-county level in parantheses. Predicted Region effects in terms of West used as omitted category. Ethnic fractionalization calculated using the Probabilistic ELF formula, which calculates the average probability that two voters in a polling station are from the same region. "SC FE" refers to subcounty fixed effects. "Region FE" refers to In-region effects Unadjusted R^2 reported. * p<0.10 ** p<0.05 *** p<0.01

expected as the borders are porous and tend to have people of either region and might lead to the machine learning algorithm having more variance in its probabilities for those regions. The Central region is a partial exception to this explanation since it has more ethnic diversity with the inclusion of Kampala. Despite this, there are also sub-counties away from the borders that have significant variation. This is illustrative of the within-subcounty variation that we will use for identification. These outlier sub-counties in the north and east of the country might roughly coincide with areas next to wildlife and game reserves, such as Kibale National Forest in the West and the Bokora Wildlife Reserve in the Northeast, where there are opportunities to settle and create new agricultural land (Hartter et al., 2015).

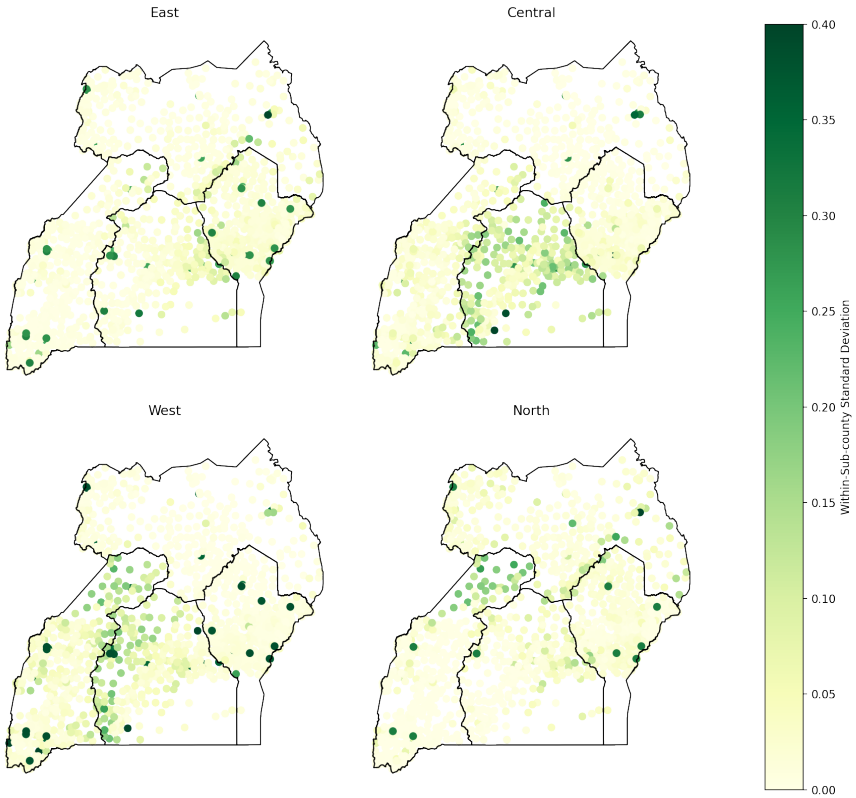

Figure 3.5. Region Prediction (within Sub-county Variation)

Note: Each marker corresponds to a sub-county. Black lines demarcate Ugandan regions. Each map gives the standard deviation of the prediction of coming from a particular region

We first discuss estimates of the effect of ethnic identity on voting outcomes without

heterogeneous effects. Switching identity from Western to any of the other regions by 1 percentage point (pp) decreases support for the president by around 0.2pp. The only significant effect in the regression is that of being predicted to be from the Central region. This suggests that polling stations with higher Central ethnic identity tend to decrease support for Museveni. This effect might be driven by Kampala and future research will include regressions without Kampala as a robustness check. But the results from Table 3.3 tell a seemingly different story than Table 3.5. When we take into account that the different regions of residence have different heterogeneous effects as in Eq. 3.4.2, this becomes much clearer.

When we run the regression associated with Eq. 3.4.2, we find that there are no changes in voting behavior as a result of being in different regions of residence. It would seem that ethnic voting behavior does not change with location, e.g. being from the North and moving to the West does not change one's voting behavior. This demonstrates that ethnic voting preferences are robust to location change. What we do see is that the *intensity* of one's decrease in support does change based on location. For instance, a polling station in the East tends to decrease support for Museveni by 0.9pp with an increase in more Eastern affiliation. But if there is more East affiliation in the North, then that number jumps to 1.41pp.

3.5 Conclusion

This paper studies the effect of ethnicity on voting behavior, in particular by disentangling the location and ethnic identity effects of voting. We estimate these effects by creating and analyzing a nationally representative data set from the 2016 Uganda presidential election. We overcome the challenge of not knowing the ethnicity of individual voters by predicting the ethnic identity of each voter by using machine learning methods to tie features of a voter's surname to their ethnic origins.

First, we estimate the effect of ethnic fractionalization on voting behavior. We find an average effect across polling stations in Uganda such that changing from no fractionalization

Table 3.5. Effect of Predicted Regions on Museveni Support

	Museveni %	Museveni %	Museveni %
Central	-30.328***	-19.103***	-16.699***
	(3.148)	(2.206)	(3.522)
North	-22.798***	-18.502***	-5.532
	(2.021)	(1.799)	(4.859)
East	-19.633***	-17.993***	-1.957
	(1.664)	(1.641)	(4.818)
Age		1.378***	0.483***
		(0.172)	(0.092)
Female		19.586***	19.803***
		(1.481)	(1.109)
Voter Turnout		0.694***	0.516***
		(0.024)	(0.024)
Percent of Invalid Votes		0.410***	0.018
		(0.058)	(0.022)
N	26814	26814	26814
R^2	0.13	0.36	0.21
SC FE	No	No	Yes

Note: Clustered standard errors at the sub-county level in parantheses. Predicted Region effects in terms of West used as omitted category. "SC FE" refers to subcounty fixed effects. Unadjusted R^2 reported. Location Effects calculated using the delta method. * p<0.10 ** p<0.05 *** p<0.01

Table 3.6. Region-specific Prob. ELF Effect on Museveni Support

	Museveni %	Museveni %	Museveni %
Intercept	55.920***	-85.460***	
	(3.287)	(6.405)	
InRegion==North	114.255***	95.988***	
	(22.120)	(15.235)	
InRegion==East	154.462***	244.448***	
	(39.754)	(34.312)	
InRegion==Central	48.781***	59.995***	
	(5.063)	(3.891)	
p_{North}	16.088***	30.663***	18.427***
	(5.111)	(4.325)	(5.101)
p_{East}	-7.519	42.733***	35.620***
	(16.511)	(11.257)	(10.045)
$p_{Central}$	88.855***	102.333***	41.404***
	(17.614)	(15.383)	(11.100)
InRegion==North x p_{North}	-129.391***	-127.641***	-147.262***
	(22.805)	(16.341)	(35.505)
InRegion==East x p_{North}	-234.645***	-317.895***	-77.065***
	(42.404)	(36.346)	(21.896)
InRegion==Central x p_{North}	-123.260***	-84.173***	-42.638***
	(17.295)	(13.957)	(8.068)
InRegion==North x p_{East}	-96.141***	-120.631***	-141.380***
	(27.292)	(19.642)	(36.215)
InRegion==East x p_{East}	-140.214***	-288.618***	-89.897***
	(43.887)	(36.770)	(20.950)
InRegion==Central x p_{East}	-55.453***	-90.543***	-69.405***
	(18.866)	(12.507)	(11.689)
InRegion==North x $p_{Central}$	-313.488***	-129.379***	-167.598***
	(72.169)	(39.946)	(59.354)
InRegion==East x $p_{Central}$	-276.950***	-340.606***	-80.836***
	(51.655)	(44.085)	(26.289)
InRegion==Central x $p_{Central}$	-147.697***	-164.153***	-84.205***
	(18.219)	(15.687)	(11.467)
N	26814	26814	26814
R^2	0.22	0.48	0.25
SC FE	No	No	Yes
Controls	No	Yes	Yes

Note: Clustered standard errors at the sub-county level in parantheses. Predicted Region effects in terms of West used as omitted category. "SC FE" refers to subcounty fixed effects. Unadjusted R^2 reported. * p<0.10 ** p<0.05 *** p<0.01

($ELF = 0$) to complete fractionalization ($ELF = 0.75$) increases a polling station's support for President Museveni's by 0.20 percentage points. Our results suggest that ethnic fractionalization has a positive effect on support for Museveni, which is independent of the location of the polling station.

