Berkeley Economic Review

VOLUME X

| Content

Professor Interviews

Research Papers

Cover

Chelsea Gomez-Moreno

| Policies

Submission Policy

- **Format:** Please format your submission as a Microsoft word document or Latex file. We would prefer submissions double-spaced, 12-point, Times New Roman font. Please use the Chicago Manual of Style. Please name your file using the following format: "FirstnameLastname-PaperTitle".
- **Length:** No maximum length for theses or original research papers.
- **Plagiarism:** We maintain a strict zero tolerance policy on plagiarism. According to the University of North Carolina, plagiarism is "the deliberate or reckless representation of another's words, thoughts, or ideas as one's own without attribution in connection with submission of academic work, whether graded or otherwise." For more information about plagiarism—what it is, its consequences, and how to avoid it—please see the UNC's website.
- **Citations:** Use Chicago format for in-text citations and your works cited page. Include a separate works cited page at the end of your work.
- Deadline (subject to change): The deadline for Spring 2021 is Tuesday, February 16, 2021 at 5:00pm PST. Any submissions after the deadline will be considered for our Fall 2021 Edition.
- **Send to:** editor@econreview.berkeley.edu with the subject line "Fall 2020 Journal Submission: [Name], [Paper Title]"
- Please direct any questions you may have to the same email address as above. We will reach out after the above deadlines regarding your acceptance status to our journal. Please note that we no longer accept any op-eds, essays, or articles for consideration.

Disclaimers

The views published in this journal are those of the individual authors or speakers and do not necessarily reflect the position or policy of *Berkeley Economic Review* staff, the UC Berkeley Economics Department and faculty, or the University of California, Berkeley in general.

Copyright Policy

| Staff

Editors in Chief

Parmita Das*
Selena Zhang*

Peer Review

Charlie McMurry*
Anne Fogarty
David Lyu
Edgar Hildebrandt Rojo
Inga Schulz
Kevin Wang
Oliver McNeil
Prakash Srivastava
Ronnie Ghose
Sarah Baig
Sarah Bui
Vatsal Bajaj
Yutao He

Research and Editorial

Staff Writers

Ally Mintzer
Ani Banerjee
Davis Kedrosky
Grace Jang
Jennifer Jackson
Katya Yudina
Konnor von Emster
Najati Nabulsi
Peter Zhang
Ria Bhandarkar
Savr Kumar
Sze Yu Wang
Thomas So
Vasanth Kumar

* *Officers*

| Staff

Editors

Abhishek Roy
Amanda Yao*
Amanda Zhang
Andrew Babson*
Aydin Mahrarramov
Chazel Hakim
Courtney Fung
Kaley Kranich
Kareena Hargunani
Kat Xie
Lucia Dardis
Pallavi Murthy
Pedro de Marcos
Raina Zhao
Raksha Sen
Sean O'Connell

Layout and Design

Angela Li
Anita Megerdichyan
Chelsea Gomez-Moreno*
Christina Xu

Web Design

Aswin Sezhiyan*
Eric Lin
Mandy Zhang
Shikhar Srivastava
Suryaprakash Vengadesa

** Officers*

| Staff

External Affairs

Outreach

Anthea Li
Chelsie Park
Fianna Ton
Henry Isselbacher
Josh Lee
Millie Khosla
Nga Man Arina Sze
Pearleen Wang*
Samarth Bhutani
Vidushee Mishra

Marketing

Caroline Wang
Ihina Ganguly*
Katrina Toups
Rachel Lee
Ryan Kwon
Yash Rajwanshi

Professional Development

Liz Flores*
Sophia Li*

Additional Staff

Finance

Rishi Modi*

Social Chair

Douglas Koehler*

** Officers*

| From The Editors' Desk

Dear BER Reader,

On behalf of the 77 staff members of Berkeley Economic Review's six departments and executive team, we are proud to present the Fall 2020 volume of our namesake journal, Berkeley Economic Review.

As the leading platform uplifting the voices of undergraduate economists around the world, Berkeley Economic Review curates the best research papers to present in our journal. We received an overwhelmingly enthusiastic response from the global undergraduate population this semester, which made our job of selecting the best undergraduate research in the field of economics both challenging and rewarding. Our talented Peer Review team has diligently worked to select some of the best papers and we are delighted to share them with you.

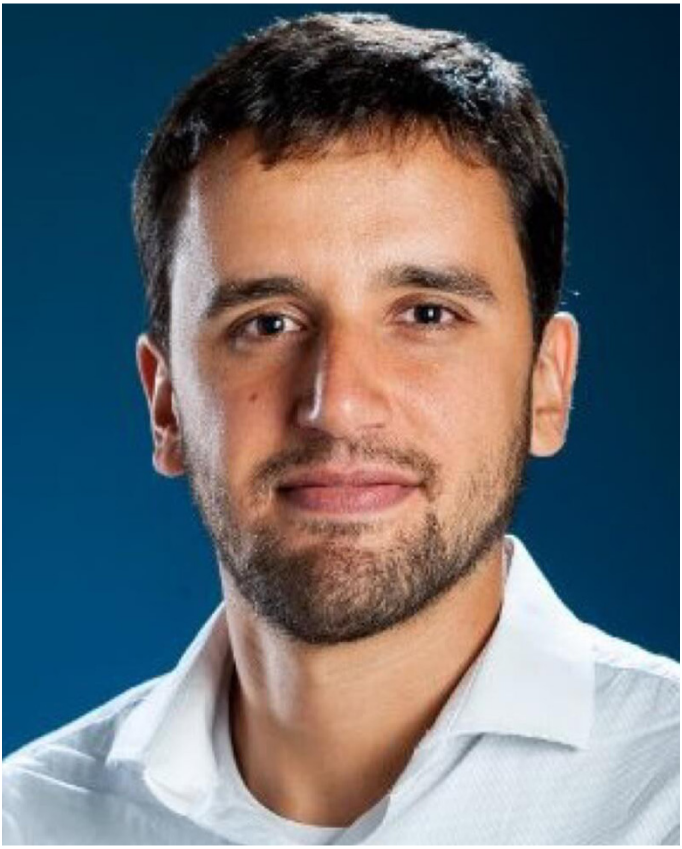
These papers are a testament to the strength and resilience of the global undergraduate economics community in persevering through unprecedented times in world history to pursue their research interests. We hope that you will be similarly moved by the content of our journal and that, within these pages, you find new perspectives on economics issues affecting the globe.

Without further ado, we present to you an important milestone in our journey, the 10th volume of Berkeley Economic Review.

Best,
Parmita Das & Selena Zhang

Editors-In-Chief

Berkeley Economic Review



Professor Fernando Hoces de la Guardia

Interviewed by ***Peter Zhang***

Fernando Hoces de la Guardia is a Postdoctoral Fellow at the Berkeley Initiative for Transparency in the Social Sciences (BITSS) hosted by the Center for Effective Global Action (CEGA). His research interest is evidenced-based policy, particularly in increasing the transparency and reproducibility of policy analysis. Dr. Hoces wishes to highlight the role of CEGA and BITSS in making Open Policy Analysis (OPA) a team effort. BER Staff Writer Peter Zhang interviewed Dr. Hoces over Zoom on October 27th, 2020.

Interviewer: How were you initially drawn to economics and public policy?

Hoces de la Guardia: A long time ago, in Chile—I did my undergrad and master's in Chile, where you don't just go to college, you go to a career right from the start—I started on a track that's called Commercial Engineering, which is basically a dual major between business and economics. I started thinking of going along the business track because I was seeking financial stability. But, when I was two years into it, I discovered economics and realized that you can use this rigorous language to talk about social issues— and you can make a career out of it! So that was my first encounter.

Interviewer: How has learning in Chile influenced how you think of economics today?

Hoces de la Guardia: The career I had in Chile was what you would call a traditional track in economics—doing a master's and following a standard path. Then I spent a total of four years working in the public sector. That's what heavily influenced my thinking about public policy. I describe myself as a policy economist; my training plus my experience was in doing economic analysis for public policy, and that heavily shaped the way that I think about how to bring more rigor into policy debates.

One of the main things is that when I was in Chile, I was thinking that basically economics and public policy is this exercise of doing rigorous analysis to answer policy questions. But I was somewhat surprised that I had the opportunity to serve under two different governments—one from the left and one from the right. When I was doing pretty much the same job, the perceived effectiveness of my work or the perceived rigor of my work was very different under different administrations. That was a motivating factor.

On top of that, I migrated to the US, 7-8 years ago, permanently because my wife is here. When I came here, I was like, "ok, being a policy economist is pretty much a translatable skill." I thought of this as something that's pretty tradable; you can move from one country to another and pretty much the same thing.

And I realized it's really not the case. There's a lot, especially in policy economics, that entails credentialing and the authoritativeness of how you say things and how you're perceived. In Chile, I had some that authority, some of those credentials, and here I did not have the eloquence; I did not have the connections; I did not have anything. I was sort of stripped from that.

I realized "wow! There's a big premium around that," and that means that there is a lot of noise around how policy economists are conducted. There's a lot of space for opaqueness. There's a lot of space of ambiguity, and that was one of my biggest motivations to say, "this whole exercise can be way more transparent, and I suspect that there will be big benefits out of it."

Interviewer: That's a really interesting journey. Given this background that you have in policy economics, looking at both the left and right in Chile and then moving to the United States, how has all of this influenced the work you do right now?

Hoces de la Guardia: After realizing that there's a large amount of opaqueness in policy analysis, there's something a little bit choking or discouraging. When I was drawn to economics, I was drawn by this idea of bringing rigor to public points. I was not driven to bring in rigor to academia. I think academia has rigor, and that's great.

But when I started looking at how this idea of rigor to public policy is brought in practice, I saw that there's a little bit of contempt coming from the academic world to the policy world. It's like, "ah, that's like less rigorous." It's totally fine that it's less novel, but there's no reason why it should be less novel.

And this contempt, I've seen it in Chile. I've seen it here, from the academic world and toward policy analysis world, is what heavily motivates me to say: "wait a minute, to say these analyses are as important, if not more important, than academic work, and we should have exactly the same standards of rigor—how do we do that?

Then, I saw what was going on in the reproducibility crisis in science and all the discussion about how to bring more transparency and reproducibility in science, and I saw the parallel. I draw a lot from the language that they were using, and the tools and the solutions. That's what influenced my decision to bring these ideas into policy analysis.

Interviewer: Could you talk a bit about open policy analysis? What is it and how does it connect to that goal?

Hoces de la Guardia: With the Open Policy Analysis initiative, what we're doing is that we're trying to promote the use of open science practices into policy analysis. It's basically asking for policy analysis to be more transparent and reproducible in a systematic way. We're doing that by building a framework, building a conceptual idea of how to carry out open policy analysis.

Another part is to carry out some of these open policy analyses in practice. We call these open policy analysis projects. So we did one with the wealth tax, with [Professors] Emamual Saez and Gabriel Zucman. We're finishing one with deworming interventions, with [Professor] Edward Miguel and Evidence Action [an NGO]. We're building a pipeline, one about unemployment insurance in a collaboration with the Berkeley Initiative for Young Americans.

Next, we're looking to build a community of practice around open policy analysis. We're not the only ones working on this idea. There's other people who are doing great work. We want to bring everyone together to create a critical mass of people interested in these ideas.

Interviewer: Could you dive in a little bit more into one of these projects?

Hoces de la Guardia: The way that we articulate the framework of Open Policy Analysis is to have 3 high level principles. One principle is open output. Another principle is open analysis. Another one is open materials.

By open output, what we mean is to move away from the traditional report where you put several results in a report and you give the policymakers 10 scenarios. This is the famous economist that says, "on the one hand, this thing, on the other hand, another thing." We want to give the policymakers what from the analyst's perspective is the best representation of the facts. There should be one scenario. That's the idea of open output: to commit to one output, but also to make a clear connection between how that output changes when you change the underlying assumption. So that the open output will be at one fixed output but also in an interactive way such that you can change the underlying assumptions and see how the output changes. That first principle results in an application where you can change the underlying assumption and see how the results change.

The second component is to have an open analysis. By that we mean to have an exhaustive description of how the analysis was carried out a) in narrative form but also b) with equations to add more clarity as to how this analysis is carried out, but c) on top of that to add code, so [sic] combine everything using principles of computer science of literate programming in what's called dynamic documents, to basically combine everything in one place such you can see all the analysis, all the code, and all the equations in a transparent way.

The last one, open materials, is that all these materials should be in an open repository that allows external viewers to reproduce those components with minimal effort. So just do a few clicks, and you get the analysis using open source software with a minimal number of restrictions like having a computer.

Interviewer: In a lot of fields, open sourcing data and methodologies is becoming more common. You're trying to build this practice—why isn't this already commonplace?

Hoces de la Guardia: I thought quite a bit about that and I can only speculate. I think that we're in a suboptimal equilibrium in policy analysis.

First of all, there's a good amount of policy analysis out there that is credible, that is rigorous . . . serious, let's say. But a large fraction is much more advocacy, painting the target right after you shoot the bullet. You justify the conclusion ex-post. I would call all those policy analyses noise.

In the unfortunate equilibrium we're in, those who produced credible policy analysis have little individual incentives to go with a fully open, transparent and reproducible analysis, because locally they are perceived as credible. The tough part is that the larger community—particularly policy analysts that have different ideologies—is less likely to view them as credible.

We end up having parallel worlds, where we have a think tank on the left and a think tank on the right who address the same empirical question and conduct the same analysis to address the same quantitative issues and they get radically different answers. This is what Charles Manski started writing about in the early 2010s.

The unfortunate component of the equilibrium, based more on what I've seen in Chile, is that there's a premium around credentials and authority. The people who have the authority, who have the perceived seriousness, do not want to relegate so easily. The people who are considered serious policy analysts or policy economists have little incentive to disclose everything.

Interviewer: You've painted a world where ideology divides and institution incentives keep in place this poor equilibrium, so what are the next steps to upset that?

Hoces de la Guardia: I see a parallel between that and the open science movement. The open science movement has gained a lot of traction over the last 5-10 years. The incentives are somewhat similar. The incumbent, prestigious institutions, have little incentive to open up materials and open up their work because they're perceived as credible. But if there's a critical mass of discontent with the current practice of science, you will have innovators in the open science space who will do it differently. That's what has happened in the last 5-10 years: a push towards open data, open code, computational reproducibility, replicability of experiments. That's something that has changed radically in the last 10 years.

After the creation of the new principles of open science, there was a push to get the buy-in of funders, to get the buy-in of journals, and that spurred the movement. This is where CEGA—the largest is the Center for Effective Global Action and the initiative is Berkeley Initiative for Transparency in the Social Sciences [BITSS]—plays a key role. The center is focused on bringing the best evidence possible to alleviate poverty around the globe. With that motivation, CEGA created BITSS to bring more credibility to social science. The next logical step is to move from evidence policy change to policy analysis to improve the credibility of policy analysis. That's where CEGA and BITSS are playing a key role.

Interviewer: Let's say that we create this shift in policy analysis. How do you envision this shifting policy?

Hoces de la Guardia: We want to create an environment where we only have these credible policy analyses around. Now, there are a multitude of facts lying around. That ultimately creates a situation where policymakers can choose their own facts.