Second, we estimate the effect of ethnic identity itself on voting behavior. Relative to identifying with the President's ethnic identity of West, support for the President decreases most with the ethnic identity of Central, followed by North and East. Polling stations with a higher proportion of voters with Central identities tend to reduce support for Museveni. When looking at the location of the polling station, we find that ethnic voting preferences are robust to location, but that some residence regions tend to amplify those preferences. This finding shows that ethnic voting preferences may be robust to location factors.

By employing a machine learning approach to construct identity measures of individuals in a nationally representative data set, this study is able to disentangle ethnicity's identity and location effects on voting behavior across a country. This is important as it adds a new dimensions to the discussion of ethnic voting.

3.6 Appendix A: The Probabilistic ELF Formula

For a given polling station, we define its Probabilistic ELF as the average probability that two voters in the polling station are from the same region. We define the Probabilistic ELF formula as:

$$\widehat{ELF} = \frac{1}{n(n-1)} * \sum_i \sum_j (1 - \widehat{West}_i \widehat{West}_j - \widehat{North}_i \widehat{North}_j \tag{3.6}$$

$$- \widehat{East}_i \widehat{East}_j - \widehat{Central}_i \widehat{Central}_j) \tag{3.7}$$

Chapter 4

Downstream Market Power and Commune Privatization

4.1 Introduction

Market power in various forms can impede the welfare benefits of land reform. The consequences of ignoring, for example, land concentration has been witnessed in land reforms such as in Ukraine (Plank, 2013), Zimbabwe (Moyo, 2011), and in Colonial Chile (Conning, 2001). What is less studied, but which is just as important, is how market power in the supply chain affects the success of land reform and the rate of the privatization of common lands. But agriculture is different in developing country contexts because the process of privatization comes with strategic factors that standard decentralized smallholder agriculture does not. Agriculture with a common land component is different because some feature of agriculture is shared across members. Members can pool costs and reduce fixed costs, or risk-share, but at the expense of receiving less revenue.

For instance, take a cooperative with farmers of different skills, and that share in the profits of production. When they produce, suppose that there is no division between "worker" and "manager" as in Lucas Jr (1978), and that skills "mix" in some way to get at some average productivity in the cooperative. Total output for this cooperative would be lower if they stayed together than if some members were to break up and become separate

smallholder farmers. Suppose that this was made possible by a privatization reform and that it was costless. Privatization would reduce to a strategic decision about whether the reduction in costs from leaving the cooperative (given the number of members it has) was worth the increase in revenue (given that higher-skilled members would increase revenue in the cooperative for a lower-skilled player). This decision is sensitive to downstream competition, since downstream competition partially determines revenue.

An example of a cooperative institution that has these strategic dynamics and is affected by downstream pressure is the Mexican *ejido*. I will present a model of *ejido* privatization that explains the tensions that affect the decision to privatize. This model will only consider the strategic considerations behind common land in the *ejido* and we will abstract away from any political economy considerations or tragedy of the commons problems. Instead, we will assume that the *ejido* has no private parcel land, for the sake of simplicity and clarity. In reality, *ejidos* of this sort do exist, but they make up a small fraction of all *ejidos*. The *ejido* in this model will be thought of as a cooperative institution in which there is heterogeneity in ability across members, and with rules that prioritize egalitarianism. In the case of a privatization reform, those with the highest ability have an incentive to exit first, since they would be the most competitive. However, members must cope with various infrastructure challenges, such as poor road quality, remoteness and distance from large cities, as well as high altitude, which would lead to costs that can be reduced by sharing the burden in the cooperative. These challenges can also serve as a way of reducing the market power of downstream processors, potentially protecting them from higher mark-ups.

I plan on acquiring data from INEGI (Instituto Nacional de Estadística y Geografía), which will include the Ejidal Census with a panel of Ejidos across multiple rounds, as well as supplementary data in the form of the Mexican economic census. I will focus my analysis on maize and maize processing. The Economic Census data provides information on firms and a detailed list of industry codes that I can use to see if the firm is a maize processor or not. The other advantage of the INEGI data is that it can provide information on location

and market share. This will enable me to create a spatial measure of market power with which I can match firms to *ejidos*, infer which processors the *ejidos* are connected to, and measure the mark-ups faced by each *ejido*. I will use the 1996 grain price shock as a policy event to evaluate my hypothesis (Light and Shevlin, 1998).

Using this data, I will present two main specifications that will allow me to capture the effect of market power on privatization. The first will use the global price of corn as an instrument for mark-ups to see if that affects the share of households that have formally partitioned and privatized their land in Mexico. Secondly, I will use the roll-out of *PROCEDE*, the land reform in Mexico, along with mark-ups and the price shock as a policy event to run a triple-difference in difference estimation.

The paper is organized as follows: Section 2 lays out the structure and context of the model. Section 3 will set up the cooperative exit game and its Nash equilibrium. Section 4 will introduce the downstream processor and their optimization problem, as well as the main insights from the model. Section 5 will set up the empirical setting, estimating mark-ups and my estimation strategies. Section 6 concludes.

4.2 Background

The *ejido* is a cooperative agricultural institution that came into existence after the Mexican revolution in the early 20th century. These reforms were an attempt to alleviate large inequalities in land ownership up to that point, and seize power from wealthy landowners that could potentially rise up in counterrevolution. The property rights specified by Article 27 of the Mexican Constitution were set up so that farmers could not could not sell or lease land to anyone else, making *PROCEDE* and its formal titling of land an important step in the privatization process.

Mexico undertook its privatization reforms starting in 1994 (Janvry et al., 1997). There were two main features of Mexican land reform, *PROCEDE* and *dominio pleno*. The RAN (Registro Agrario Nacional) was created to formally register all *ejido* land in INEGI, as well as the boundaries of each part of the *ejido*, whether it was private or common, and

give a partial title of the land to *ejidatarios* (Perramond, 2008). PROCEDE allowed the registration of land in the *ejido* and gave *ejido* members or *ejidatarios* a formal title on land, while *dominio pleno* created a formal process for partitioning and dividing the land in the *ejido* into private hands. PROCEDE and *dominio pleno* was rolled out nationwide from 1992-2006. Although over 90% of *ejidos* were certified with PROCEDE, only 10% has taken part in *dominio pleno*. And in areas such as Chiapas and Oaxaca, *PROCEDE* has been completely rejected (Barnes, 2009). *PROCEDE* has been shown to increase maize productivity (McArthur, 2016), and has led to increased migration due to its land tenure risk reduction (De Janvry et al., 2015).

Agrarian laws mandate collective decision-making for *ejidos* which require all *ejido* members to have a general assembly in charge of setting rules, settling disputes and even re-allocating usufruct (usage) rights (Barnes, 2009). Rules include outlawing land sales, rental contracts and the administration and allocation of common land (Dower and Pfutze, 2013). There tend to be three different "types" of *ejido* land tenure regimes: an urban zone, common land, and individual agricultural parcels. Common use areas usually cover twice as much land as the individual agricultural parcels and are usually made up of forest or pasture land (Barnes, 2009). In 7% of cases of *PROCEDE* registered land, all land is held in common.