Interviewer: Alternative facts!

Hoces de la Guardia: It has been accentuated in the U.S. over the last four years, but this is not something specific to the last four years. I think it relates this contempt from academia to policy analysis. It's kind of agreed that policymakers can choose whatever fact they want, when discussing minimum wage, wealth tax, funding higher education, or whatever issue is under debate. It's kind of agreed that now policymakers can choose whatever they want. What I see is that as open policy becomes a norm, policymakers will be forced to face the best representation of the facts. If the policy analyst community, supported by the academic community, can say this is "the best representation of the facts as we know them today," then policymakers will be forced to say, "I support the minimum wage, and someone opposes the minimum wage, and we both agree on what the facts are." That will shed light on the normative preferences of the different policymakers. What do they stand for?

That's our agenda for the next 10 years.

Interviewer: Suppose I'm a pessimist and I think that, even with consensus among policy analysts, someone like President Donald Trump still won't care.

Hoces de la Guardia: I thought quite a bit about that and that is a possible scenario. My response would be that the way I see things—and this is obviously painted by my current initiative—is that under a world where there is open policy analyses, it's much harder for people like Donald Trump to exist. Some people out there might still give some credibility to the different facts laid on the table. And it's precisely because we do not have 100% agreement of what's the best representation of analysis.

I don't think OPA is going to prevent something like Donald

Trump from happening in the future. I think it's going to make it less likely. It will shed light on when there is a blatant divergence from the facts.

This is not a magic bullet. But, I think it is something of a step in the right direction to prevent the spread of misinformation.

Interviewer: If I am a policy researcher who is interested in your project and I want to get involved or follow your work, where could I do that?

Hoces de la Guardia: Definitely check out the Berkeley Initiative for Transparency in the Social Sciences. That's the initiative that is supporting open policy analysis. Also check out the Center for Effective Global Action.

If you want to be particularly involved in open policy analysis, you can visit the background on our webpage and you can be involved as an undergraduate research assistant each semester. You can suggest a policy analysis for us to open up or for us to collaborate with others who open up. You can also subscribe to BITSS news so you're up to date on everything we're doing.

Interviewer: Great. Thank you for your time!

Professor Kim

*Interviewed by **Sze Yu Wang***

Professor Kim is an Associate Professor of Economics at the Hong Kong University of Science and Technology, and an Assistant Professor at Cornell University. His research focuses on development economics, health economics, and education economics.

Interviewer: I'd like to begin by speaking about your personal journey in economics. What was it that initially drew you to economics, and what experiences influenced this interest?

Kim: I started off as a medical doctor, and I practiced for 2-3 years. One thing that really affected me was that, at the end of my time in medical school, doctors went on strike in Korea. I was shocked! How can doctors go on strike while the patients are dying? How is this possible? Of course they didn't shut down the ER and the ICU, but this was still a shock to me. I then realized that it was to do with issues in health policy, and I saw how important public policy is.

In my last year as a medical student, I also worked at a breast cancer clinic. What I noticed was that those from rich areas with higher education and better income came in for screenings more often, and were able to afford expensive cancer treatment. For poor people, it wasn't like that. One day, a woman came in. She looked around 65 years old, with dark skin—not dark skin from sun-tanning, dark skin from working in the sun. I looked through her documents and saw that she had been referred from a hospital in the rural area. When I touched her breast, I immediately realized that her axillary nodes were already super protruded and expanded. Anybody, even a medical student like me, could quickly realize that this was terminal cancer. But she asked me: "Do I have cancer?"

That made me very sad. She was scared, and I don't think she could have lived long. South Korea is a pretty equal society compared to other countries, but I could see how, based on level of education and socioeconomic status, cancer survival was definitely not equal. I felt this was unfair, and I wanted to study the effects of public policy and inequality.

Think about the Chinese famine in the 1960s and 70s. Bad public policy can kill millions of people. Let's check how many people died from COVID in the US: 234,000 people. How many people died in Korea? 400. How many in Hong Kong? Less than that. How many lives can I save as a medical doctor in one lifetime? At most 200, maybe 300. But as an economic policy-maker, the number of lives I can help is much bigger than one single doctor.

Interviewer: What research projects are you working on now, and how are you collecting your data?

Kim: My research involves two types of approaches. The first one we often call primary data collection with randomized controlled trials (RCT), which is where I collect the data myself. When I was a medical doctor, randomized controlled trials were everywhere. But when I first started economics, RCT was not popular. With this approach, I may provide treatment for one group, and leave another as they are.

For example, in one paper I published two years ago in Science, we randomly provided tuition money for female secondary school students in Malawi. We then followed these people in the long run. One of our major findings was that the group that received the tuition money showed much better decision making quality, and made much better and much more careful investments. The implications of better economic decisions are huge right? We were able to show how secondary school education improved the quality of life for these women, which as far as I know, is the first causal evidence of secondary school education. Most research in this area focuses on primary school education.

Another approach in my research is to make use of natural experiments and large comprehensive data, which usually means looking at interventions created by the government. To study these effects we need to find a suitable control group. Government policy only affects certain groups of people based on income, residential area, family structure, age, and so on. The key is that most of these policies usually have a cut-off point. For example, if a government introduces a policy based on your income, there will be a group of people just above the cut-off, and a group of people just below the cut-off. These people are super similar, but are treated completely differently, which we can use to our advantage. This is one way we use to estimate the effects of government policy.

Interviewer: You've done a lot of interesting work in Malawi and Ethiopia. How did you first get involved with that, and what inspired you to look at those places in particular?

Kim: Fifteen years ago, I didn't have any connections in Ethiopia and Malawi. One of my friends randomly invited me: "Byrant Kim can we go to Africa together?" (laughs) I followed him, and we visited Malawi and Ethiopia. The reason he invited me was to visit the hospitals there, and surprisingly, one of the best hospitals in Malawi was actually run by Korean missionaries. They asked me if there was any research I could do in their hospital, and that was how our project started. One study we did was the cash-transfer project for secondary school female students which I mentioned, and another was on the impact of male circumcision on risky sexual behaviour and HIV prevention. These two programs were implemented in Malawi almost 10 years ago, and I still follow up on these people to see how the program affected their lives.

In Ethiopia, we introduced some health and nutrition programs. Unfortunately, we had to stop these projects because of serious political protests in the area. More than 500 people were killed nearby, and one of the American postdocs from U.C. Davis was killed on a road that we often took. It could have been myself, my wife, or anyone on our team. It was too dangerous, and we had to evacuate permanently.

Interviewer: How do you think the COVID-19 pandemic might disproportionately affect people in lower socio-economic classes?

Kim: Within the US, there is clear evidence that black people and people in lower socioeconomic classes are more likely to be infected by COVID-19. Just look at the incidence and mortality of COVID-19 by level of income, or by residential area—the effects are obvious. But I think that education could have much bigger negative consequences that unequally affect people.

What are the consequences of not coming to school? There is lots of evidence that schooling has important impacts on mortality, longevity, crime rates, and future income. It also affects non-cognitive abilities, because school is basically where you learn to talk to people. People are losing these opportunities. Richer students are able to receive education through other channels, but for poorer students, school is everything. The infection rate of course is unequal, but I think a more serious impact could be driven by an unequal opportunity of education. This must be studied in the future.

Interviewer: What about people in developing countries?

Kim: Good news! COVID-19 differs by age group. Under 65, the mortality rate is very close to influenza. Influenza infection is tough, but it is very unlikely to kill you. If you are older than 65, however, things change a lot. The mortality rate is almost 10 times more serious. Thanks to the age structure of developing countries, it seems like their mortality rate is much lower than in European countries, where most of the infected are elderly people. Developing countries have worse healthcare, which definitely has an impact, but because of their youthful age structure things are not as serious as in Western countries. Look at Ghana: 48,000 cases but only 320 deaths, which is pretty low.

Interviewer: You started with a degree in medicine, you travelled all over the world, and you've worked as an ER doctor, a public health physician, and now an economist. What advice might you have for college students who find that their interests transcend the typical boundaries of their major?

Kim: You guys are 21, 20. Don't be afraid to explore other areas. When I first studied economics I was 25, which is still young, but older than you guys. I decided to invest at least two years studying economics, even though I didn't know how capable I was, or even how much I truly liked it. By investing these two years, I learned that this was what I wanted to do.

Take advantage of UC Berkeley's flexible system for choosing majors and taking courses. Explore yourself! Every summer you have freedom, so explore the world. If I were you guys I would spend one summer as a backpacker. Maybe it's difficult during COVID-19, but after it is gone, it's definitely worth a try. I backpacked to Europe, Africa, and the Middle East, which could be dangerous nowadays, but when I was your age the world was a lot safer in terms of terrorism. Take advantage of each summer and study yourself. Find what you like, and what makes you happy. Oh and try to find a girlfriend! Don't be too shy. (laughs)

Professor Steven Vogel

Interviewed by ***Ally Mintzer***

Steven Vogel is Chair of the Political Economy Program, the Il Han New Professor of Asian Studies, and a Professor of Political Science at the University of California, Berkeley. He specializes in the political economy of advanced industrialized nations, especially Japan.

Interviewer: I'd like to first talk about your personal journey and how it has led to your research on political economy. Can you describe your background and what experiences led you to discover your passion for market governance and Japanese politics?

Vogel: I was taken to Japan at age 13 against my will by my parents. I was going to study at an American high school to pretend I wasn't in Japan, but the school would have forced me to study Japanese. So I figured if I had to study Japanese, I might as well go all the way and attend a Japanese high school. When I came back to the US a year later, I had reverse culture shock, and eventually returned to Japan to finish high school. I majored in international affairs at Princeton University and spent the summer between my junior and senior year interning for a member of the Japanese Diet, which is their parliament. This is what got me interested in Japanese politics. I hit the jackpot because this legislator was then appointed as Defense Minister. After I graduated, I went back to Japan as a reporter for the Japan Times, and then moved to France as a freelance reporter, editor, and research assistant.

Interviewer: What prompted the transition from being a reporter to getting a PhD?

Vogel: When I left college, I had a feeling I would be going back to graduate school eventually in California, but I did not know what discipline. My father is an academic, so that probably influenced me. I loved being a reporter, but my one frustration was that whenever I became interested in a new topic, I had to move on. Now as an academic, I have the exact opposite problem: once I launch a major research project, I'm stuck with it for the long haul. I find that my personality falls somewhere in between the two. I like the excitement of new topics, but reporting was too surface-level. I was three years out in the field when I came back to graduate school, but I figured if I wanted to do a PhD, I had better do it then. I did miss reporting because it is an exciting job for someone in their 20's, just absorbing information constantly.

Interviewer: I feel like I'm in a similar position. I would love to attend graduate school some day, but I do not know which discipline yet.

Vogel: It's great to first get a job right out of college, and I generally don't encourage my students to go to graduate school right away. Obviously, the more interesting the job the better—something where you're just absorbing information like reporting or consulting. Those are great first jobs that teach you what you like and don't like. It's hard to figure that out in college because you're developing skills like taking tests—that you may never do again. Once you're in the professional world, you'll realize, hey, I'm great at presentations and terrible at memos, for example. Within a year or two you can really get a direction; at least that's what I find with my former students.

Interviewer: So how did you decide to get your PhD in political science?

Vogel: My interests spanned from politics to economics to philosophy. The way I decided my discipline, which wasn't the most logical, was that I figured if I did a PhD in economics or philosophy and did not like it, I would be stuck. But with political science, I could hedge my bets and do some philosophy and economics too. Even at the time I didn't think it was a rational decision, but what surprises me looking back is that I actually knew my interests very well.

Interviewer: When did you realize you wanted to teach political economy and Japanese politics?

Vogel: When I was doing my PhD, I did not think of myself as a Japan specialist. I started in international relations and ended in comparative political economy, although I didn't realize it at the time. My dissertation was on the deregulation movement that started in the 1970's in the US and spread throughout the world, focusing mainly on Britain and Japan. I figured Japan would be easy since I had the language skills. I focused on telecommunications and finance initially, then extended my dissertation to include broadcasting, utilities, and transport, also in France, Germany, and the US. By that point, I was knee-deep in political economy. I applied for comparative and Japanese politics jobs, and was hired to teach Japanese politics at UC Irvine. Ironically, I had never actually taken a class in Japanese politics, but I was well-equipped to teach since I went to high school there, worked in politics myself, and was a reporter there. I have been teaching comparative political economy and Japanese politics ever since!

Interviewer: What are some areas you are researching right now?

Vogel: I'm kind of at a crossroads. After I finished my *Marketcraft* book, I did pieces on Japanese labor, Japanese corporate governance, the regulatory roots of inequality, and entrepreneurship in the United States. I've been writing a lot of op-eds because I'm worried about the state of our country, and I try to give commentary on policy. I think my next project will be on economic inequality, very broadly defined, across industrial countries. I still haven't determined which countries or how theoretical or empirical it will be. The demands for data will probably not be as high as my first book on deregulation and my second book *Japan Remodeled,* which involved more extensive original research.

Interviewer: You mentioned your recent book *Marketcraft, How Governments Make Markets Work*. Can you explain what you mean by "marketcraft?"

Vogel: "Marketcraft" is my word for market governance, which is how markets are structured by governments, firms, and individuals, including laws, regulations, business practices, and social norms. Markets are inherently governed; they don't work without rules. It sounds pretty simple, but it's amazing how much both intellectuals and policymakers just ignore that and say that we need less government and more markets—as if those two came together. I used the term marketcraft because I was trying to evoke statecraft, which is similarly critical for the welfare of a nation. For instance, what was the greatest economic success story and the greatest economic failure in the United States over the past four decades? For failure, it would be the global financial crisis or Covid-19, and the success would be the digital revolution. Both the financial crisis and the digital revolution were products of marketcraft, which gives you a sense of the scale of the consequences of marketcraft. Amazing things can happen if you get marketcraft right, and terrible things can happen if you get it wrong.

Interviewer: You're also co-chair of the UC Berkeley Network for a New Political Economy to develop a new intellectual paradigm as an alternative to neoliberalism. Neoliberal is a word that is so ubiquitous yet many describe it differently. So first, how would you define neoliberalism, and what would a post-neoliberal society look like?

Vogel: I was hoping for easy questions! I have avoided the term neoliberal until recently and instead used "market liberal" for precisely that reason. There's so much literature on it; it's like a code word with multiple meanings. I believe there are five or six different main definitions of neoliberalism, but two prominent ones include an intellectual movement beginning in the 1930's and 1940's with Friedrich Hayek and Milton Friedman and a political project beginning with Margaret Thatcher and Ronald Reagan. In a nutshell, I would say that neoliberals believe that it is desirable to have less government and more markets. This ideology has caused incredible damage for analysis, for policy, and for public welfare.

An intellectual replacement for neoliberalism would be market institutionalism. I'm not pretending I came up with this; there are entire subfields in economic sociology, economic geography, or institutional economics that treat markets as institutions, not natural institution-free spaces. In terms of policy, market institutionalism suggests a "predistribution" agenda.

Do we let markets do their thing and then have a redistributive tax or welfare policy, or do we design markets to achieve better outcomes from the get-go? A real market liberal like Hayek or Friedman would go bonkers, saying we should not mess with the free market! I would counter that there is no such thing as a free market, as markets are inherently governed. If you believe in equity or sustainability, those goals should not be outside the realm of market design. This idea totally transforms the progressive agenda as you're not only addressing social regulation, tax policy, and welfare spending but how to redesign markets for a more equitable and sustainable market economy as well. To be concrete, this would include changing labor regulations to balance the power between employers and employees. It would also mean reimagining corporate governance so corporations maximized the welfare of a broader range of stakeholders rather than a narrow range of shareholders and stock options-maximizing executives.

Interviewer: You teach a course at UC Berkeley on market governance and the digital economy. It was recently announced that the DOJ is filing an antitrust lawsuit against Google. What are your thoughts on this and the future of possible Big Tech break-ups?

Vogel: I have a lot of thoughts about antitrust. I'm not sure I'm ready to dive into the weeds of Google's exact practices, but it is clear that we need more aggressive antitrust enforcement. If you look at the US economy over the past few decades, there has been a gradual increase in market concentration which has led to weaker macroeconomic performance. I also believe aggressive antitrust policy would be good politically, which is controversial because most antitrust experts in economics say you shouldn't consider market power as a political problem. Yet a high concentration of market power correlates with a high concentration of political power, which is not good for our political system.

There is a valid question about what to do with Big Tech. These companies became so big because they offered innovative services, lots of people like them, and their network effects are not inherently bad. But if they are leveraging their market power into anti-competitive practices, that's a problem. The US should become much tougher on merger approvals; next time Facebook wants to buy Instagram, agencies should think twice. The authorities need to consider not only how big these companies are at the time, but their potential to become a competitor—whether the acquisition is an attempt to prevent competition. One should be cautious with breakups, however; would the harm of breaking them up outweigh the benefits? I don't have a great answer, but I lean towards some action because there is pretty good evidence of anti-competitive practices by Big Tech firms. This might be slightly radical, but it's probably a good principle that you can run a marketplace or participate in a marketplace, but you can't do both.

Interviewer: How is it teaching here at UC Berkeley where you got your PhD?

Vogel: Oh, I love Berkeley! We do not hire our own PhD's, so I had to do six years of hard exile before they let me come back. It was great to come back because I knew the department, a lot of my own professors were still around, and the department was an intellectual match.

My first fall of my PhD, wanting to come to California, I was actually deciding between here and Stanford. Again, not necessarily rationally, I chose Berkeley partly because it had a bigger department and I figured someone could help me determine my interests. Berkeley seemed very foreign at first, being from the East Coast. To make matters worse, one of my high school friends from Japan was starting his MBA at Stanford. I used to get calls from him on a daily basis saying, "I don't know whether to go to the Dean's social, you know, or this champagne lunch with the President." That did not sound like my experience. So I was jealous for about three or four days and then I had an epiphany, realizing Berkeley is really what I wanted. It took me a little while to figure this out, but I haven't looked back since. Pre-Covid, I used to go to Sproul Plaza at lunchtime and just soak up all the energy. I became a fan very quickly.

Beyond Damages: Explaining Variation in States' FDI Loss from Adverse ISDS Rulings

Connor M. Brennan

Georgetown University

Table of Contents

Table of Contents

1 Introduction

Bit by bit, investor-state dispute settlement (ISDS) provisions have become troublesome for states seeking to attract foreign direct investment (FDI). When a state loses an ISDS case, foreign investors similarly lose confidence in that state's promises to protect foreign investors, thereby dissuading future FDI. However, this loss of FDI inflow varies greatly: Kyrgyzstan lost 60.4% of its FDI following its 2013 loss in Beck v. Kyrgyzstan while Bolivia[1] lost only 15.6% of its FDI after its 2014 loss in Guaracachi v. Bolivia. Why do states lose more FDI after losing some ISDS than after losing others?

Although ISDS provisions within bilateral investment treaties (BITs) are intended to increase FDI, a rich literature has developed to demonstrate that states lose FDI in the event of ISDS arbitration, both when a case is initiated and when it is decided against them[2]. Put simply, states' BIT promises are only valuable if they are kept. However, the nature of ISDS cases has changed considerably in the last decade. Instead of alleging direct expropriation, which occurs when a state seizes legal ownership of an asset without due compensation (i.e., nationalization of a factory), the majority of ISDS cases now allege indirect expropriation. [3]Indirect expropriation cases arise when states, instead of directly seizing ownership of an asset, indirectly eliminate an asset's value through regulatory means[4]. Recent literature suggests that states lose different amounts of FDI when different types of cases are initiated against them[5]. However, there has been no research on the variation of FDI loss after cases are decided[6]. My paper

1 Here, Kyrgyzstan lost its case on grounds of direct expropriation while Bolivia lost its case on grounds of indirect expropriation. I will discuss these two types of claims at length in my paper.
2 Allee and Peinhardt 2011.
3 Pelc 2017.
4 Friedman Prager and Popova 2019.
5 Kerner and Pelc M.S., 2019.
6 This hole in the literature is due to both the recency of these emergent trends and the focus in the literature on more immediate regulatory chill when cases are initiated. Scholars such as Jennifer Tobin (2018) argue that firms initiate ISDS

explains this variation and fills the gap in the literature about different losses of FDI upon losing different ISDS cases.

I argue that states lose more FDI after losing indirect expropriation ISDS cases than after direct expropriation cases. Oftentimes, direct expropriation only compromises the property rights of a single investor for a stated public purpose, minimizing concern among other unrelated investors. Meanwhile, guilty verdicts in indirect expropriation cases concern regulation pertinent to a wide variety of industries. Indirect expropriation cases concern issues ranging from discriminatory taxation and licensing requirements to public health and environmental policy, all of which affect a wide range of foreign investors. So, when a state loses an indirect expropriation case, the case itself pertains to matters affecting a much wider scope of investors than the loss of a direct expropriation case[7]. Therefore, if a state loses an indirect expropriation case, it will lose a significantly greater magnitude of FDI inflow than what is lost after losing a direct expropriation case. However, regulation that attracts indirect expropriation challenges may have real public policy goals, such as when Philip Morris challenged Uruguay's requirements that cigarette packaging include large health warnings. Thus, states are faced with balancing the benefit of domestic policy against the threat of FDI loss should those policies be challenged as indirect expropriation.

The evidence bears out the intuition of my hypothesis that losses of indirect expropriation cases cause a greater reduction of a country's FDI inflow. In a linear fixed-effects model, I will examine the 60 ISDS cases that allege either indirect or direct expropriation by running regressions of these respective cases and their verdicts against respondent countries' FDI inflows. Specifically, I will be using the

cases to impose immediate costs on states, forcing states to abandon their policies and deter others from enacting similar proposals.

7 The importance of ISDS cases for other investors is the information about the government's action that is conveyed in the rulings, not necessarily the actual penalties involved.

model developed in Todd Allee's and Clint Peinhardt's work on the subject[8], which has been cited as the benchmark of statistical analysis of the impact of international investment arbitration on FDI[9]. I will use data from the World Bank's World Development Indicators and the Global Economy Project to update control variables in their model, such as GDP growth, capital account openness, population growth, and property rights, among others. Through this analysis, I find evidence to support my argument, which I will then illustrate using case studies that examine specific examples of investors' reactions to the issues implicated in both direct and indirect expropriation cases.

If states lose more FDI after guilty verdicts in indirect expropriation than in direct expropriation cases, there is a policy implication in both scenarios. Since direct expropriation rulings harm states' FDI inflow to a lesser degree, a state has more latitude to challenge those rulings unilaterally. Meanwhile, to avoid the significant reductions in FDI from losing indirect expropriation ISDS cases, states should dedicate more legal resources to fighting off indirect expropriation cases than direct expropriation cases. If indeed a state does lose on indirect expropriation, governments should have a greater incentive to attempt to mitigate the subsequent FDI loss. This could include spending considerable resources to explain the regulations in question or attracting more foreign investment through costly and credible policy incentives. When considering new regulations, countries should weigh the different magnitudes of FDI loss, associated with losses of different types of ISDS cases, against the costs of allaying the ruling's impact[10].

This paper proceeds in five sections. First, I review the existing literature on the relationship between ISDS cases

8 Allee and Peinhardt 2011.

9 Kerner and Pelc M.S., 2019.

10 Alternatively, states should more actively seek the help of NGOs and intergovernmental partners in indirect expropriation cases to avoid losing. An example of this is in Philip Morris v. Uruguay, where Uruguay's legal defense was largely funded by Bloomberg Philanthropies, the Bill & Melinda Gates Foundation, and the Pan American Health Organization. See Kelland 2015 and Mitchell 2014.

and respondent states' FDI. Second, I argue that the type of ISDS case lost is the primary determinant of a respondent states' consequent loss of FDI. I then provide empirical support through statistical analysis, as well as present case studies to illustrate my argument's causal mechanism. Finally, I consider some limitations of my argument and discuss the implications that my findings bear for policymakers.

2 Literature Review

Comparing states' FDI inflows and their ISDS records is nothing new. Beginning this academic debate, Allee and Peinhardt (2011) found that states lose considerable FDI inflow at two specific points in ISDS arbitration: when cases are initiated against them, and when (or if) the case is ruled against the state[11]. Their model found that a single ISDS case can offset the gains of FDI from signing a BIT. Therefore, Allee and Peinhardt concluded that an ISDS claim against a state signals that the threat of arbitration will not constrain the state from violating their treaty obligations. Intuitively, their argument makes sense—in line with credible commitment theory, a state's credibility to uphold their BIT promises only goes as far as their demonstrable adherence to their assurances[12]. That is to say, if a state has an ISDS case initiated against them, investors may react to some degree based on what is known about the case. However, if a state loses that ISDS case, it is a verified affirmation that the state betrayed their commitments. Allee and Peinhardt conclude that ISDS is more than just a means by which foreign investors can recover damages from unjust state policy; it is also a signaling mechanism for a larger audience of prospective investors.

Not only are ISDS cases taken to represent a state's future propensity to backtrack on their promises to foreign investors, but the potential consequences of facing ISDS

11 Allee and Peinhardt 2011
12 Stevens and Cooper 2009.

arbitration also influence state policy[13]. Confirming Allee and Peinhardt's initial findings, Aisbett et al. (2015) developed a hypothesis in which "BITs [act] as deterrents" of policies adverse to foreign investment[14]. According to their paper, states weigh the benefits of various policies against the costs of an ISDS challenge to that policy—especially if any such rulings go against the state[15]. When these potential costs deter a state from implementing such policies, especially after similar experiences of other states, the literature defines this as "regulatory chill.[16]" The Aisbett et al. hypothesis was reflected in a 2016 study by Van Harten and Scott, in which policymakers acknowledged that they consider the costs of potential investment arbitration when making policy decisions[17].

Although the work of Allee, Peinhardt, and Aisbett et al. on the loss of FDI inflow ensuing ISDS arbitration is informative, their studies mistakenly treat all ISDS cases equally. In fact, Kerner and Pelc (2019) use the same models to show that different cases cause different losses of FDI following ISDS initiation[18]. Specifically, they find that states experience an investment slowdown following the initiation of direct expropriation ISDS cases, while the initiation of indirect expropriation ISDS cases cause no similar effect[19]. Having affirmed that the type of case can affect the subsequent loss of FDI at the initiation stage of ISDS, Kerner and Pelc unfortunately do not explore whether or

13 There are a number of significant costs with being brought up before ISDS, including considerable legal fees, damages regularly in the hundreds of millions of USD, and, as demonstrated, loss of FDI. It is the variation in the third components of these costs, loss of FDI, that is the subject of my investigation.
14 Aisbett, Busse, and Nunnenkamp 2017, 122.
15 Tobin 2018 illustrates this cost by demonstrating that plain packaging ISDS litigation has delayed or deterred similar policies in several states.
16 Tienhaara 2017.
17 Van Harten and Scott 2016. See interview excerpt.
18 Kerner and Pelc M.S., 2019.
19 Along with other authors such as Janeba (2019), Kerner and Pelc (2019) allege a rise of frivolous ISDS litigation intended to dissuade regulatory action rather than compensate for real damages. While these authors argue that more low-merit cases reduce the signaling power of ISDS, they do not examine fluctuations in foreign investment after an arbitral body rules that an investor's claim has considerable merit.

not this discrepancy is replicated when cases are decided, leaving a hole in the literature.

This hole is especially pertinent in light of Allee and Peinhardt's assertion that the ruling of a case provides investors a clear verdict on a state's true culpability and hence causes a market reaction of its own[20]. With respect to indirect expropriation, there is a considerable space for rulings to clarify whether or not a state acted unjustly. As the Organization for Economic Cooperation and Development notes, "the line between the concept of indirect expropriation and governmental regulatory measures not requiring compensation has not been clearly articulated and depends on the specific facts and circumstances of the case," which are revealed not when a claim is initiated but, rather, when a case is concluded[21]. Most BITs remain very vague about what exactly constitutes indirect expropriation[22], allowing arbitrators to determine its interpretation on a case-by-case basis. For example, the German model BIT (and, until 2012, the US model BIT) defines indirect expropriation as measures which would be "tantamount to expropriation," the Canadian model BIT describes it as measures "that have an effect equivalent to direct expropriation without formal transfer of title or outright seizure;" and the Italian model BIT does not even bother to define the term[23]. Meanwhile, definitions of direct expropriation are fairly detailed[24]. Dolzer and Stevens (1995) note that this ambiguity on indirect expropriation leaves

20 Allee and Peinhardt 2011.
21 OECD 2004.
22 Of course, vagueness is in part a purposeful design of indirect expropriation clauses, as they are meant to cover all alternative ways that a state may compromise the value of foreign investors' assets. See Tienhaara 2017.
23 Complete copies of these BITs can be found on UNCTAD's International Investment Agreements Navigator through their Investment Policy Hub. See Investment Dispute Settlement Navigator | UNCTAD Investment Policy Hub.

24 As Pelc (2015) notes, ISDS was originally established in response to waves of nationalizations in the 70s and 80s. In fact, indirect expropriation cases were rather uncommon until it was alleged in Metalclad v. Mexico in 1992, and since 2000 indirect expropriation claims have been featured in 70% of ISDS cases. For a full database of ISDS cases, please see UNCTAD's Investment Dispute Settlement Navigator through their Investment Policy Hub.

foreign investors with insufficient information to judge whether or not a pending case signals a threat to their investments and, as such, insufficient information to decide whether or not to continue business operations in a host country[25].

Nondescript indirect expropriation clauses not only cause uncertainty for investors but also leave policymakers in a state of ambiguity. Scholars have described at length how legitimate public policy action might leave states vulnerable to indirect expropriation ISDS challenges—including all the costs of litigation. Jennifer Tobin (2018) describes how Uruguay and Australian cigarette plain packaging laws, intended as public health policy, allegedly indirectly expropriated the intellectual property of companies who could no longer include their trademark on their products[26]. In Indonesia, the government's attempts to develop its mining industries from extraction to processing through export controls quickly attracted an ISDS challenge by Newmont Mining, claiming that the trade policy stalled production at their mines and therefore indirectly expropriated lost revenues[27]. Uncertain whether or not regulation is permissible within their BIT obligations, states must weigh the benefits of such policy against its likelihood to attract ISDS challenges, and the costs associated with litigating and losing a case.