Concurrently with PROCEDE, Mexico was also going through a period of economic liberalization with a period of decreased regulation and agricultural price controls, as well as the advent of NAFTA (The North American Free Trade Agreement). NAFTA opened the doors for more severe monopsony effects in trade, and the effect of imperfect markets on *ejido* privatization has not been studied.[1] Concentration in processing firms has been shown to increase prices for consumers and reduce prices to farmers in Mexico (Murphy, 2006). Since the inception of NAFTA in 1992, maize prices paid to local farmers have

[1]Recent literature has focused on the institutional elements that explain the persistence of cooperative agriculture, despite reform laws. These include looking at power structures (Hoch, 1986), the management of common property (Alix-Garcia et al., 2005), and cooperation through coalition creation (Bianco and Bates, 1990). Most of this literature assumes that the root of why the *ejido* persists comes from the need to manage common resources and the challenges that arise when those resources need to be partitioned during privatization.

declined, while tortilla prices have went up in real terms. This is at least partly due to the economic liberalization that NAFTA brought, which created a less regulated market where market power began to increase (Nadal, 2000; Rogers and Sexton, 1994). Although NAFTA has improved the Mexican economies in some regards, the effect on the rural sector has been negative (Barnes, 2009). The Mexican context of the mid-1990s provides the perfect setting to think about how monopsony power affects privatization during land reforms.

4.3 The Model

The model begins by positing an economy with a fixed endowment of land $\bar{T}$ and N members. I abstract away from any political economy effects having to do with an *ejido* assembly, and assume that the *ejidatarios* are driven purely by profit motive. Suppose that all land is owned by the cooperative, and is characterized by three features:

1. Each producer in the cooperative is endowed with a productivity parameter, β_i.
2. Members of the cooperative have a title to an equal share of land, but produce together and receive an equal share of the profits from production. When producing together, their managerial abilities will "mix" and the cooperative will produce at its average ability. This is a simplification of how many real *ejidos* use their common land. Many of them have open access common land that is not cultivated. But in many instances, *ejido* members run some agricultural, forest, or pasture operation and allocate profits based on rules in the general assembly. I abstract away from more complicated sharing rules by just giving each member an equal share.
3. There is a privatization reform that allows a producer to costlessly leave the cooperative. If they choose to exit, then they take their share of the land and begin to produce privately with the same technology, but being able to use their own $\bar{\beta}$.

Suppose for now, that this economy takes prices as given and must spend a fixed, transportation-related cost to get their goods to market. Let this fixed cost be called θ. θ is related to contracting costs and is a proxy for the remoteness of an area, and distance

from marketplace. For example, remote areas might have a harder time finding a trucking company to send their goods to market.

Suppose the cooperative can sell their good at price p_f. Production technology is characterized by a function $f(T)$, where T is land, and with $f' > 0$ and $f" < 0$. It is important to note that we assume decreasing returns to scale for the production function and this is assumed primarily to capture any incentive problems that would make it difficult to deal with larger plots of land. For instance, large forest held in common can be subject to too much deforestation or pasture land can be over-grazed. Future work will model these incentive issues directly. It would be useful to define cooperative profits in terms of the conditional expectation of β below some threshold, β^*. Additionally, assume that M members have already left the cooperative leaving it with $\bar{T}(1 - \frac{M}{N})$ units of land. Cooperative profits are then:

$$\pi_c = p_f E(\beta|\beta \leq \beta^*) f(\bar{T}) f(1 - \frac{M}{N}) - r\bar{T}(1 - \frac{M}{N}) - \theta(D)$$

where $r > 0$ is the rental rate of land. The profits for a single farmer would then be: $\frac{\pi_c}{N-M}$

If a member decides to break away, they will take their share of land and produce privately. The profits for such a farmer would then be:

$$p_f \beta f(\frac{\bar{T}}{N}) - r\frac{\bar{T}}{N} - \theta(D)$$

Figure 4.1 illustrates the payoffs for cooperative and private producers and shows the relationship between the private returns of a member and their returns if they broke away. The blue line gives the private returns to a member of the cooperative, which increases with more members. This is due to the fact that there are more members with higher ability in the cooperative, and more people to share in the burden of θ. But also, higher membership means that there is more land on which to produce. The red contour shows

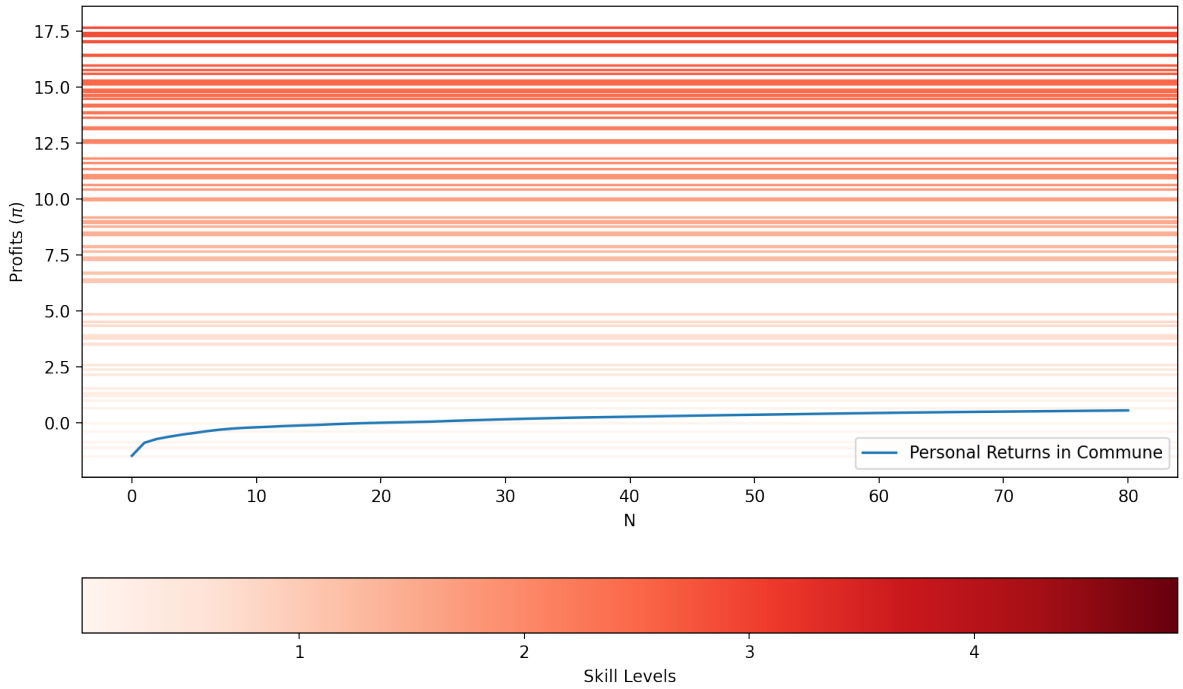

Figure 4.1. Private and Cooperative Profits

the private profits for a given skill level, if they were to leave the cooperative. The darker the color, the higher the skill being pictured.

Figure 4.1 also illustrates how we might go about intuiting a pure strategy Nash equilibrium for this game. If personal cooperative returns for some $N - M$ members is higher than the number of members with lower private returns, then that is a candidate for a stable equilibrium in which those $N - M$ members decide to stay in the cooperative, and the rest produce privately. In this particular figure, there is no candidate pure strategy Nash equilibrium. There is one crucial idea that makes this equilibrium possible: the idea that there is an ordering to members' strategies in the exit game.

4.4 The Cooperative Exit Game

The Cooperative exit game is a simultaneous game played by all the members of the cooperative. Each player decides whether to stay in the cooperative or leave and take part in private production. The condition for a member with some ability β, to stay in the

cooperative (for a given set of choices made by the other players) is:

$$\frac{p_f E(\beta|\beta \leq \beta^*) f(\bar{T}) f(1-\frac{M}{N}) - r\bar{T}(1-\frac{M}{N}) - \theta(D)}{N-M} \geq p_f \beta f(\frac{\bar{T}}{N}) - r\frac{\bar{T}}{N} - \theta(D) \quad (4.1)$$

A member will stay in the cooperative if their payoff from staying is higher than from leaving. Private profits are strictly increasing in skill and the average cost of the fixed cost θ is decreasing in N. The following proposition will show that the equilibrium to this game will simply require finding some β^* member who is indifferent between staying and leaving. They will define the partitioning of the skill distribution and give a rule as to who stays and who goes.

Proposition 1 *If the highest skilled player, with skill $\bar{\beta}$, stays in the cooperative, it cannot be that some player with skill $\bar{\beta} - \epsilon$ with $\epsilon > 0$ will decide to leave.*

Proof: Suppose to the contrary, that this is possible. This means that for the highest skilled player:

$$\frac{p_f E(\beta|\beta \leq \bar{\beta}) f(\bar{T}) f(1-\frac{M}{N}) - r\bar{T}(1-\frac{M}{N}) - \theta(D)}{N-M} > p_f \bar{\beta} f(\frac{\bar{T}}{N}) - r\frac{\bar{T}}{N} - \theta(D)$$

while simultaneously meaning that:

$$\frac{p_f E(\beta|\beta \leq \bar{\beta}) f(\bar{T}) f(1-\frac{M}{N}) - r\bar{T}(1-\frac{M}{N}) - \theta(D)}{N-M} \leq p_f (\bar{\beta} - \epsilon) f(\frac{\bar{T}}{N}) - r\frac{\bar{T}}{N} - \theta(D)$$

But this would mean that $\bar{\beta} < \bar{\beta} - \epsilon$, by monotonicity, a contradiction.■

Proposition 1 shows that the cooperative stays together as long as the highest skilled member stays in the cooperative. As soon as they leave, however, this has an effect on the average skill in the cooperative and the number of members that share profits. The

solution to the problem is then to find the member that is indifferent between staying and leaving.

Theorem 1 *The Nash equilibrium for the cooperative Privatization game is characterized by β^* which satisfies the following equation:*

$$\frac{p_f E(\beta|\beta \leq \beta^*) f(\bar{T}) f(1-\frac{M}{N}) - r\bar{T}(1-\frac{M}{N}) - \theta(D)}{N-M} = p_f \beta f(\frac{\bar{T}}{N}) - r\frac{\bar{T}}{N} - \theta(D) \quad (4.2)$$

where, given p_f, r, and θ, all $\beta < \beta^*$ will stay in the cooperative and everyone else breaks away and produces privately.

Proof: Suppose, to the contrary, that the split at β^* is not a Nash equilibrium. Then it must be that there exists at least one player with $\beta < \beta^*$ who would prefer to leave or a $\beta > \beta^*$ who would prefer to stay. But we know by Proposition 1, that if the highest skilled players chooses to stay, then all other players choose to stay as well. Since the highest skilled player in the cooperative is $\beta \leq \beta^*$, the player cannot come from those that are in the cooperative. Then there must be a player with $\beta > \beta^*$ that would rather stay in the cooperative. Call this $\hat{\beta}$. But if $\hat{\beta}$ would prefer to be in the cooperative, then by Proposition 1, all $\beta \in [\beta^*, \hat{\beta}]$ would also want to stay in the cooperative as well, making $\hat{\beta}^*$ a new split and a new Nash equilibrium satisfying Equation 4.2 and where all players would not deviate.■

We can see an example of the equilibrium in Figure 4.2, where the intersection of the two curves from Equation 4.2 intersect to give the profits of the indifferent member.

4.4.1 Comparative Statics

Now that we have a Nash Equilibrium, I explore two main comparative statics before heading to the next section. Based on Theorem 1, we know implicitly that there is a function $\beta^*(p_f, \theta(D))$ that governs the relationship between the price of the raw good, transportation-related contracting costs, and the equilibrium size of the cooperative.

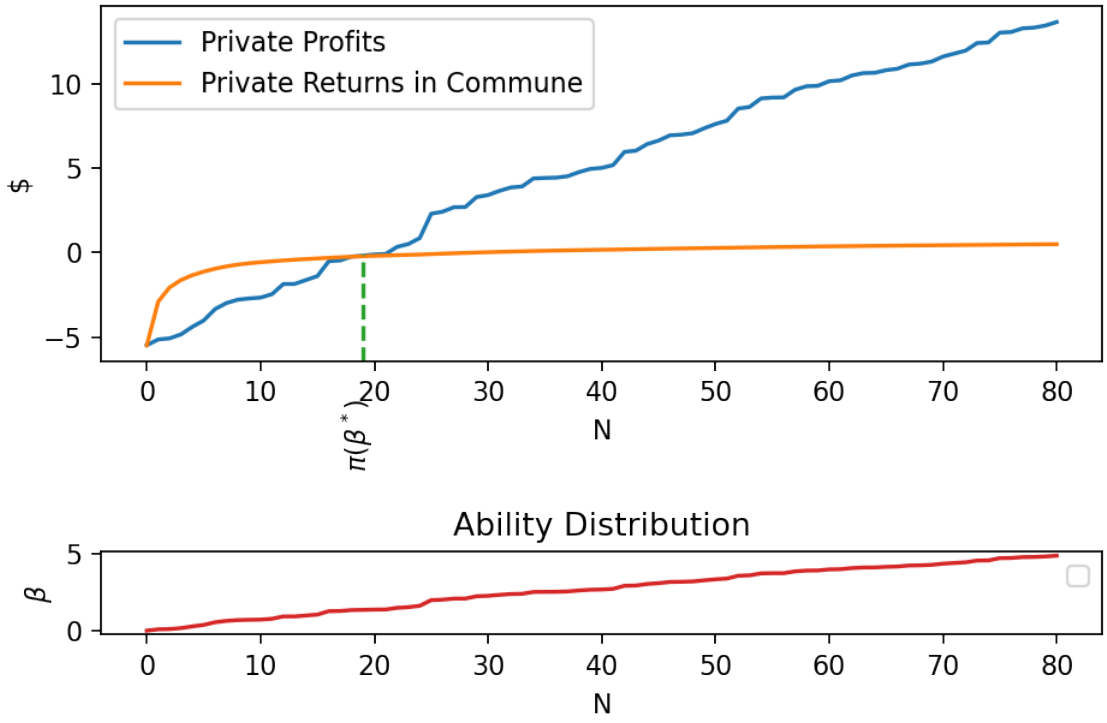

Figure 4.2. Nash Equilibrium Payoff Function

Firstly, we will look at the effect of transportation-related contracting costs on cooperative size. This is important as it gives a glimpse into what we might expect as we head further from city centers. The second is looking at how the price offered to the cooperative affects its size. This will be important when analyzing the decisions of the processor.

Regardless of the distribution of ability, a cooperative would shrink as there would be nothing to incentivize higher skilled members to stay. When we allow for $\theta(D) > 0$, however, we can show that higher transportation-related contracting costs tend to create larger cooperatives. Since these costs are shared across members in the cooperative, there comes a point when it becomes more profitable to stay in the cooperative even if a member's share of the profits might not be as high as in the transportation-related costs' absence. Figure 4.3 shows the effect of transportation-related contracting costs on β^*. Costs increase the β^* cutoff and create larger cooperatives in equilibrium.

Another important comparative static that is important for the analysis is understanding

how the price of the good sold by the cooperative affects the equilibrium. If transportation costs are held constant, an increase in p_f will tend to create higher potential revenues for higher ability members. Higher prices create an incentive to leave and so, as a result, tends to decrease the size of the cooperative.