Nothing in the existing literature explains why states lose different amounts of FDI inflows when they lose ISDS cases. Could the difference lie in the type of case?

3 Argument

My argument proceeds in three parts. Firstly, I contend that direct expropriation cases do not convey actionable information for many foreign investors. Secondly, I argue that the regulatory nature of indirect expropriation cases makes the information conveyed by any ruling against a given state applicable to a wider scope of foreign investors. Thirdly, I argue that indirect

25 Dolzer and Stevens 1995, 99.
26 Tobin 2018.
27 van der Pas and Damanik 2014.

expropriation rulings convey more actionable information than direct expropriation cases, due to the ambiguous legal definitions of indirect expropriation.

The FDI impact of an ISDS ruling against a state is a function of two variables: (1) how widely applicable the issues of the case are to investors and (2) how much additional information that ruling provides to investors. When a case is decided against a state, it is clear that the state was willing to violate its BIT promises; but just how much investors care about that violation depends on how threatened their investments are. As Heil and Robertson (1991) describe, the degree by which a market signal influences an investor's behavior depends on just how relevant the information is to their business operations[28]. Applying this idea to my puzzle, I argue that the violations involved in indirect expropriation cases pertain to a larger audience of foreign investors than the violations involved in direct expropriation cases. Furthermore, I argue that rulings of indirect cases provide investors with information they did not have pre-ruling; combined with the greater applicability of indirect cases to a larger audience of investors, a "loss" verdict consequently causes a greater magnitude of FDI reaction. If a state loses an ISDS case alleging indirect expropriation, then it will see a greater reduction in FDI inflow than it does in losing a case alleging direct expropriation. The stronger signal sent by an indirect expropriation loss in regard to a state's conduct raises concerns for the business operations of a wider audience of foreign investors.

3.1 *Variables*

In my argument, the dependent variable is a country's annual FDI inflow, as published by UNCTAD in millions of US dollars[29]. Variation in the independent variable, whether the state loses an indirect expropriation ISDS case or a direct expropriation ISDS case, causes notable variation in the dependent variable. I also define the loss of a case, in either

28 Heil and Robertson 1991.
29 UNCTAD "FDI Statistics."

indirect or direct expropriation, as when an arbitral body explicitly rules in favor of the investor. This contrasts with other analyses[30], which have counted settled cases as losses and assumed the reputational loss of a state losing or settling a case to be the same. Considering the wealth of literature that suggests that developing countries may back down from ISDS challenges and settle to avoid paying the considerable legal fees, I reject those other analyses in favor of my narrower definition of a "loss."[31] Imperative to my discussion, the impartial and ruling of an arbitral body in indirect expropriation cases offers considerably more information upon which foreign investors act, given the ex ante ambiguity of what policy behaviors actually constitute indirect expropriation[32]. When cases are settled, it would be unclear to foreign investors whether the state indeed violated its BIT promises, or if it merely found the cost of settlement to be less than the cost of ongoing arbitration. Because of this, I consider cases explicitly lost by a state.

3.2 *Causal Logic: Where Direct Expropriation Falls Short*

Cases of direct expropriation could give rise to two potential fears among foreign investors: either that their assets could be similarly expropriated, and that any such takings will not be compensated[33]. However, if these fears do not apply to a large audience of foreign investors, states will not lose much FDI.

I will first address fears of uncompensated expropriation from direct expropriation cases. Oftentimes states and

30 This methodology was followed in both Allee and Peinhardt (2011) and Aisbett et al. (2017).
31 Bonnitcha, Lauge, and Poulsen 2017.
32 While some forums for ISDS, such as ICSID, are transparent and publish detailed findings of cases, others such as UNCITRAL do not disclose such information. Regardless of a decision's detail on the proceedings, they almost always release a basic ruling of whether or not a state is found guilty, which will be crucial to my argument.
33 Picht and Stüven 1991.

investors use ISDS not to conclude whether or not an uncompensated direct expropriation has occurred, but rather to determine a fair amount of compensation for that seizure. Kerner and Pelc (2019) report that these ISDS cases are fairly common[34]. In cases where ISDS is used as a pricing mechanism instead of as a justice system, a state is entirely willing to recognize and compensate the investor for the expropriation; there is simply a disagreement between the state and investor as to what is the fair rate of compensation.

In regard to the fear of future takings, outright expropriations usually have specific justifications. As a result, these specific public purposes narrow the number of investors who fear that such a policy would be used against their assets in the future. States usually provide detailed evidence to justify how a particular seizure serves that specific public interest, as in Unglaube v. Costa Rica (2008) when Costa Rica seized the beachfront property of a German national to protect threatened local sea turtle populations[35]. While a direct expropriation case concerns foreign investors whose holdings fit the specific set of circumstances in the case, it presents little risk to assets that do not match such a narrow description. Few investors would fear losing their assets from the threatened seizure of a narrow range of assets.

3.3 *Causal Logic: Indirect Expropriation Rulings' Impact on a State's FDI Inflow*

On the other hand, indirect expropriation cases involve regulatory behavior more frequently encountered by a wide range of industries. The voluminous log of ISDS cases[36] reveals which areas of state policy are liable to be ruled as indirect expropriation. Saar Papier v. Poland (1994) involved

34 Kerner and Pelc M.S., 2019.
35 Kerner and Pelc M.S., 2019.
36 UNCTAD Investment Dispute Settlement Navigator publishes brief summaries of most disputes and links relevant news articles for further investigation.

trade policy with export controls; Bear Creek Mining v. Peru (2014) concerned environmental and safety regulation; Stans Energy v. Kyrgyzstan (I) (2013) challenged licensing policy; and Yukos Universal v. Russia (2005) fought the country's tax code. In all of these cases, the overarching state policy threatened a large swath of business operations. In losing an indirect expropriation case, a state damages its credibility to exercise this regulatory power within its BIT obligations. As follows, the signal of regulatory risk applies to the broad range of foreign investors whose assets are potentially threatened by that policy. It is logical that, with more investors reacting, states consequently lose more FDI inflow.

3.4 Why Wait Until a Ruling?

Why would investors react to indirect expropriation cases only after a case's ruling, rather than upon its initiation? I attribute this disparity to the lack of a formal definition of indirect expropriation. Since most BITs do not clearly distinguish legitimate regulatory action and indirect expropriation, the facts of these cases can be complex and are left to the interpretation of arbitral bodies on a case-by-case basis. Investors simply do not have enough information to react before the ruling of indirect cases. Until they are decided, investors have little reason to believe that most indirect expropriation claims are valid, since indirect expropriation ISDS claims have a success rate of only 13%[37]. An ISDS ruling of indirect expropriation against a state would be significant: because states rarely lose these cases, losses are therefore worthy of attention when they happen. In contrast, the higher success rates and more detailed legal definitions of direct expropriation send a clearer signal to investors of a state's conduct upon the initiation of an ISDS case. This means that the amount of information revealed by an indirect expropriation ruling against the state is much more than that of a direct expropriation ruling. With additional information to act upon, investor reactions will be more robust.

37 Pelc 2015.

3.5 Hypothesis

I hypothesize that governments will experience a greater reduction of FDI flows following the loss of an investor-state dispute settlement regarding indirect expropriation than would be experienced after the loss of an investor-state dispute settlement regarding direct expropriation, ceteris paribus.

3.6 Competing Explanation

An obvious counterargument is that the particular facts of an ISDS case determine a state's loss of FDI inflow rather than the type of case at hand. It seems reasonable that the specifics of a case, by demonstrating how drastically a state violated its BIT promises, would modulate the magnitude of FDI deterred from any foreign investor. This explanation may nuance investors' reactions, but it wields less general explanatory value. The different types of state behavior involved in direct and indirect expropriation produce disproportionate scopes of foreign investors acting upon any such information. No state's loss of a direct expropriation case, no matter how flagrant the seizure, could deter as wide of an audience of foreign investors as could the loss of an indirect expropriation case. Regardless of how clear, convincing, and egregious a direct expropriation case is, its salience among the wide population of foreign investors is limited by the state's purpose for seizing any such property[38]. Meanwhile, regulatory behavior involved in indirect expropriation concerns a wide scope of investors. Foreign investors may react differently to case specifics, but the audience responding to direct expropriation cases will still be dramatically narrower compared to the audience reacting to indirect expropriation cases.

38 In defence of proposals to nationalize American banks amid the financial crisis, HuffPost published an article titled, "Nationalization: It's Not Scary, It's All Around You." This article included a quote from economics Nobel Laureate Paul Krugman that nationalization is "as American as apple pie." See Sirota 2011.

4 Data and Methodologies

4.1 *The Model*

To test my hypothesis against empirical evidence, I use a linear fixed-effect model with 658 observations to examine the respective FDI impacts of losing an indirect expropriation ISDS case and of losing a direct expropriation ISDS case. I control for other variables that may affect FDI flows and make each country a panel of the regression. I use UNCTAD's comprehensive database of ISDS cases to see how many indirect or direct expropriation cases a country loses in any given year, and regress these losses and control variables against counties' FDI flows. My dataset spans from 1990 to 2018, and it does so for several reasons. Firstly, many of the countries that have lost these cases are of the former Soviet Union or its communist satellite states, so it makes sense to begin the series when many of these countries began accepting foreign investment. Secondly, 99.902% of ISDS cases recorded in UNCTAD's database occur after 1990. Finally, indirect expropriation cases only began arising in the 1990s, and we need both cases to make any meaningful comparisons.

4.2 *Independent Variable*

As previously stated, the independent variable of my research is whether a state loses an indirect expropriation or direct expropriation ISDS case. In my statistical analysis, I consider the number of these respective cases that a country loses in any given year with the variables INDIRECT EXPROPRIATION ISDS LOSS and DIRECT EXPROPRIATION ISDS LOSS. I limit these variables to cases that allege either only indirect expropriation or direct expropriation to exclude cases with several types of claims and rulings. I do this to assure that my analysis exclusively compares the effects of losing indirect versus direct expropriation and does not introduce confounding variables of how investors react to claims and rulings upon additional alleged breaches such as umbrella clause, performance requirements, or full protection and security. Cases must also be dated according to when arbitral bodies render a decision that shows a clear verdict in the case with regards to whether or not a state violated its BIT obligations. As such, I date cases by their final award and verdict, unless an interim award has already been rendered, in which case I use the date of the interim award.

Following these parameters, our model includes 60 cases lost by 25 countries between 1990 and 2018.

4.3 Dependent Variable

My dependent variable is LAGGED PERCENT CHANGE FDI NET INFLOW, measured annually at the country level. Following previous studies, I use an inverse hyperbolic sine transformation on nominal FDI net inflows to avoid skew in our measure of FDI. I then follow the transformation given by Aisbett, et al. in their study of ISDS losses on states' FDI, where the "transformation of a variable y is

$$\Theta^{-1}\sinh^{-1}(\Theta y) = \Theta^{-1}\text{-}\ \ln(\Theta y + (\Theta^{2}\ y^{2} + 1)^{1/2}),$$

where I set $\Theta = 1$[39]. This transformation of FDI inflows has several advantages over the logarithmic transformation used in the studies of Allee and Peinhardt (2011)[40] and Kerner and Pelc (2019)[41], not the least of which is that it allows me to examine zero and negative values of FDI inflow. This is particularly important when considering some of the FDI inflows in our dataset, such as when Hungary saw a net FDI inflow of -14,797.345 million US dollars in 2014[42]. I then find the percentage change of this variable from year to year in

39 Aisbett et al. (2017) offer thorough elaboration on the inverse hyperbolic sine transformation, including its origin and use, on pages 129 and 130.
40 Allee and Peinhardt 2011.
41 Kerner and Pelc M.S., 2019.
42 Analysis from Santander Bank suggests that large public debt was a considerable deterrent to Hungarian FDI. See "Hungary: Foreign Investment." Foreign investment in Hungary - Santandertrade.com.

order to yield the variable used in this analysis.

For LAGGED PERCENT CHANGE FDI NET INFLOW, we see a maximum value of 11.9, a minimum value of -10.3, and an average value of 7.1.

4.4 *Control Variables*

To reinforce my statistical regression's ability to accurately capture the FDI impact of losing these different types of ISDS cases, I include five control variables that could also influence the net FDI inflow of a country. By controlling for these other variables, I preempt objections that the variation examined results from variation in these other variables. Moreover, these additional variables create a more robust test of statistical significance against which to set my hypothesis.

The first control variable included is YEARS SQUARED to control for nonlinearity in time by providing a proxy for inflationary pressures. Currencies usually experience inflation, but the variation in currencies used by foreign investors would make a more specific variable challenging to compute, and I square the value so that the differences are not unitary. The second control variable is BILATERAL INVESTMENT TREATIES, which is the number of bilateral investment treaties that a country has in place during a given year according to the UNCTAD Investment Policy Hub. Previous work has shown that more bilateral investment treaties tend to attract more foreign investment[43]. The third control variable is GDP GROWTH as a measure of a country's annual economic growth, taken from the World Bank's World Development Indicators, since foreign investors are more likely to invest in fast-growing economies[44]. Our fourth control variable, POLITICAL RIGHTS, is taken from Freedom House's Political Rights ratings and serves as a general control for democracy on FDI[45]. It is important to note that this index ranges from one (weak rights) to seven (strong rights). Scholars disagree whether higher levels of authoritarianism or liberal democracy promote

43 See Bhasin Manocha 2016, Frenkel and Walter 2019, and Lee and Johnston 2016.
44 "World Development Indicators." DataBank. World Bank.
45 "Economic Data." TheGlobalEconomy.com. The Global Economy.

FDI[46], but, because it is prevalent in analyses of FDI, I nevertheless control for it. Finally, I include a fifth and final control variable TOTAL WORLD FDI, as published by UNCTAD, to capture transitory changes in aggregate foreign investment in any given year as well as the increase in global investment.

Table 1:

	(1)
	F.LAGGED CHANGE IN FDI NET INFLOW
Primary Variables	
INDIRECT EXPROPRIATION ISDS LOSS	-3.339*
	(-2.42)
DIRECT EXPROPRIATION ISDS LOSS	-0.285
	(-0.16)
Control Variables	
YEARS SQUARED	0.0000741**
	(2.98)
BILATERAL INVESTMENT TREATIES	0.0465**
	(2.65)
GDP GROWTH	0.338***
	(4.66)
POLITICAL RIGHTS	0.302
	(1.29)
TOTAL WORLD FDI (*in millions*)	-0.188*
	(-2.10)
CONSTANT	-282.7**
	(-2.86)
N	658

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.5 *Empirical Results*

The results of my analysis are shown in Table 1. The loss of

46 The literature is intensely divided on this topic. Truman and Emmert (2004) and Tomashevisky (2017) link authoritarian regimes to greater FDI inflow on account of favorable conditions for investors and political incentives. Meanwhile, Anyawu and Yaméogo (2015) and Li and Resnick (2003) hypothesize that liberal democracy attracts more FDI by upholding a rule of law.

an indirect expropriation case is negatively correlated and significant (p = 0.02), revealing that indirect expropriation rulings consistently deter considerable foreign investment to a state. Meanwhile, the loss of direct expropriation is insignificant (p = 0.887) and, moreover, negatively correlated at only a fraction of that of a loss of indirect expropriation. This indicates that states do not consistently lose FDI after losing a direct expropriation case and, when they do, the FDI loss tends to be much less than what follows the loss of an indirect expropriation case. This difference becomes especially apparent when graphing 95% confidence intervals for the marginal effects of these ISDS losses, as is done in Table 2.