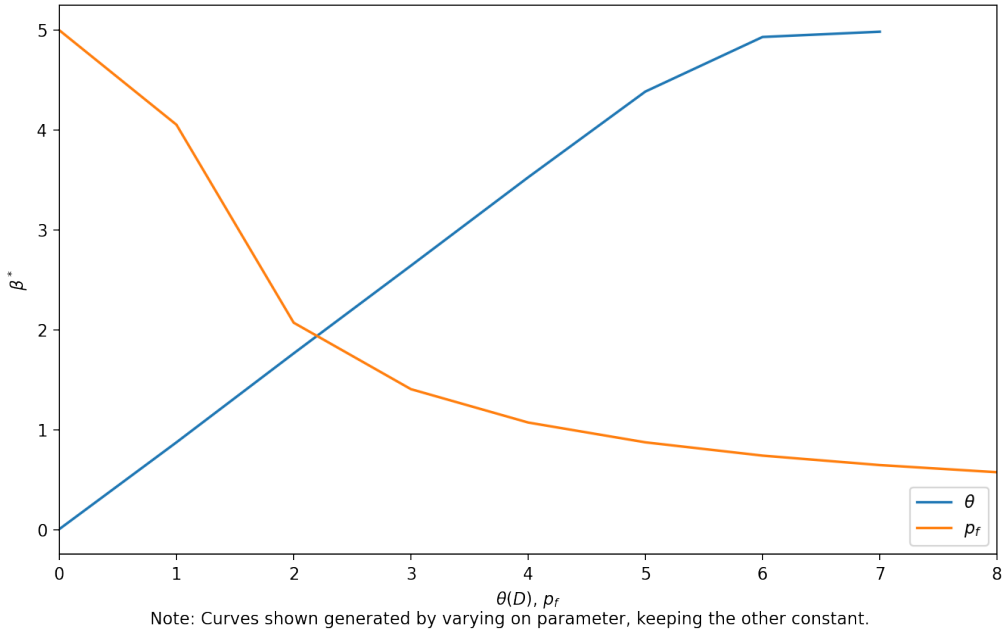

Figure 4.3. The Effect of Transportation-Related Contracting Costs and Price on Cooperative Size

4.5 The Processor

When prices are exogenous and both cooperative members and private producers are subject to the same price, all else equal, a positive price shock tends to create more privatization. On the flip side of this, lower prices lead to higher cooperative membership, as the incentive to leave becomes lower. In the next section, we will explore what happens when a processor with monopsony power has a first-mover advantage in a larger cooperative exit game.

4.5.1 Processor with Monopsony Power

Suppose that the international price of a processed good is p_r. The processed good needs the raw good provided by the cooperative in order to create it. The processor knows that

the cooperative exit game exists and acts as a Stackelberg leader. Specifically, this means that they are are aware of $\beta^*(p_f, \theta(D))$ and maximize their profit function, subject to knowledge of that function. For simplicity, let $E(\beta|\beta \leq \bar{\beta}) = \bar{\beta}$:

$$\begin{aligned} \max_{p_f}(p_r - p_f) \cdot f(\frac{\bar{T}}{N}) \left[f(N-M)\bar{\beta}(\beta \leq \beta^*) + M\bar{\beta}(\beta > \beta^*)\right] \\ s.t. \\ p_f \geq \bar{p} \\ \beta^* = \beta^*(p_f, \theta(D)) \\ M = M(\beta^*(p_f, \theta(D))) \end{aligned} \tag{4.3}$$

$M(\beta^*(p_f, \theta(D)))$ is the number of members that decide to break away. This function comes naturally as a result of the form of the equilibrium; since members leave in order, β^* simultaneously defines M as well. Another feature of this problem to note is that there is some lower bound for the price that the monopsonist can choose, $\bar{p}$. $\bar{p}$ is meant to proxy for a limit on the pricing power of the monopsonist. This primarily comes from the fact that pricing too low might attract more processors to the market, as there is a potential competitive fringe waiting to enter if there is an opportunity (MacDonald, 1986).

The first order condition for this problem boils down to an expression relating prices to cooperative output and privatization. Let Q^* be the total output of the economy, i.e. $f(\frac{\bar{T}}{N}) \left[f(N-M)\bar{\beta}(\beta \leq \beta^*) + M\bar{\beta}(\beta > \beta^*)\right]$. Then we can write the first order condition like so:

$$\frac{p_r - p_f}{p_f} \geq \frac{1}{\epsilon^{Q^*}_{\beta^*} \cdot \epsilon^{\beta^*}_{p_f}} \tag{4.4}$$

These first order conditions tell us that mark-ups in an economy with a monopsony processor must be equal to two quantities: $\epsilon^{Q^*}_{\beta^*}$ or the *output elasticity of privatization* and $\epsilon^{\beta^*}_{p_f}$, the *price elasticity of privatization*. If $\frac{p_r - p_f}{p_f} > \frac{1}{\epsilon^{Q^*}_{\beta^*} \cdot \epsilon^{\beta^*}_{p_f}}$, then the $p_f = \bar{p}$. This

means that at a certain point, if the elasticity of supply becomes low, the monopsonist is constrained and must price at $\bar{p}$.

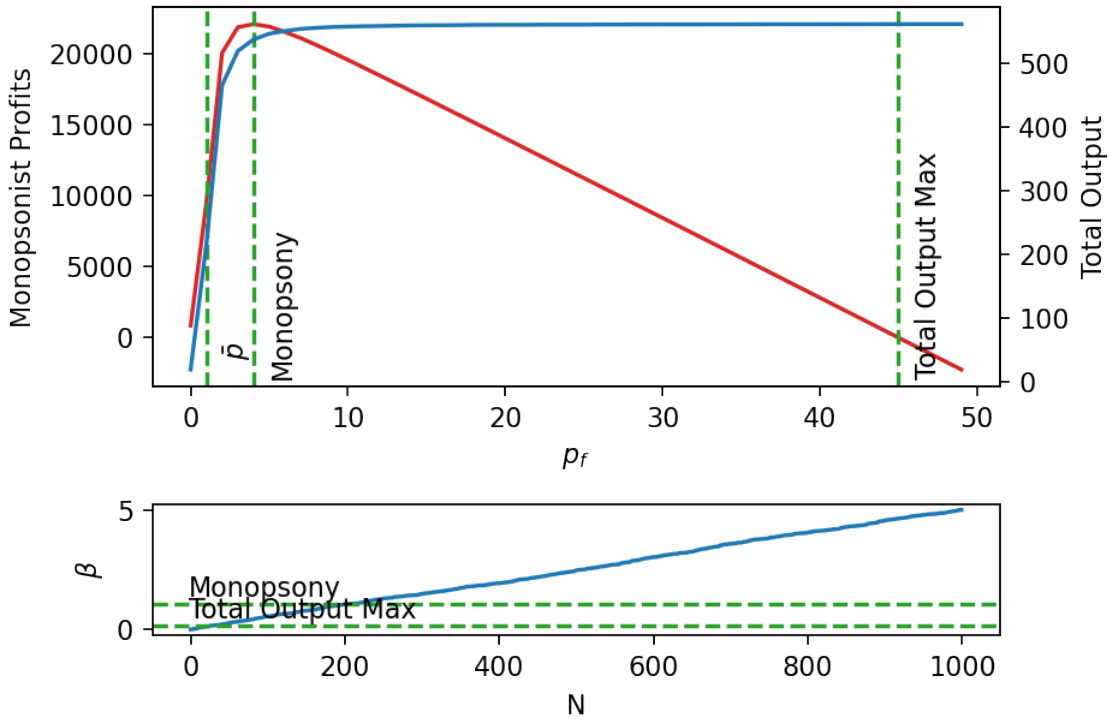

Figure 4.4. The Monopsony Processor Equilibrium

Figure 4.4 illustrates two basic ideas. The first is that higher privatization raises output. This is seen by how higher prices tend to induce privatization. Since the production function is decreasing returns to scale, this means that there is a higher marginal product of land at lower quantities of land used. Note that under constant returns to scale, this would not happen. There wouldn't be any difference between producing in a cooperative or separately. This can be seen in the objective function if you assume a constant returns to a scale technology, such as $f(t) = t$. The objective function would reduce to: $\frac{\bar{T}}{N} \sum_{i=1}^{N} \beta_i$. The main lever of optimization used by the processor in this problem is controlling the amount of total output through their effect on privatization; this is only possible when there is inefficiency to privatization, which would not be the case under constant returns to scale.