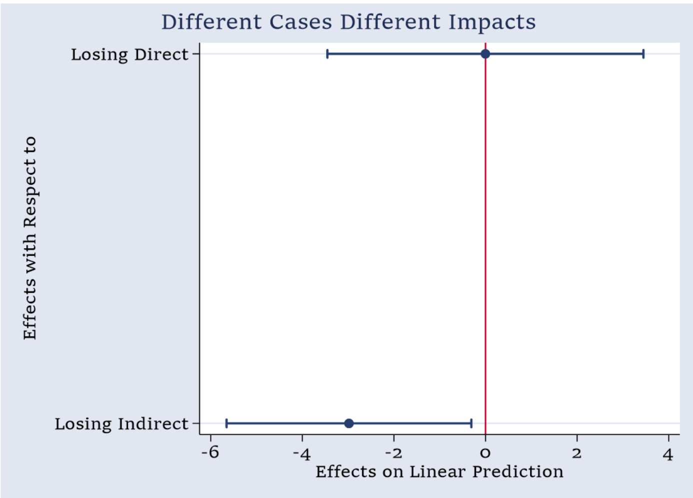


TABLE 2: 95% Confidence Intervals for marginal effects of indirect and direct expropriation ISDS rulings on states' FDI inflow.

The model also finds all but one included control variable statistically significant. YEARS SQUARED, the number of bilateral investment treaties a country signs, and a country's economic growth are all found to be significant and positively correlated with a state's FDI inflow. Meanwhile, although significant, total world FDI shows a negative correlation. Finally, my model finds that a country's measure of political rights is positively correlated with a state's lagged net FDI inflow but statistically insignificant. Again, it is important to note that my measure of political rights judged stronger political rights as lesser values, so this analysis therefore says that a country's measure of democracy is negatively correlated to its FDI inflow, albeit insignificant[47].

4.6 Discussion of Empirical Results

As hypothesized, the model shows a clear difference of effect to states' FDI inflows between losing indirect expropriation ISDS cases and losing direct expropriation ISDS cases. My findings reveal empirical support for my hypothesis: states tend to lose more FDI inflow following losses of indirect expropriation ISDS cases than after losses of direct expropriation ISDS cases. The model shows a negative relationship between a state's net FDI inflow and losses of both an indirect expropriation and direct expropriation case, but the relationship is statistically significant only for the losses of indirect expropriation cases. Moreover, the model directly predicts that indirect expropriation rulings harm a state's FDI inflow much more than direct expropriation rulings do.

This difference in significance levels falls in line with the legal clarity component of my logic. BITs' more explicit definitions of direct expropriation afford a greater degree of clarity for foreign investors to judge whether or not a state violated its BIT obligations even before an ISDS ruling. This follows the analysis of Kerner and Pelc (2019) which found that states lose FDI following the initiation of direct expropriation cases. States do not consistently see FDI loss

47 This finding may be best explained by Bak and Moon (2016), who assert that authoritarian rulers encourage FDI as it helps insulate their political stability. While it is beyond the scope of this paper, I encourage future inquiry into whether foreign investors consider different types of cases differently for different regime types.

after losing direct expropriation cases, as those losses have largely already occurred by the time a ruling is released. This contrasts indirect expropriation cases for which I argue that foreign investors do not have enough information to react until the definitive decision of an arbitral body against the state. In conclusion, states consistently lose FDI following the loss of indirect expropriation cases because this is when investors have the information necessary to react. Conversely, these FDI losses already proceed the rulings upon direct expropriation.

Another noticeable finding is that the negative coefficient attached to the loss of an indirect expropriation case is more than ten times that of direct expropriation cases. This difference is especially pronounced in the visualization of Table 2. This finding supports my argument that the regulatory nature of indirect expropriation cases lends to a broader scope of foreign investors reacting than do specific seizures of direct expropriation cases. Logically, the more foreign investors who react, the greater the loss of FDI inflow for a state, which my model confirms.

4.7 Descriptive Examples

To illustrate this analysis with qualitative examples, I compare two ISDS cases: ADC v. Hungary (2003), which alleged the direct expropriation of an airport terminal, and Yukos Universal v. Russia (2005), which alleged indirect expropriation of a British oil company through tax law.

Precipitating ADC v. Hungary (2003)[48], the Cypriot company ADC Affiliate Limited had entered into a contract with Hungary to build and operate two airport terminals at Budapest-Ferihegy International Airport. However, in December 2001, the Hungarian Minister of Transport ordered all flight-related operations of the capital's airport to be taken under state control—including those of ADC. Investors hardly needed an arbitral ruling to identify Hungary's action as a direct expropriation, considering the

48 ADC Affiliate Limited and ADC & ADMC Management Limited v. Republic of Hungary, ICSID Case No. ARB/03/16.

government quite openly seized ownership of the airport terminals. Nonetheless, this policy was of limited concern to the majority of foreign investors, most of whom were unaffected by the seizure. In fact, Hungary's FDI inflow continued to increase both after the case's initiation and ruling against the state. In light of the September 11 attacks that year, the Hungarian government justified the takeover with specific security concerns[49], especially since the airport was near the capital city of Budapest. That justification narrowly limited the foreign investors concerned by the case as most assets do not fall into the specific set of circumstances that justified the seizure in question, and, in turn, the case's ruling was of little consequence for Hungary's FDI.

This contrasts with Yukos Universal v. Russia (2005) which was ruled as indirect expropriation in 2014 by UNCITRAL[50]. Prior to this case, Russia's annual FDI had nearly doubled between 2009 and 2013[51]. UNCITRAL's Investment Policy Hub describes the case as follows:

> "Claims arising out of a series of actions undertaken by the respondent against Yukos Oil Company, including arrests, large tax assessments and liens, and the auction of the main Yukos facilities, among others, which allegedly led to the bankruptcy of the company and eliminated all value of claimant's shares in Yukos.[52]"

Following this case's initiation in 2005, Russian FDI inflow actually increased into 2006. However, the opposite occurred after the case's decision. Arbitrators ruled that

49 In fact, the December 2011 Report of the Council of Europe's Committee on Economic Affairs and Development gave a green light to such policies, remarking "The rather draconian measures advocated in this report are likely to leave the reader with an eerie sentiment that we are approaching a 'Big Brother' society. We certainly are at airports. However, it must be remembered that this is a state of affairs not of our choosing." See Council of the European Union 2001.
50 Yukos Universal Unlimited (Isle of Man) v. The Russian Federation, UNCITRAL PCA Case No. AA 227.
51 "FDI Statistics." UNCTAD.
52 Investment Dispute Settlement Navigator | UNCTAD Investment Policy Hub.

the Russian Federation unfairly issued audits and fines that forced the Yukos Oil Company into bankruptcy[53]. Perhaps more than any other policy area, tax policy—from payment to compliance—affects just about any legitimate economic activity[54]. When ISDS verified that Russia abused a state policy that could impact any foreign investment, that signal of risk applied to the broad audience of investors who could likewise be subject to the same abuses of the tax code. Following this case's decision, Russia's 2015 FDI inflow fell by 59% compared to 2014. [55,56]

4.8 *Robustness Check*

Do my results bear out even under a different methodology? To find out, I ran my variables through a Tobit panel regression with a lower bound at 0. I did this because negative values of net FDI inflow are not common, but, when they do occur, they are rather large[57]. By censoring these values, I ensure that they do not distort the regression and that I am examining the effects under less-exceptional circumstances.

Table 3 exhibits the results of this test. Rather than cast doubt upon my argument, the robustness check actually reveals convincing evidence for my hypothesis. Losses

53 Yukos Universal Unlimited (Isle of Man) v. The Russian Federation, UNCITRAL PCA Case No. AA 227. Further complicating matters in this case, assets of Yukos Oil Company were quickly auctioned off to Gazprom and Rosneft, two state-owned Russian oil companies.

54 See Artemenko, Aguezarova, and Porollo 2017.
55 "FDI Statistics." UNCTAD.
56 It is difficult to completely isolate the cause of this drop in FDI, as Russia's annexation of Crimea occurred in early 2014. However, I believe this case study still holds illustrative value as the invasion preceded the Yukos Universal v. Russia and, therefore, we can expect foreign investors to have reacted to the invasion before they reacted to the case. This seems plausible, as Russia's FDI inflow had fallen by 45% by the end of 2014 and fell another 59% by the end of 2015.
57 It is likely that such large drops of FDI are due to exceptional circumstances that may be difficult to account for in a model. I therefore drop these so we can examine the relationship between ISDS rulings and states' FDI under more ordinary conditions.

of indirect and direct expropriation cases retain their negative relationships to net FDI inflow, but, whereas the

	(1)
	F. LAGGED CHANGE IN FDI NET INFLOW
F. LAGGED CHANGE IN FDI NET INFLOW	
Primary Variables	
INDIRECT EXPROPRIATION ISDS LOSS	-4.446*
	(-2.30)
DIRECT EXPROPRIATION ISDS LOSS	-0.0117
	(-0.01)
Control Variables	
YEARS SQUARED	0.0000849**
	(2.79)
BILATERAL INVESTMENT TREATIES	0.0498*
	(-2.45)
GDP GROWTH	0.349***
	(-4.75)
POLITICAL RIGHTS	0.253
	(1.05)
TOTAL WORLD FDI (*in millions*)	-0.204
	(-1.77)
CONSTANT	-325.1**
	(-2.69)
/	
sigma u	1.080
	(1.54)
sigma e	9.282***
	(22.09)
N	658

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

loss of a direct expropriation case becomes less significant, the loss of an indirect expropriation case becomes even more significant. Moreover, the loss of FDI associated with losing an indirect expropriation case becomes even more pronounced compared to in the linear-fixed effect model, while drawing closer to zero for the loss of direct expropriation.

TABLE 3: Tobit panel regression with lower bound at 0. The effects of different ISDS losses on states' net FDI inflows.

5 Limitations and Implications

I began this research with the simple question of why certain cases cause states to lose more FDI than others. I argued that this inconsistency arises from the types of cases. Specifically, that the loss of an indirect expropriation causes a state to lose more FDI than the loss of a direct expropriation case. Although statistical analysis and real-world examples demonstrate strong evidence for my hypothesis, my argument is limited because it only considers cases that exclusively claim indirect or direct expropriation. Unfortunately, my analysis does not explore these dynamics under different types of complaints, such as violations of umbrella clauses or minimum standard of treatment.

Such a level of analysis may be tricky. The majority of ISDS cases allege multiple breaches against a state, so it can be difficult to isolate the effect of any one alleged breach among multiple. However, I believe that my findings have some level of generality, since the inclusion of either an indirect expropriation or direct expropriation complaint in any given case should influence its FDI cost in a similar manner to my argument.

Another limitation present in my research may be the types of investments a state can lose. Even if a state loses on indirect expropriation, it is not always possible for foreign investors to walk away, despite the risk conveyed. If their assets involve substantial sunk costs, they may have more to lose by withdrawing their investments[58]. As such, states' FDI loss may also be related to the liquidity of their industries. Although this may keep foreign investors from pulling out of a country, however, it is also entirely plausible that an indirect expropriation ruling would more strongly deter new investors from committing those sunk costs in the future[59]. While future research should examine the costs of being brought before ISDS arbitration at a higher level of granularity, my findings introduce several important implications for policymakers.

First, the cost of an indirect expropriation ISDS case does

58 Crasnic, Kalyanpur, and Newman 2016.
59 Zeelenberg and Van Dijk 1997.

not only come down to lawyers and legwork. If a state loses a case, it can cost them their reputations with many foreign investors. However, it is unclear whether policymakers realize this. Studies such as Van Harten and Scott (2015) suggest that policymakers are more focused on the legal costs of arbitration than how investors will interpret a loss[60]. Also, past studies—which do not consider cases' rulings—such as Kerner and Pelc (2019), may mislead policymakers by suggesting that indirect expropriation cases do not harm states' FDI. However, if indirect expropriation cases are indeed ruled in favor of the claimant, my analysis shows that the cost is real and considerable. By following the existing literature and disregarding the costs of indirect expropriation claims, states may find themselves facing adverse rulings that damage their reputation with foreign investors.

Second, if states consider the FDI cost of losing indirect expropriation ISDS cases, it adds another layer to the existing literature regarding regulatory chill. As explained earlier, past literature suggests that policymakers are not always sure whether new regulation may be considered indirect expropriation. Policymakers must weigh the regulation's benefit against any costs of it attracting ISDS claims[61]. Existing literature has largely ignored FDI in this cost calculation. However, since these regulatory measures are often the target of allegations for indirect expropriation allegations, my argument presents an additional FDI cost for states to evaluate when considering regulatory behavior. Factoring in this additional and considerable potential cost of FDI, policymakers may find that the risk of new regulation outweighs its benefits and accordingly fail to implement it. While my first implication is that states who disregard this cost may mistakenly damage their credibility with foreign investors, my second implication is that considering these costs intensifies ISDS regulatory chill[62].

60 See Van Harten and Scott 2016 and Tienhaara 2017.
61 Bonnitcha 2016. See the chapter on regulatory chill and its causes on pages 113 to 132, which does an excellent job of describing these considerations.
62 Concern over regulatory chill is not limited to academic circles. While it was still under consideration, the Transatlantic Trade and Investment Partnership

However, states' resources are also limited, leading authors to point out where states simply give up in ISDS arbitration unable to muster the legal legwork. Nonetheless, my research shows that waving the white flag and accepting the loss of an indirect expropriation case has its own considerable FDI cost. While legal expenses can be substantial in the moment, it can take years for a state to rebuild its credibility with investors[63]. Therefore, if policymakers want to contain the ultimate damage done by an ISDS case, they may consider dedicating even more resources to fighting indirect expropriation allegations. If an indirect expropriation loss is indeed inevitable, states should at least play damage control and mitigate foreign investors' fears of regulatory risk by reducing other regulatory barriers or justifying the regulations in question.

Third, the lesser FDI loss for direct expropriation suggests an opportunity for states to exercise more unilateral discretion towards ISDS rulings. Because the range of concerned investors is small, the losses of FDI may altogether be less than the sizable damages an arbitral body orders a state to pay. Hence, while complying with an ISDS ruling may help ease the concern amongst these investors, if the awarded damages are greater than the FDI that would be deterred from this narrow scope investors, a state can plausibly choose the lesser of two evils and refuse to pay damages. This unilateral capacity already has precedent in the Rosatti Doctrine of Argentina, which states that international rulings (such as ISDS) are subject to the review of national courts[64]. Continuing with the Argentine example, since states face a more narrow audience of investors in direct expropriation rulings, national courts could overrule direct expropriation cases—and hence

(TTIP) faced open public protest in Germany, in part motivated by concerns over the fear that ISDS provisions would curtail regulatory sovereignty. See The Guardian 2016. Likewise, United States Senator Elizabeth Warren chided the agreement's inclusion of ISDS, writing "it would undermine U.S. sovereignty." See Warren 2016.

63 Santiso 2004.

64 Gómez 2011.

avoid paying ruling's damages—with considerably less consequence than in indirect expropriation cases.

6 Conclusion

Far from being written off, indirect expropriation ISDS cases have considerable consequences for policymakers to consider. I argued that indirect expropriation losses concern more foreign investors and reveal more actionable information than do direct expropriation losses, provoking more FDI loss for the respondent state. My hypothesis is confirmed by empirical analysis which confirms that losses of indirect expropriation cases are notably more daunting to foreign investment than losses of direct expropriation cases. As governments endeavor to strike a balance between attracting foreign investors and crafting effective public policy, they must be mindful of the costs associated with breaking their BIT obligations. There is a particular risk present in indirect expropriation cases, and it deserves to be directly stated.

7 Acknowledgements

This paper would not have been possible without the insight, guidance, and encouragement of the Spring 2020 Krogh Honors Research Seminar at the Georgetown University School of Foreign Service. I thank Advait Arun, Khai Nyugen, Channing Lee, and Charlie "Chazz" Wang for their efforts in editing and strengthening this paper. I would especially like to thank Mariana Aslund for her excellent recommendations throughout the writing process. I extend my sincerest gratitude to Professor Marc L. Busch, whose relentless energy, innovative teaching, and candid criticism created a formative educational experience. Finally, I thank my mother, father, and brother for their exceptional patience as I pitched arguments.

8 Works Cited

ADC Affiliate Limited and ADC & ADMC Management Limited v. Republic of Hungary, ICSID Case No. ARB/03/16 (October 2, 2006). http://

icsidfiles.worldbank.org/icsid/ICSIDBLOBS/OnlineAwards/C231/DC648_En.pdf

Aisbett, Emma, Matthias Busse, and Peter Nunnenkamp. “Bilateral Investment Treaties as Deterrents of Host-Country Discretion: the Impact of Investor-State Disputes on Foreign Direct Investment in Developing Countries.” *Review of World Economics* 154, no. 1 (2017): 119–55. https://doi.org/10.1007/s10290-017-0285-1.

Allee, Todd, and Clint Peinhardt. “Contingent Credibility: The Impact of Investment Treaty Violations on Foreign Direct Investment.” *International Organization* 65, no. 3 (2011): 401–32. https://doi.org/10.1017/s0020818311000099.

Anyanwu, John C., and Nadège D. Yaméogo. “Regional Comparison of Foreign Direct Investment to Africa: Empirical Analysis.” *African Development Review* 27, no. 4 (2015): 345–63. https://doi.org/10.1111/1467-8268.12152.

Artemenko, D. A., Aguzarova, L. A., Aguzarova, F. S., & Porollo, E. V. (2017). Causes of tax risks and ways to reduce them. E*uropean Research Studies*, 20(3), 453-459. Retrieved from http://proxy.library.georgetown.edu/login?url=https://search.proquest.com/docview/1944539192?accountid=11091

Bak, Daehee, and Chungshik Moon. “Foreign Direct Investment and Authoritarian Stability.” *Comparative Political Studies* 49, no. 14 (November 2016): 1998–2037. https://doi.org/10.1177/0010414016655536.

Bhasin, Niti, and Rinku Manocha. “Do Bilateral Investment Treaties Promote FDI Inflows? Evidence from India.” *Vikalpa: The Journal for Decision Makers* 41, no. 4 (2016): 275–87. https://doi.org/10.1177/0256090916666681.

Bonnitcha, Jonathan. Substantive *Protection under Investment Treaties: A Legal and Economic Analysis*. Place of publication not identified: Cambridge Univ Press, 2016.

Bonnitcha, Jonathan, Lauge N. Skovgaard. Poulsen, and Michael Waibel. *The Political Economy of the Investment Treaty Regime*. New York, NY: Oxford University Press, 2017.

Council of the European Union. *Air transport and terrorism: how to enhance security?,* Doc. 9296. (Report of the Committee on Economic Affairs and Development, Brussels, December 2001). https://assembly.coe.int/nw/xml/XRef/X2H-Xref-ViewHTML.asp?FileID=9574&lang=EN

Crasnic, Loriana, Nikhil Kalyanpur, and Abraham Newman. “Networked Liabilities: Transnational Authority in a World of Transnational Business.” *European Journal of International Relations* 23, no. 4 (September 2016): 906–29. https://doi.org/10.1177/1354066116679245.

Dolzer, Rudolf, and Margrete Stevens. *International Centre for Settlement of Investment Disputes: Bilaleral Investment Treaties*. The Hague: M.

Nijhoff, 1995.

"Economic Data." TheGlobalEconomy.com. The Global Economy. Accessed March 15, 2020. https://www.theglobaleconomy.com/download-data.php.

"FDI Statistics." UNCTAD. Accessed March 13, 2020. https://unctad.org/en/Pages/DIAE/FDI Statistics/FDI-Statistics.aspx.

Frenkel, M., & Walter, B. (2019). Do bilateral investment treaties attract foreign direct investment? the role of international dispute settlement provisions. *The World Economy*, 42(5), 1316-1342. doi:http://dx.doi.org/10.1111/twec.12743

Friedman, Mark W, Dietmar W Prager, and Ina C Popova. "Expropriation and Nationalisation." Global Arbitration Review. Debevoise & Plimpton, 2019. Accessed March 3, 2020. https://globalarbitrationreview.com/chapter/1142558/expropriation-and-nationalisation.

Gómez, Katia Fach. "Latin America and ICSID: David versus Goliath?" *Law and Business Review of the Americas* 17, no. 2 (March 2011): 195–230. https://doi.org/10.2139/ssrn.1708325.

Heil, Oliver, and Thomas S. Robertson. "Toward a Theory of Competitive Market Signaling: A Research Agenda." *Strategic Management Journal* 12, no. 6 (1991): 403–18. https://doi.org/10.1002/smj.4250120602.

"Hungary: Foreign Investment." Foreign investment in Hungary - Santandertrade.com. Accessed May 8, 2020. https://santandertrade.com/en/portal/establish-overseas/hungary/foreign-investment.

""Indirect Expropriation" and the "Right to Regulate" in International Investment Law", OECD Working Papers on International Investment, 2004/04, OECD Publishing. http://dx.doi.org/10.1787/780155872321

International Investment Agreements Navigator | UNCTAD Investment Policy Hub. UNCTAD. Accessed January 8, 2020. https://investmentpolicy.unctad.org/international-investment-agreements/model-agreements.

Investment Dispute Settlement Navigator | UNCTAD Investment Policy Hub. UNCTAD. Accessed January 8, 2020. https://investmentpolicy.unctad.org/investment-dispute-settlement.

Kelland, Kate. "Gates and Bloomberg Create $4 Million Fund to Fight Big Tobacco." Reuters. Thomson Reuters, March 18, 2015. https://www.reuters.com/article/us-health-tobacco-fund/gates-and-bloomberg-create-4-million-fund-to-fight-big-tobacco-idUSKBN0ME24C20150318.

Kerner, Andrew, and Krzysztof J. Pelc, "Do Investor-State Disputes Harm FDI?" M.S., Michigan State University, 2019. https://www.semanticscholar.org/paper/Do-Investor-State-Disputes-Harm-FDI-Kerner-Pelc/2a9c0a2563d90f14ddbc210f2aafeec9609a9044

Lee, C., & Johnston, N. P. (2016). Improving reputation BIT by BIT: Bilateral investment treaties and foreign accountability. International

Interactions, 42(3), 429-451. doi:http://dx.doi.org/10.1080/03050629.2016.1128429

Li, Quan, and Adam Resnick. "Reversal of Fortunes: Democratic Institutions and Foreign Direct Investment Inflows to Developing Countries." *International Organization* 57, no. 1 (2003): 175–211. https://doi.org/10.1017/s0020818303571077.

Mitchell, Cristina. "PAHO/WHO: PAHO/WHO Supports Uruguay in Defending Tobacco Control Policies against Tobacco Industry Challenge." Pan American Health Organization / World Health Organization. Pan American Health Organization, November 5, 2014. https://www.paho.org/hq/index.php?option=com_content&view=article&id=10179:2014-uruguay-defending-tobacco-industry-challenge&Itemid=1926&lang=en.

Pelc, Krzysztof J. "What Explains the Low Success Rate of Investor-State Disputes?" *International Organization* 71 (June 2017): 559–83. https://doi.org/10.1017/S0020818317000212.

Picht, Hartmut, and Volker Stüven. "Expropriation of Foreign Direct Investments: Empirical Evidence and Implications for the Debt Crisis." *Public Choice* 69, no. 1 (1991): 19–38. https://doi.org/10.1007/bf00123852.

Santiso, C. (2004). RE-FORMING THE STATE: GOVERNANCE INSTITUTIONS AND THE CREDIBILITY OF ECONOMIC POLICY: [1]. *International Public Management Journal*, 7(2), 271-298. Retrieved from http://proxy.library.georgetown.edu/login?url=https://search.proquest.com/docview/218852321?accountid=11091

Sirota, David. "Nationalization: It's Not Scary, It's All Around You." HuffPost. HuffPost, May 25, 2011. https://www.huffpost.com/entry/nationalization-its-not-s_b_170150.

Stevens, Charles E., and Joseph T. Cooper. "A Behavioral Theory of Governments' Ability to Make Credible Commitments to Firms: The Case of the East Asian Paradox." *Asia Pacific Journal of Management* 27, no. 4 (February 2009): 587–610. https://doi.org/10.1007/s10490-009-9166-z.

Tienhaara, Kyla. "Regulatory Chill in a Warming World: The Threat to Climate Policy Posed by Investor-State Dispute Settlement." *Transnational Environmental Law* 7, no. 2 (2017): 229–50. https://doi.org/10.1017/s2047102517000309.

Tobin, Jennifer L. "The Social Cost of International Investment Agreements: The Case of Cigarette Packaging." *Ethics & International Affairs* 32, no. 2 (June 2018): 153–67. https://doi.org/10.1017/s089267941800028x.

Tomashevskiy, Andrey. "Investing in Violence: Foreign Direct Investment and Coups in Authoritarian Regimes." *The Journal of Politics* 79, no. 2 (2017): 409–23. https://doi.org/10.1086/688356.

"TTIP Protesters Take to Streets across Germany." The Guardian. Guardian News and Media, September 17, 2016. https://www.theguardian.

com/business/2016/sep/17/ttip-protests-see-crowds-take-to-streets-of-seven-german-cities.

Tuman, John Peter, and Craig F. Emmert. "The Political Economy of U.S. Foreign Direct Investment in Latin America: A Reappraisal." *Latin American Research Review* 39, no. 3 (2004): 9–28. https://doi.org/10.1353/lar.2004.0060.

Van der Pas, Hilde, and Riza Damanik. "The case of Newmont Mining vs Indonesia." (Policy briefing, Indonesia for Global Justice, Transnational Institute, EU-ASEAN FTA Network, Brussels, November 2014).

Van Harten, Gus Nadine, and Dayna Nadine Scott. "Investment Treaties and the Internal Vetting of Regulatory Proposals: A Case Study from Canada (Research Paper 26)." Toronto: Osgoode Hall Law School, April 19, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2700238.

Warren, Elizabeth. "The Trans-Pacific Partnership Clause Everyone Should Oppose." The Washington Post. WP Company, February 25, 2015. https://www.washingtonpost.com/opinions/kill-the-dispute-settlement-language-in-the-trans-pacific-partnership/2015/02/25/ec7705a2-bd1e-11e4-b274-e5209a3bc9a9_story.html.

"World Development Indicators." DataBank. World Bank. Accessed March 3, 2020. https://databank.worldbank.org/source/world-development-indicators#.

Yukos Universal Unlimited (Isle of Man) v. The Russian Federation, UNCITRAL PCA Case No. AA 227, Final Award (July 18, 2014). https://www.italaw.com/sites/default/files/case-documents/italaw3279.pdf

Zeelenberg, Marcel, and Eric Van Dijk. "A Reverse Sunk Cost Effect in Risky Decision Making: Sometimes We Have Too Much Invested to Gamble." *Journal of Economic Psychology* 18, no. 6 (1997): 677–91. https://doi.org/10.1016/s0167-4870(97)00029-9.

Combatting Coronavirus Effectively: Determinants of Health Status Outcomes and Inequalities in the European Union

Annamara Rimmer

University of Cambridge

Abstract

This paper addresses the determinants of health status outcomes and inequalities within the European Union. In the aftermath of the Coronavirus pandemic, a population's health status has become of even greater importance to policymakers, and the necessity of further study into the determinants of health status outcomes and inequalities has thus become evident. An aggregate health production function is considered here for the 28 European Union member countries and for a sample of 12 non-European nations for comparative purposes. Various methodologies are used to mitigate potential issues of heterogeneity and endogeneity, including panel data, instrumental variables, and non-parametric regression. Whilst acknowledging the important limitations with the available data, particularly when drawing conclusions in the aftermath of the Covid-19 pandemic, our results find that risk factors, access to care, and healthcare resources are all statistically significant factors in determining health status outcomes in the European Union. These findings are pleasingly robust to different proxy measures for health status being used which allow for greater focus on more qualitative aspects of health status.

Acknowledgements

Many thanks are due to the York Health Economics Consortium, particularly Dr Nick Hex and Dr Matthew Taylor for their support and advice during this project. I am also very grateful for the support I received throughout this year from Dr Juan Block.

Table of Contents

1 Introduction

The health status of a population represents a complex and multifaceted blanket term that can be measured by a wide range of factors, including life expectancy, mortality rates, morbidity from specific diseases, and quality of life (National Center for Health Statistics, 2017).

Consequently, we define health status outcomes as a measure of the change in the health status of an individual or group that can be attributed to an intervention, noting that health status in and of itself is a multidimensional concept, and thus, a difficult issue to pin down. Even prior to the current Coronavirus pandemic, health status outcomes have become increasingly important as one of the major concerns amongst policymakers in developed countries in recent years. This is reflected by the fact that expenditure on health care accounted for 9.6% of GDP in the EU as a whole in 2017, up from 8.8% in 2007 (OECD, 2019). With the continued spread of the Coronavirus pandemic, this trend will likely only continue and even accelerate.

As a policy issue, health status outcomes and health inequalities are becoming increasingly complex, and the additional catalyst of Covid-19 has only made this all the more evident. Mental health issues have started to play an increasingly important role, with the total cost of mental ill-health estimated at over 4% of GDP across the 28 European Union member states (OECD, 2019). Issues of inefficiency and waste have further come to light, with the OECD estimating that up to 20% of health expenditure in the EU is wasteful and could be reallocated to better uses. The issue of health inequality is arguably more prevalent than ever before, with the Coronavirus pandemic making only too evident the way socioeconomic status and individual health outcomes interact.