Under constant returns to scale, it is clear that the optimal choice by the monopsonist

would be to choose $p_f = \bar{p}$. However, with decreasing returns to scale and a cooperative, the processor would rather choose a higher price, but a much lower price than if it was to maximize total output. The monopsonist processor's main tradeoff is that paying a higher price leads to more efficient production, but at the cost of needing to pay higher and higher prices to induce privatization. Based on the curvature we see in Figure 4.3, the effectiveness of p_f at inducing privatization becomes more and more difficult. The monopsonist stops at that point where the benefits of inducing an extra cooperative member to privatize is no longer worth its cost.

4.5.2 Monopsony Processor Location Choice and Countervailing Market Power

Although the model is not explicitly spatial, we can still say something about where a monopsonist might choose to locate if we assume that location choice is at least partly driven by transportation costs and contracting.

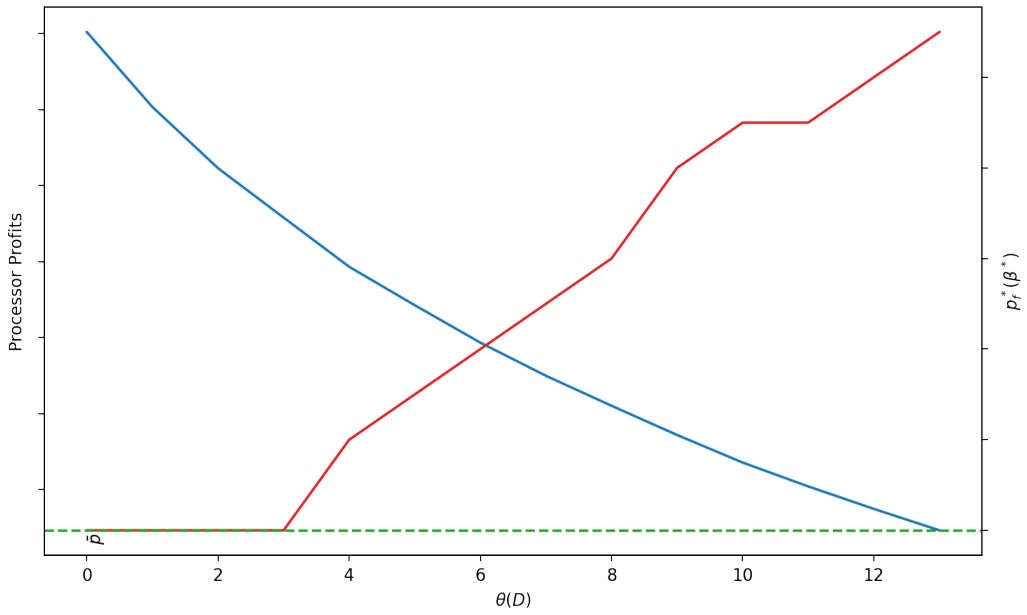

Figure 4.5. The Effect of Transportation-Related Contracting Costs on Processor Profits

Figure 4.5 illustrates how equilibrium profits changes with higher transportation-related

contracting costs. We can see that these costs reduce processor profits and this is due to the fact that higher costs mean that it becomes more difficult to induce privatization. As a result, processors would tend to locate in less remote areas with more developed infrastructure.

Another way to think of θ, though, is as a way to build in a counter-vailing force against the market power of the processor. We can see here that as $\theta(D)$ goes down, then the optimal price that the processor offers increases, showing that their market power and mark-up goes down with higher θ. There are tradeoffs then to cooperatives in high θ areas. On the one hand, there is more protection against monopsony and market power, but at the cost of significantly higher contracting costs for finding a someone to transport goods.

4.5.3 Monopsony Processor Price Passthrough

It is important to note how price will be transmitted from an international market to a cooperative. In a competitive world, p_r would be passed through completely to the cooperative, i.e. that $p_r = p_f$. From Equation 4.4, we can see that the amount of passthrough will depend on the elasticities of privatization:

$$\frac{dp_r}{dp_f} = 1 + \epsilon_{p_f}^{\beta^*} \epsilon_{\beta^*}^{Q^*} (1 + \delta_{p_f}^{\beta^*, p_f})$$

where $\delta_{p_f}^{\beta^*, p_f}$ is the *elasticity of the price elasticity of privatization*. From this expression we can see that pass-through will be less than 1. This means that monopsony processors will tend to pass-through less than the full price to the cooperative, leading to lower privatization than under competition. This is most interesting in the case where we might analyze a price shock to see its effects on privatization. Similar areas that only differ by the mark-ups of their downstream processors will tend to have different effects on privatization.

4.6 Empirical Approach

The model presented in this paper shows how different rates of privatization can be explained by differing mark-ups of downstream processors. Downstream processors with high market share will tend to pass-through less of a price increase, leading to smaller rates of privatization than expected under competition. The challenge in this context is finding an identification strategy that will give a causal estimate of the effect of market power on privatization.

The data that I have access to are the Mexican Economic Censuses which are take every five years from 1989-2007. This data will provide me with industry codes and locations of processing firms. On the *ejido* side, I will have access to a mix of Ejidal and Agricultural Censuses from 1991-2007, which provides input and profit information for *ejidos*, as well as whether they have taken part in PROCEDE and *dominio pleno*. The data that I have access to will also include GPS coordinates of firms, which will allow me to match *ejidos* with processing firms.

I will also use data from FRED (Federal Reserve Economic Data) on global corn prices to produce price shocks of corn that I will either use as instruments in an IV regression or as a policy event in a triple difference-in-difference strategy. The 1996 grain price shock (Light and Shevlin, 1998) provides an exogenous policy event that was caused by a drought in the Midwestern United States in 1995-1996, making it a plausibly exogenous policy event for Mexico.

In the subsections that follow, I will first outline how to estimate mark-ups of the processors that *ejidos* face. After that I will discuss several strategies for using these mark-ups, in conjunction with municipality level *dominio pleno* rates to estimate the effect of market power on privatization.

4.6.1 Estimating Markups

Our independent variable of interest needs to be estimated as market share isn't a good enough proxy for firm mark-ups. I take the definition of markups from De Loecker and

Eeckhout (2017), which uses a method similar to Olley and Pakes (1992). We start with looking at firms in the economic census:

$$q_{it} = \beta_v v_{it} + \beta_k k_{it} + \omega_{it} + \epsilon_{it}$$

where q_{it} is log of deflated sales for firm i at time t, v_{it} are variable inputs, k_{it} are capital inputs and ω_{it} is an unobserved productivity term, which is an AR(1) process. Using the two-step approach as in Olley and Pakes (1992), we define unobserved productivity as a function of variable inputs and a set of controls. We can identify ω_{it} if we assume that variable inputs respond to productivity shocks, that lagged variable inputs do not, and that lagged variable input use is correlated to current period variable input use. Markups can then be estimated, by calculating:

$$\hat{M}_{ist} = \beta_v \frac{S_{ist}}{C_{ist}}$$

where S_{ist} is sales and C_{ist} is the total variable cost of production.

4.6.2 Price Shocks as an Instrument

If we were to merely regress privatization shares in municipality m, on markups (setting aside the fact that mark-ups are estimated and standard errors would need to be corrected), we would not achieve identification because there are many reasons why mark-ups would be correlated with the error term. The error term in this regression would include any processor-level or *ejido*-level factors that would be correlated with privatization.

For example, this could be through contracting costs, θ, as discussed in previous sections. Processing firms would tend avoid remote areas where there would be infrastructure or transportation challenges. This is exactly those features that would cause there to be larger cooperatives. We can try to control for this through a variety of proxies: altitude measures, number and type of roads within some radius of a given *ejido*. However, there are several factors that cannot be easily accounted for.

To get around these issues, I present a model that will use a large price shock in a processed good, like tortillas. The basic empirical approach will first consider the following equation:

$$Priv_{ist} = \gamma_m M_{ist} + \theta TransCosts_{ist} + X_{ist}\beta + \alpha_e + \alpha_s + \varepsilon_{irt}$$

where $Priv$ is the share of privatization in *ejido* i, region/market r at time t. M is the markup of the market that *ejido* i is connected to, $TransCosts$ are a measure of transportation costs, X are a set of control variables, α_e are *ejido* fixed effects, and α_s are municipality fixed effects. What I'd like to explore is the effect of a change in mark-ups that is precipitated by a price shock in maize.