In the aftermath of the Coronavirus pandemic, further analysis of the determinants of both health status outcomes and health inequality is of the utmost importance to policymakers. At this current moment, there is little data on the impact of the Coronavirus on health status variables such as life expectancy, predicted years of life lost, or self-reported health measures. Thus, to have any hope of using rigorous statistical analysis to draw inferences for policymaking in this time of truly unprecedented crisis, we must turn to the historical data. This paper therefore looks to make a unique contribution to the existing body of literature on the determinants of health status by considering a more comprehensive model than previous contributions in this area, using an enhanced database collated from a variety of reputable sources, and synthesising this with a rigorous analysis of the determinants of health status differentials and health inequality across the European Union member states.

1.1 Theoretical Underpinnings

From a purely economic and theoretical perspective, health status of a population is of interest to policymakers as a determinant of human capital, and thereby, the rate of economic growth, as depicted in the Augmented Solow model (Mankiw et al, 1992). We follow Barro (1997) in formulating this neoclassical growth model in the following general form:

$$\dot{y} = f(y, y^*)$$

where $\dot{y}$ denotes the growth rate, y the current level of output, and y^* the steady-state level of output, all in terms of effective workers.

The growth rate of output correlates negatively to the current level of output since we assume decreasing marginal returns to capital per effective worker. On the other hand, growth correlates positively to the steady-state level of output, which in turn is determined by factors that are a function of health status, such as the quantity and quality of the labour supply, the rate of fertility, and the savings rate (Barro, 1997). Better health improves productivity and increases labour quality, as well as creating the incentive to supply a higher quantity of labour. Therefore, through affecting both the quality and quantity of labour supplied, health status is a determinant of the growth rate of an economy. Thus, from an economic theory perspective, understanding the factors that determine health status is crucial for enabling policymakers to formulate policies that encourage consistent economic growth. This comprises one of the principal motivations behind a desire to understand health status determinants.

This theoretical perspective fails to account for the social and moral considerations as to why the health status of a population is important for policymakers. However, all of these motivations, be they economic, social, or moral considerations, have only become all the more powerful in light of the Coronavirus pandemic we are currently facing. This paper thus aims to offer further illumination on the determinants of health status outcomes and inequalities within the European Union in order to draw policy inferences as to where the most effective and targeted response should be directed in order to have the maximum impact in combating the current pandemic.

1.2 Literature Review

A plethora of literature has been written on the determinants of health status outcomes across both Europe and the OECD using a wide range of data sources, however, few of them consider as broad a range of health care factors as we aim to analyse here. Early literature tended to focus heavily on a health production function approach, with health expenditure and health resources being the principal variables of interest. Over time, the important role of risk factors was also recognised, and a broad consensus on the inputs into this health production function was reached, namely health care resources per capita, a vector of life-style factors, and a vector of socio-economic factors. However, within these categories, there is further debate regarding the specific inputs into the production function, and few works consider a truly comprehensive model, often preferring to focus on a smaller selection of variables for brevity. Horrace, Shaw and Vogel (2005) found pharmaceutical expenditure per capita, the population's age distribution and risk factors such as alcohol consumption per capita and a population's obesity rate all to be significant causal factors affecting health status. Pocas and Soukiazis' (2011) analysis placed a greater weight on risk factors and socioeconomic factors, concluding that income per capita, education, healthcare resources per capita, and a vector of risk factors were all significant determinants.

More recently, the literature in this area has begun to move towards methods of analysis which account for efficiency considerations. Additionally, given the rising issues of inefficiency and wasted health care resources in health systems in many developed nations, data envelopment analysis (DEA) has proved to be especially popular. Joumard et al. (2008) used data envelopment analysis in conjunction with a panel regression approach and found that both methods consistently suggested that population health status could be greatly improved in most OECD countries, whilst keeping health care inputs constant. Joumard's DEA results suggested that potential efficiency gains in the health care sector might be large enough to raise life expectancy at birth by 2 years on average across the OECD, whilst her panel regression analysis suggested that risk factors have the most statistically significant impact on life expectancy (as a proxy measure for health status), as well as socioeconomic variables, such as education and income per capita.

The more recent work of Antunes, Pocas, and Soukiazis (2020) concluded that per capita income, education levels, risk factors and medical staff per capita (a proxy for healthcare resources) were all significant determinants. Despite being arguably the most comprehensive work to date in this area, this paper continues the trend in the literature of taking a very broad overview approach towards healthcare resources. This paper aims to take a more in-depth approach by considering variables proxying availability and accessibility of care, quality of care, and quantity of health care resources, and thus, offer a unique contribution to the existing literature in this regard.

However, even when considering more recent contributions, much of the existing literature can be criticised for an excessive focus on quantitative aspects of health status, using conventional health proxies such as mortality rates or life expectancy. It is evident why this is the case, given the increased availability and quality of data for these measures relative to measures of the more qualitative aspects of health status. However, this excessive focus on “living longer” rather than “living better” has meant many works fail to capture the multidimensionality of health status and the importance of quality of life factors. This paper will therefore aim to consider both quantitative and qualitative measures of health status, in the hopes of fully capturing the many multifaceted aspects of a population’s health status, and again offering a unique contribution to the academic discussion on this issue.

2 Method

2.1 Data

Data were collected for all 28 European Union member states from 1980-2018 and for 12 OECD countries that were not European Union members for comparative purposes. The year 1980 was chosen as a starting point for our data for several reasons. Firstly, the 1980s are generally considered to mark a turning point of the rightwards movement of the median voter in most developed countries, and the accompanying move in policy. Although there have been movements along the political spectrum since then, this was arguably the most recent significant swing and as such, it is hoped that attitudes towards health policy will have remained broadly consistent throughout. Secondly, of course, for some of the less developed nations in the European Union, data collection prior to 1980 proved to be quite difficult and the quality of the data could not be ensured.

Initial data collection came from the OECD Health Statistics Database 2019; however, a significant challenge was securing data for non-OECD countries that were EU members. This data was principally obtained from Eurostat, the World Bank Data Bank, and the World Health Organisation's European Health Information Gateway. Some data had to be scaled to ensure that data from different databases were consistent, however, during this process care was given to maintain data integrity. All currency values were measured in real terms with constant purchasing power parities (PPP), using the OECD base year. In all cases, except for our self-reported dependent variables, measures other than self-reported values were used to avoid attenuation bias due to measurement error (Wooldridge, 2009).

The results of the Variance Inflation Factor test on initial regressions showed that collinearity was an issue, with some variables having VIF values far exceeding 10. Thus, it was determined that further data collection should be undertaken to mitigate this. Subsequently, further data were collected from the United Nations' Statistics Division, the World Health Organisation, the World Bank and Eurostat. Finally, access was obtained to the Survey of Health, Aging and Retirement in Europe (SHARE) database, which proved to be an invaluable source of additional data for this project.

However, even after further data collection, there remained the issue of missing data. Given that this was the case for the non-OECD countries that were EU members, it is possible that bias in our results could have arisen from systematic missing data. This would be the case if the probability of the data being missing was related to the values the data take and would imply that the Gauss-Markov assumption of random sampling was violated. Given the potential for bias here, multiple imputation using multivariate normal regression patterns was used to accommodate arbitrary missing values as a means of mitigating the missing data problem. Throughout this, care was given to maintain data integrity

Variable notation	Description	Significance	Mean	Standard Deviation	Minimum	Maximum
LE	Life expectancy at birth for the total population	Dependent variable. Proxy for health status capturing the **quantitative** aspects. Chosen because of quality and quantity of data available for this indicator.	76.01	3.69	66.8	83.4
Smok	Percentage of population aged 15+ who are daily smokers	Variable capturing one of the **risk factors**. Expected coefficient sign is **negative** due to health implications.	27.08	6.69	10.4	50.5
Drink	Litres of alcohol per capita amongst the population aged 15+	Variable capturing one of the **risk factors**. Expected coefficient sign is **negative** due to health implications.	11.02	2.48	4.29	20.1
Obes	Percentage of population classified as obese (measured)	Variable capturing one of the **risk factors**. Expected coefficient sign is **negative** due to health implications.	16.558	4.4311	6.5	28.9
Pol	Mean annual exposure to PM.25 air pollution, mg per cubic metre	Variable capturing one of the **risk factors**. Expected coefficient sign is **negative** due to health implications.	16.43	5.33	5.86	31
OOP	Out-of-pocket expenditure on healthcare per capita, US$ PPP (constant prices, constant PPPs)	Variable capturing **access to care**. Expected coefficient sign is **negative**: higher OOP expenditure indicates lacking access to care, due to the financial burden.	415.2	299.04	14.46	1370
CT	CT Exams per 1, 000 population	Variable capturing **access to care**. Expected coefficient sign is **positive**: more CT exams implies greater access	23.34	47.03	12.19	273.5
Phys	Practising physicians (FTE equivalent) per 100, 000 population	Variable capturing **health resources**. Expected coefficient sign is **positive**: more physicians indicate greater resources.	292.47	75.64	110.1	630
Grad	Medic graduates per 100, 000 population	Variable capturing **health resources**. Expected coefficient sign is **positive,** as with physicians.	10.45	4.76	3.25	32.91
HEx	Current expenditure on health per capita, US$ PPP (constant prices, constant PPPs)	Variable capturing **health resources**. Expected coefficient is **positive**: greater expenditure implies increased quantity/better quality resources	1715.93	1296	183.5	5986
Dis	Inpatient care discharges per 100, 000 population (all hospitals)	Variable capturing **quality of care**. Expected coefficient is **positive:** more hospital discharges indicates better quality care.	12, 332	753.9	6, 000	28, 936.3
Stay	Inpatient care average length of stay (all hospitals)	Variable capturing **quality of care**. Expected coefficient is **negative**: shorter hospital stays indicates better quality care (OECD 2019).	10.71	3.41	4.8	23.2
Urb	Degree of urbanisation, percentage of population	Control variable. Expected coefficient is **negative** due to health implications.	70.59	12.37	42.79	98.01
Pov	Percentage of population living on $3.20 a day (constant prices, constant PPPs)	Control variable. Expected coefficient is **negative** due to health implications.	1.7	2.39	0	11.7
Educ	Average years of schooling amongst the population aged 25+	Control variable. Expected coefficient is **positive:** more education leads to better health literacy and thus, better health behaviours in the population.	9.54	1.79	3.72	13.2
GDP	GDP per capita, US$ PPP (constant prices, constant PPPs)	Control variable. Expected coefficient is **positive:** wealthier populations can afford better healthcare.	25, 168.6	16, 899	4747.3	105, 342
CompEd	Compulsory years of education as mandated by law	Instrumental variable for Educ in our IV regressions	9.33	1.81	6	13
Tour	Tourism receipts per capita (constant prices, constant PPPs)	Instrumental variable for GDP in our IV regressions	159.6	239.9	53.7	1625

2.2 *Model Specification*

We take a similar approach to that of Joumard et al. (2008) in formulating our model. However, we develop this by including additional regressors of interest. Our model is thus:

$$\ln HS_{i,t} = \beta_0 + \beta_1 \ln Smok_{i,t} + \beta_2 \ln Drink + \beta_3 \ln Obes_{i,t} + \beta_4 \ln Pol_{i,t} + \beta_5 \ln OOP_{i,t} + \beta_6 \ln CT_{i,t} + \beta_7 \ln Phys_{i,t} + \beta_8 \ln Grad_{i,t} + \beta_9 \ln HEx_{i,t} + \beta_{10} \ln Dis_{i,t} + \beta_{11} \ln Stay_{i,t} + \beta_{12} \ln Urb_{i,t} + \beta_{13} \ln Pov_{i,t} + \beta_{14} \ln Educ_{i,t} + \beta_{15} \ln GDP_{i,t} + \varepsilon_{i,t}$$

Here ($HS_{i,t}$) is a measure of the population health status in country i at period t, alternatively:

- Life expectancy at birth (LE): this summarises the mortality pattern that prevails across all age groups of a population.

- Potential years of life lost (PYLL): a measure of the years of life lost in a population due to premature deaths. Based on the assumption that the average age of death is 75, this measure gives greater weight to deaths at younger age and less weight to deaths at older age.

- Percentage of population that considers themselves to be in good health (Good)

There are different measures of population health status, a more in-depth or philosophical discussion of which is beyond the scope of this paper. However, we feel that by considering each of these measures, which cover both quantitative and qualitative aspects of health status outcomes, we will be able to gain a sufficiently in-depth picture of health status and the multidimensionality of the associated issues to draw conclusions regarding its determinants and make inferences for policy implications from these results.

Stepwise regression using backwards selection and rejecting variables that are insignificant at a 15% significance level (we allow some flexibility with the convention of 10% here due to the endogeneity that we believe likely to be present in our model) was used on the variables for which data were collected. However, in order to ensure a parsimonious model, the author chose to select the most relevant of these variables to include in the analysis. This could have potentially introduced omitted variable bias into the regression, though in our later models, we take steps to deal with the issues that may arise from this choice.

Following the convention in the literature, all variables are in logarithmic form. Other variations of this functional form such as level-level, log-level or having the variables which are measured in years or percentages in level form (as is convention) and all others in logarithmic form were considered alongside the conventional approach of logarithmic form. Pleasingly, our results were robust to these different functional forms, thereby affording us greater confidence in our conclusions. Thus, as this log-log approximation allows the interpretation of results in terms of elasticity and also increases the normality of the distribution of the residuals, we continue to use this form throughout our analysis.

2.3 Econometric Methodology

Initial pooled OLS regressions give reasonably high R-squared values, indicating that our model is appropriate in terms of its explanatory power for health status outcomes. Following additional data collection, we find that no variable has a VIF value that exceeds 5.96 and thus, we conclude that collinearity is no longer a problem. Using Ramsey's RESET test, we see that a linear functional form is not appropriate and that higher-order terms could potentially better account for nonlinearities in the population regression function. A few variations of the model were considered with higher-order terms of variables that would intuitively be most likely to exhibit diminishing marginal returns, however, Ramsey's RESET test still indicated that our model had omitted variables (indeed, this is the case even when we include all the variables indicated as appropriate by our earlier stepwise regression). Thus, for brevity we do not pursue this possibility further. Both the Breusch-Pagan test and the White test indicate that heteroscedasticity is present. Hence, we use White's robust standard errors throughout.

To deal with unobserved heterogeneity amongst European Union member countries, we take a panel data approach. This allows us to account for unobserved, time-constant factors that affect ln(HS). If the unobserved effect is correlated with the idiosyncratic error, then pooled OLS will suffer from heterogeneity bias. Thus, a panel data approach is needed. The Hausman specification test indicates that the fixed effects (FE) method is preferred. This is consistent with the prevailing approach in the literature. Under the FE assumptions, the FE estimator is unbiased. However, these assumptions may not hold, and hence, whilst the FE approach represents a useful starting point with panel data, we must develop this further.

When we conduct Wooldridge's test for serial correlation, the null hypothesis of no first-order autocorrelation is soundly rejected. Thus, the assumption that the idiosyncratic errors are serially uncorrelated across t does not hold. We therefore use the FGLS method with panel data to correct for panel specific AR (1) serial correlation. However, FGLS is only a valid method of estimation if we have strictly exogenous regressors. As we judge that endogeneity is likely to be an issue within our model, further analysis is necessary.