I will instrument firm mark-ups with the international price shock in maize:

$$M_{ist} = P_t\alpha + \theta TransCosts_{irt} + X_{irt}\beta + \alpha_e + \alpha_s + \nu_{irt}$$

However, an issue with this identification strategy is that it relies on a singular price shock having a strong enough effect that it changes privatization dynamics across time. More plausibly, a shift in the price distribution across time would cause more long-term effects on privatization, but this would threaten the exclusion restriction for using price as an instrument.

4.6.3 Triple-Difference in Difference Estimation

Another way I can estimate the effect of market power on *ejido* privatization, is through assuming that the *ejidos* in the same region subject to different mark-ups have the *same trends* in their privatization rates across time. In this case, we would compare *ejidos* before and after a large price shock in corn (such as the 1996 grain price shock in the U.S.A.) to those *ejidos* eligible for PROCEDE across different levels of mark-ups. This regression specification would then be:

$$Priv_{irt} = \gamma_m M_{ist} + \gamma_p Pre_t + \gamma_{PROC} PROCEDE_{st}$$
$$+\delta_{mp} M_{ist} \cdot Pre_t + \delta_{m,PROCEDE} M_{ist} \cdot PROCEDE_{st} + \delta_{p,PROCEDE} Pre_t \cdot PROCEDE_{st}$$
$$+\alpha M_{ist} \cdot Pre_t \cdot PROCEDE_{mt}$$
$$+X_{ist}\beta + \alpha_e + \alpha_s + \varepsilon_{ist}$$

This would allow me to estimate a causal effect of *ejido* privatization at different levels of mark-ups in the country. The largest drawback to this strategy is that the parallel trends assumption might not be satisfied, and any estimate would then be biased. But in contrast to the instrumental variables approach, we can test for parallel trends or even make them conditional on observable covariates, since the triple difference estimator is a special case of the standard difference-in-difference estimator (Olden and Møen, 2020).

4.7 Conclusion

In this paper, I present a model of *ejido* privatization. Transportation costs and geography are important factors when considering privatization reform. Monopsony power induces less privatization and incomplete price pass-through leads to positive price shocks being attenuated. There has been an extensive literature on how land concentration affects land reform, but none have considered how market power in the supply chain affects privatization and welfare. This arguably more important to explore because monopsony power in processing is a constant reality in agriculture and its interaction with land reform is not well understood.

This paper also presents an empirical approach to test the implications of the model. This involves using econometric techniques to estimate firm markups, and then use them in conjunction with an exogenous price instrument to investigate the rate of privatization. The context that was used was the Mexican *ejido* privatization that started in 1994 and continues on to this day. Subsequently, I presented a triple-difference approach to the estimation. With that I provide multiple estimation strategies that will plausibly give me a causal estimate, although both with drawbacks as discussed. Although the data is not

available yet, I have plans to acquire them as soon as it is feasible.

References

K. G. Abraham and S. N. Houseman. The importance of informal work in supplementing household income. 2019.

D. J. Aigner. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59, 1973.

A. Alesina and E. Zhuravskaya. Segregation and the quality of government in a cross section of countries. *American Economic Review*, 101(5):1872–1911, 2011.

A. Alesina, R. Baqir, and W. Easterly. Public goods and ethnic divisions. *The Quarterly journal of economics*, 114(4):1243–1284, 1999.

A. Alesina, A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg. Fractionalization. *Journal of Economic growth*, 8(2):155–194, 2003.

Y. Algan, C. Hémet, and D. D. Laitin. The social effects of ethnic diversity at the local level: A natural experiment with exogenous residential allocation. *Journal of Political Economy*, 124(3):696–733, 2016.

J. Alix-Garcia, A. De Janvry, and E. Sadoulet. A tale of two communities: explaining deforestation in mexico. *World Development*, 33(2):219–235, 2005.

H. K. Altinyelken, S. Moorcroft, and H. Van Der Draai. The dilemmas and complexities of implementing language-in-education policies: Perspectives from urban and rural contexts in uganda. *International Journal of Educational Development*, 36:90–99, 2014.

G. Barnes. The evolution and resilience of community-based land tenure in rural mexico. *Land Use Policy*, 26(2):393–400, 2009.

L. I. O. Berge, K. Bjorvatn, S. Galle, E. Miguel, D. N. Posner, B. Tungodden, and K. Zhang.

Ethnically biased? experimental evidence from kenya. *Journal of the European Economic Association*, 18(1):134–164, 2020.

T. Berger, C. Chen, and C. B. Frey. Drivers of disruption? estimating the uber effect. *European Economic Review*, 110:197–210, 2018.

P. Bharadwaj, W. Jack, and T. Suri. Fintech and household resilience to shocks: Evidence from digital loans in kenya. Technical report, National Bureau of Economic Research, 2019.

B. Bhusal, M. Callen, S. Gulzar, R. Pande, S. A. Prillaman, and D. Singhania. Does revolution work? evidence from nepal's people's war. 2020.

W. T. Bianco and R. H. Bates. Cooperation by design: Leadership, structure, and collective dilemmas. *The American Political Science Review*, pages 133–147, 1990.

H. Binswanger and M. Rosenzweig. Wealth, weather risk and the composition and profitability of agricultural investments. *Economic Journal*, 103(416):56–78, 1993.

D. A. Black, M. C. Berger, and F. A. Scott. Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association*, 95(451):739–748, 2000.

J. E. Blumenstock, N. Eagle, and M. Fafchamps. Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. *Journal of Development Economics*, 120:157–181, 2016.

H. Boogaard, R. van der Wijngaart, D. van Kraalingen, M. Meroni, and F. Rembold. Asap water satisfaction index. 2018.

E. Carlson. Ethnic voting and accountability in africa: A choice experiment in uganda. *World Politics*, 67(2):353–385, 2015.

W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. In *Proceedings*

of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, volume 161175. Citeseer, 1994.

M. K. Chen, M. Keith Chen, J. Chevalier, P. Rossi, and E. Oehlsen. The value of flexible work: Evidence from uber drivers, 2017.

J. Conning. Latifundia economics. Technical report, 2001.

C. Cook, R. Diamond, and P. Oyer. Older workers and the gig economy, 2019.

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

J. De Loecker and J. Eeckhout. The rise of market power. Technical report, Princeton mimeo, 2017.

S. Dercon. Income risk, coping strategies, and safety nets. *The World Bank Research Observer*, 17(2):141–166, 2002.

P. C. Dower and T. Pfutze. Specificity of control: The case of mexico's ejido reform. *Journal of Economic Behavior & Organization*, 91:13–33, 2013.

B. Eifert, E. Miguel, and D. N. Posner. Political competition and ethnic identification in africa. *American Journal of Political Science*, 54(2):494–510, 2010.

D. Farrell, F. Greig, and A. Hamoudi. Bridging the gap: How families use the online platform economy to manage their cash flow. *Available at SSRN 3481471*, 2019.

A. Gelan. Trade liberalisation and urban–rural linkages: a CGE analysis for ethiopia, 2002.

GeoNetwork. Geonetwork opensource portal to spatial data and information. 2007.

F. Golooba-Mutebi and S. Hickey. The master of institutional multiplicity? the shifting

politics of regime survival, state-building and democratisation in museveni's uganda. *Journal of Eastern African Studies*, 10(4):601–618, 2016.

J. Hartter, S. J. Ryan, C. A. MacKenzie, A. Goldman, N. Dowhaniuk, M. Palace, J. E. Diem, and C. A. Chapman. Now there is no land: a story of ethnic migration in a protected area landscape in western uganda. *Population and Environment*, 36(4): 452–479, 2015.

S. L. Hoch. *Serfdom and Social Control in Russia: Petrovskoe, a Village in Tambov*. The University of Chicago Press, 1986.

J. D. Huber. Measuring ethnic voting: Do proportional electoral laws politicize ethnicity? *American Journal of Political Science*, 56(4):986–1001, 2012.

N. Ichino and N. L. Nathan. Crossing the line: Local ethnic geography and voting in ghana. *American Political Science Review*, pages 344–361, 2013.

W. Jack and T. Suri. Risk sharing and transactions costs: Evidence from kenya's mobile money revolution. *American Economic Review*, 104(1):183–223, 2014.

E. Jackson. Availability of the gig economy and long run labor supply effects for the unemployed. *Job Market Paper*, 2019.

A. d. Janvry, G. Gordillo, E. Sadoulet, et al. *Mexico's second agrarian reform: household and community responses, 1990-1994*. 1997.

H. Kazianga and Z. Wahhaj. Will urban migrants formally insure their rural relatives? family networks and rainfall index insurance in burkina faso. *World Development*, 128: 104764, 2020.

D. Koustas. Consumption insurance and multiple jobs: Evidence from rideshare drivers. *Unpublished working paper*, 2018.

R. S. Kovats, M. J. Bouma, S. Hajat, E. Worrall, and A. Haines. El niño and health. *The Lancet*, 362(9394):1481–1489, 2003.

E. Kramon, J. H. Hicks, S. Baird, and E. Miguel. Deepening or diminishing ethnic divides? the impact of urban migration in kenya. 2019.

D. Lagakos, A. M. Mobarak, and M. Waugh. The welfare effects of encouraging Rural-Urban migration, 2018.

L.-F. Lee and R. H. Porter. Switching regression models with imperfect sample separation information–with an application on cartel stability. *Econometrica: Journal of the Econometric Society*, pages 391–418, 1984.

G. N. Lesetedi. Urban-rural linkages as an urban survival strategy among urban dwellers in botswana: the case of broadhurst residents, 2003.

J. Light and T. Shevlin. The 1996 grain price shock: how did it affect food inflation. *Monthly Lab. Rev.*, 121:3, 1998.

A. F. Llitjos and A. W. Black. Knowledge of language origin improves pronunciation accuracy of proper names. In *Seventh European Conference on Speech Communication and Technology*, 2001.

R. E. Lucas. Internal migration in developing countries. *Handbook of population and family economics*, 1:721–798, 1997.

R. E. Lucas Jr. On the size distribution of business firms. *The Bell Journal of Economics*, pages 508–523, 1978.

J. M. MacDonald. Entry and exit on the competitive fringe. *Southern Economic Journal*, pages 640–652, 1986.

P. Mateos. Classifying ethnicity using people's names. In *Proceedings to the Social Statistics*

and Ethnic Diversity Conference. Quebec Inter-University Center for Social Statistics and the Institut National d'É tudes Dé mographiques. Montreal, Canada, 2007a.

P. Mateos. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4):243–263, 2007b.

T. McArthur. Direct measurement of efficiency gains from land titling: Procede's effect upon the productivity of mexican agriculture. Technical report, 2016.

A. McKay and P. Deshingkar. Internal remittances and poverty: Further evidence from africa and asia. 2014.

T. B. McKee, N. J. Doesken, J. Kleist, et al. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston, 1993.

A. Michuda. Urban labor supply responses to rural drought shocks: Uber in uganda. 2020.

L. Monasterio. Surnames and ancestry in brazil. *PloS one*, 12(5):e0176890, 2017.

S. Moyo. Land concentration and accumulation after redistributive reform in post-settler zimbabwe. *Review of African Political Economy*, 38(128):257–276, 2011.

V. A. Mueller and D. E. Osgood. Long-term impacts of droughts on labour markets in developing countries: Evidence from brazil, 2009.

P. Mukwaya, Y. Bamutaze, S. Mugarura, and T. Benson. Rural-urban transformation in uganda. *Journal of African Development*, 14(2):169–194, 2012.

S. Murphy. Concentrated market power and agricultural trade. *Ecofair trade dialogue, Discussion papers*, 1, 2006.

A. Nadal. The environmental social impacts of economic liberalization on corn production in mexico. 2000.

C. Nakalembe. Characterizing agricultural drought in the karamoja subregion of uganda with meteorological and satellite-based indices. *Natural Hazards*, 91(3):837–862, 2018.

S. Namyalo and J. Nakayiza. Dilemmas in implementing language rights in multilingual uganda. *Current Issues in Language Planning*, 16(4):409–424, 2015.

R. no. 36996-UG Poverty Reduction and A. R. Economic Management, Southern Africa. Uganda poverty and vulnerability assessment report. 2006.

B. Ogwang, T. Guirong, and C. Haishan. Diagnosis of september-november drought and the associated circulation anomalies over uganda. *Pakistan Journal of Meteorology*, 9 (2), 2012.

OIM. Migration in uganda–a rapid country profile 2013. 2015.

A. Olden and J. Møen. The triple difference estimator. *NHH Dept. of Business and Management Science Discussion Paper*, (2020/1), 2020.

G. S. Olley and A. Pakes. The dynamics of productivity in the telecommunications equipment industry. Technical report, National Bureau of Economic Research, 1992.

E. P. Perramond. The rise, fall, and reconfiguration of the mexican ejido. *Geographical Review*, 98(3):356–371, 2008.

V. Pervouchine, M. Zhang, M. Liu, and H. Li. Improving name origin recognition with context features and unlabelled data. In *Coling 2010: Posters*, pages 972–978, 2010.

C. Plank. Land grabs in the black earth: Ukrainian oligarchs and international investors. *Land concentration*, 184, 2013.

D. N. Posner. Measuring ethnic fractionalization in africa. *American journal of political science*, 48(4):849–863, 2004.

Y. Qu and G. Grefenstette. Finding ideographic representations of japanese names written in latin script via language identification and corpus validation. In *Proceedings of the*

42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 183–190, 2004.

F. Rembold, M. Meroni, F. Urbano, G. Csak, H. Kerdiles, A. Perez-Hoyos, G. Lemoine, O. Leo, and T. Negre. Asap: A new global early warning system to detect anomaly hot spots of agricultural production for food security analysis. *Agricultural systems*, 168: 247–257, 2019.

E. Riley. Mobile money and risk sharing against village shocks. *Journal of Development Economics*, 135:43–58, 2018.

R. T. Rogers and R. J. Sexton. Assessing the importance of oligopsony power in agricultural markets. *American Journal of Agricultural Economics*, 76(5):1143–1150, 1994.

M. R. Rosenzweig and O. Stark. Consumption smoothing, migration, and marriage: Evidence from rural india. *Journal of political Economy*, 97(4):905–926, 1989.

S. Shokoohyar, A. Sobhani, and A. Sobhani. Impacts of trip characteristics and weather condition on ride-sourcing network: Evidence from uber and lyft. *Research in Transportation Economics*, page 100820, 2020.

L. Smith, P. Norman, M. Kapetanstrataki, S. Fleming, L. K. Fraser, R. C. Parslow, and R. G. Feltbower. Comparison of ethnic group classification using naming analysis and routinely collected data: application to cancer incidence trends in children and young people. *BMJ open*, 7(9), 2017.

O. Stark and R. E. Lucas. Migration, remittances, and the family. *Economic development and cultural change*, 36(3):465–481, 1988.

E. Stobl and M.-A. Valfort. The effect of Weather-Induced internal migration on local labor markets. evidence from uganda, 2013.

R. Tangri and A. M. Mwenda. President museveni and the politics of presidential tenure in uganda. *Journal of Contemporary African Studies*, 28(1):31–49, 2010.

R. M. Townsend. Risk and insurance in village india. *Econometrica: Journal of the Econometric Society*, pages 539–591, 1994.

UBOS. The uganda national household survey 2016/17. 2017.

S. Wang, T. H. McCormick, and J. T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

D. A. Wilhite and M. H. Glantz. Understanding: the drought phenomenon: the role of definitions. *Water international*, 10(3):111–120, 1985.

R. E. Wolfinger. The development and persistence of ethnic voting. *American Political Science Review*, 59(4):896–908, 1965.