Three causes of endogeneity potentially affect our model. Omitted variable bias is likely due to the variables we omitted from those selected by stepwise regression. Measurement error is also a possibility, given the nature of some of our variables and the difficulty of collecting this data. However, the principal cause of endogeneity is likely simultaneity. Both ln(GDP) and ln(Educ) exhibit a bi-directional causal relationship with ln(HS). A wealthier country is able to spend more on health care and thus improve health status, whilst a healthier population is more likely to work productively and thus benefit from efficiency wages and economic growth in the form of higher income. The same is true for education: a more highly educated country can make more informed decisions pertaining to behaviours that affect health status, whilst a healthier population is able to obtain more education. Thus, our model will suffer from simultaneity bias. To deal with this, we employ an instrumental variable approach, both in the pooled cross-sectional and panel dimensions. In order for an instrument to be valid, it of course must satisfy both the relevance condition and the exclusion condition.

ln(Educ) is instrumented by compulsory years of education, ln(CompEd). Compulsory education is likely correlated with the years of education actually consumed in a country, and there is no a priori reason to believe there is also a bi-directional causal relationship with health status here. Alternatively, ln(GDP) is instrumented by tourism receipts, ln(Tour), since tourism directly affects income, as an activity that works to stimulate the economy, and thus will be correlated with income. However, as this activity is mostly coming from those who are not citizens of that country, there is again no a priori reason to assume the exclusion condition does not hold. The Durbin-Wu-Hausman test for endogeneity shows that both ln(Educ) and ln(GDP) were endogenous at a 5% significance level. Furthermore, for the European Union member countries, both instruments passed the Sagan/Wu-Hausman test of exogeneity and the Crag-Donald test of instrument strength when compared with Stock-Yogo critical values. We thus proceed to use these instruments as a means of dealing with endogeneity.

However, even when we run instrumental variable regressions, serial correlation of the error term is not adjusted for. Hence, we cannot state with confidence that the Classical Linear Regression Model (CLRM) assumptions hold. For this reason, we also included a non-parametric regression approach in our analysis. This effectively relaxes the assumption of linearity and substitutes it for the weaker assumption of a smooth population regression function f(X), where X is a vector of regressors. Consequently, we no longer need to assume that the CLRM assumptions hold in order for our analysis to be valid. We use non-parametric local-linear Kernel regression with improved AIC to compute the optimal bandwidth. As we are making no assumptions about the functional form of our model here, this offers an interesting alternative means of estimation.

The above process was also followed for our non-European Union countries. However, collinearity remains a problem in the comparative non-European Union model, even after further data collection. This could impact the significance on our results, and thus conclusions must be drawn tentatively. Furthermore, our instruments fail the Sagan/Wu-Hausman test of exogeneity. Although we continue to use them for comparative purposes, we note that this could potentially introduce a further source of bias into our results.

3 Results

Table 2: EU Health Production Function Model Results. Dependent Variable: *ln(LE)*

Variables	Pooled cross-section OLS **(1)**	Fixed effects **(2)**	FGLS **(3)**	GMM IV **(4)**	Fixed effects IV **(5)**	Non-parametric regression **(6)**
ln(Smok)	-0.00726*** (0.0010)	-0.00927 (0.0219)	-0.00726*** (0.00103)	-0.00875 (0.0152)	-0.00823*** (0.00987)	-0.00771 *** (0.000783)
ln(Drink)	-0.0326** (0.0164)	-0.01* (0.0055)	-0.0163*** (0.00225)	-0.0264*** (0.0095)	-0.0178*** (0.0242)	-0.0137*** (0.0291)
ln(Obes)	-0.0709*** (0.0203)	-0.0385*** (0.0106)	-0.0709*** (0.0191)	-0.0211 (0.0146)	-0.0727** (0.0299)	-0.0473 ** (0.0183)
ln(Pol)	-0.0296*** (0.0125)	-0.0102 (0.106)	-0.0298** (0.0114)	-0.0243** (0.0101)	-0.0096 (0.117)	-0.0189 ** (0.0146)
ln(OOP)	0.00397*** (0.00036)	0.000888 (0.000807)	0.003797*** (0.000371)	0.000725 (0.0048)	0.00374*** (0.000332)	0.00299 *** (0.000342)
ln(CT)	0.00207* (0.00126)	0.00323 (0.304)	0.006373*** (0.00164)	0.00475*** (0.00184)	0.00288* (0.00156)	0.00565 *** (0.0019)
ln(Phys)	0.00495*** (0.000884)	0.00668 (0.025)	0.00357*** (0.000864)	0.00173 (0.00151)	0.002544*** (0.000862)	0.00563*** (0.000798)
ln(Grad)	0.0732 (0.0117)	0.0138 (0.029)	0.0732*** (0.0115)	0.0892*** (0.0199)	0.06522*** (0.0113)	0.0874*** (0.0135)
ln(HEx)	0.000513*** (0.000092)	0.0002662 (0.000201)	0.00069*** (0.000126)	0.000525*** (0.000107)	0.000375*** (0.044)	0.000409*** (0.000121)
ln(Dis)	0.000245 (0.00035)	0.00053 (0.0.000495)	0.00012 (0.0000661)	0.000161 (0.000363)	0.000917 (0.00078)	0.000928*** (0.000103)
ln(Stay)	-0.0812*** (0.0177)	-0.0789 (0.029)	-0.015*** (0.00188)	-0.0108** (0.00433)	-0.0131*** (0.00187)	-0.0155*** (0.00158)
ln(Urb)	0.0160 *** (0.00589)	0.0856* (0.0496)	-.0272*** (0.0578)	-0.0859 (0.341)	-0.0231*** (0.00651)	-0.0251*** (0.00429)
ln(Pov)	-0.00127*** (0.000222)	0.00359 (0.0791)	-0.00489*** (0.00265)	-0.00309 (0.004)	-0.00426*** (0.000265)	-0.0047*** (0.000237)
ln(Educ)	-0.0121*** (0.00295)	0.009037 (0.102)	-0.017*** (0.00446)	0.0357 (0.0103)	0.0956*** (0.0351)	0.0873*** (0.00384)
ln(GDP)	0.0000498*** (0.000078)	0.0000226** (0.0000102)	0.0000275*** (0.000006915)	0.0000242*** (0.0000923)	0.0000498*** (0.0000561)	0.0000282** (0.000331)
Constant	76.2 *** (0.677)	76.3*** (4.91)	76.2*** (0.855)	71.3 *** (02.59)	72.9 *** (1.44)	75.9*** (0.130)
Observations	976	976	976	590	590	976
R-squared	0.849	.	.	0.856	.	0.882
Tests for Joint Significance Results (F-test and Wald test).						
Risk Factors	$F^4_{959} = 47.75$ p-value = 0.000	$F^4_{932} = 66.22$ p-value =0.000	$\chi^2_4 = 77.86$ p-value = 0.000	$\chi^2_4 = 21.42$ p-value = 0.000	$\chi^2_4 = 107.87$ p-value = 0.000	.
Access to Care	$F^2_{959} = 64.70$ p-value = 0.000	$F^2_{932} = 2.56$ p-value =0.078	$\chi^2_2 = 15.23$ p-value = 0.001	$\chi^2_2 = 11.74$ p-value = 0.003	$\chi^2_2 = 10.65$ p-value = 0.005	.
Health Resources	$F^3_{959} = 50.45$ p-value = 0.000	$F^3_{932} = 2.49$ p-value =0.059	$\chi^2_3 = 83.76$ p-value = 0.000	$\chi^2_3 = 36.41$ p-value = 0.000	$\chi^2_3 = 22.21$ p-value = 0.001	.
Quality of Care	$F^2_{959} = 16.52$ p-value = 0.000	$F^2_{932} = 1.06$ p-value =0.347	$\chi^2_2 = 3.54$ p-value = 0.174	$\chi^2_2 = 4.46$ p-value = 0.108	$\chi^2_4 = 6.42$ p-value = 0.040	.
Notes	Robust standard errors are reported in parentheses. This is deemed appropriate due to the results of the White test on a preliminary regression, reported in Table 5. *, **, *** represent 10%, 5%, and 1% significance levels respectively. In (4) and (5), *ln(Educ)* is instrumented by *ln(CompEd)* and *ln(GDP)* is instrumented by *ln(Tour)*.					
	Our constant term for (6) is the conditional mean value of *ln(LE)* in the non-parametric regression.					

Table 3: Non-EU OECD Health Production Function Model Results. Dependent Variable: *ln(LE)*

Variables	Pooled cross-section OLS **(1')**	Fixed effects **(2')**	FGLS **(3')**	GMM IV **(4')**	Fixed effects IV **(5')**	Non-parametric regression **(6')**
ln(Smok)	-0.01537*** (0.00215)	0.0471 (0.0483)	-0.0323** (0.0128)	-0.0940 (0.072)	-0.0422*** (0.00267)	-0.0749*** (0.0393)
ln(Drink)	-0.0201*** (0.00556)	-0.0749 (0.0199)	-0.0133*** (0.0038)	-0.0728 (0.104)	-0.0485*** (0.0171)	-0.0325 (0.032)
ln(Obes)	-0.066** (0.0251)	-0.011* (0.00711)	-0.092 (0.128)	-0.0186*** (0.00521)	-0.0108*** (0.00258)	-0.0207 (0.265)
ln(Pol)	-0.0192 (0.0268)	-0.0202 *** (0.0629)	-0.0222 (0.216)	-0.0145*** (0.00511)	-0.0159* (0.00922)	-0.0913 *** (0.0271)
ln(OOP)	0.000402 (0.0004015)	0.000748 (0.00118)	0.000872* (0.000449)	0.0013 (0.00105)	0.000354*** (0.0000888)	0.000941*** (0.000313)
ln(CT)	0.00194*** (0.00287)	0.007 (0.00603)	0.0024 (0.00279)	0.0217*** (0.00713)	0.0091** (0.00441)	0.00458 (0.035)
ln(Phys)	0.0228** (0.011)	0.0245 (0.0199)	0.0179 (0.0905)	0.0242* (0.127)	0.0202*** (0.0013)	0.0346*** (0.00596)
ln(Grad)	0.0205 (0.0250)	0.02247 (0.0445)	0.0662 (0.237)	0.0463 (0.0441)	0.0733* (0.0437)	0.0484 (0.0584)
ln(HEx)	0.000756 (0.127)	0.000155*** (0.000068)	0.00071*** (0.000274)	0.000428*** (0.000101)	0.000202*** (0.0000984)	0.000259*** (0.0000202)
ln(Dis)	0.000377 (0.000339)	0.000159* (0.00009)	0.000217 (0.000341)	0.000423*** (0.000116)	0.000162 *** (0.0000277)	0.000109*** (0.0000069)
ln(Stay)	-0.0126*** (0.0041)	-0.0170* (0.0085)	-0.0455* (0.00276)	-0.0382*** (0.0107)	-0.0145** (0.00586)	-0.01 (0.702)
ln(Urb)	0.0144*** (0.0235)	-0.0288*** (0.00694)	-0.0266**** (0.00201)	-0.0112** (0.00479)	-0.0307*** (0.00457)	-0.022** (0.00916)
ln(Pov)	-0.00792 (0.0309)	-0.0736* (0.0407)	-0.0533*** (0.0107)	-0.00335 (0.0454)	-0.0728*** (0.0125)	-0.0507*** (0.0185)
ln(Educ)	0.0189* (0.0108)	0.0292 (0.0260)	0.0122* (0.00644)	0.0136* (0.00734)	0.0148 *** (0.00423)	0.0111*** (0.00121)
ln(GDP)	0.000208*** (0.0000151)	0.000167*** (0.0000322)	0.000121*** (0.0000172)	0.000512*** (0.000653)	0.000218 ** (0.000102)	0.00029*** (0.000123)
Constant	73.1*** (2.869)	73.337*** (8.799)	75.9*** (4.415)	68.69*** (7.09)	71.36 *** (3.41)	73.2*** (0.25)
Observations	436	436	436	428	428	436
R-squared	0.906	.	.	0.816	.	0.918
Notes	Robust standard errors are reported in parentheses. This is deemed appropriate due to the results of White test on a preliminary regression, reported in Table 3. *, **, *** represent 10%, 5%, and 1% significance levels respectively. In (4) and (5), *ln(Educ)* is instrumented by *ln(CompEd)* and *ln(GDP)* is instrumented by *ln(Tour)*. Our constant term for (6) is the conditional mean value of *ln(LE)* in our non-parametric regression.					
	Countries for which data were collected: Australia, Brazil, Chile, Iceland, Israel, Japan, Korea, Mexico Russia, Switzerland, Turkey, United States.					

Table 4: EU Health Production Function Model Results with Different Dependent Variables						
	Dependent Variable: *ln(Good)* where *Good* = percentage of population with self-rated good health			Dependent Variable: *ln(PYLL)* where *PYLL* = potential years of life lost per 100, 000 population aged 75+		
Variables	Pooled cross-section OLS **(1'')**	FGLS **(3'')**	Fixed effects IV **(5'')**	Pooled cross-section OLS **(1''')**	FGLS **(3''')**	Fixed effects IV **(5''')**
ln(Smok)	-0.0498 (0.0467)	-0.0217* (0.0156)	-0.0137** (0.000108)	0.112* (0.00262)	0.253* (0.239)	0.0875* (0.0152)
ln(Drink)	-0.725*** (0.107)	-0.314* (0.00473)	-0.269*** (0.0144)	0.184** (0.00585)	0.413* (0.00672)	0.265** (0.095)
ln(Obes)	-0.143*** (0.0857)	-0.571*** (0.0907)	-0.509*** (0.00486)	0. 158** (0.00642)	0.106*** (0.011)	0.211 (0.0146)
ln(Pol)	-0.166** (0.057)	-0.195*** (0.0588)	-0.325 (0.346)	0.247*** (0.0033)	0.53*** (0.0816)	0.243** (0.101)
ln(OOP)	-0.0214*** (0.00162)	-0.00395*** (0.000833)	-0.0407** (0.00591)	0.0527*** (0.0131)	0.0296 (0.103)	0.00725* (0.00278)
ln(CT)	0.0590*** (0.00796)	0.0127*** (0.00416)	0.0383** (0.00588)	-0.165*** (0.00453)	-0.527 (0.566)	-0.0475** (0.0184)
ln(Phys)	0.0187*** (0.00433)	0.00133* (0.000232)	0.0655** (0.00898)	-0.883*** (0.0261)	-0.602** (0.312)	-0.173** (0.00155)
ln(Grad)	0.687*** (0.0588)	0.0721*** (0.0268)	0.122*** (0.313)	-0.199** (0.0311)	-0.305 (0.303)	-0.0892*** (0.0199)
ln(HEx)	0.00651*** (0.000593)	0.000248 (0.000373)	0.00123*** (0.00018)	-0.0032** (0.00367)	-0.0556*** (0.00404)	-0.009*** (0.000338)
ln(Dis)	0.000809 (0.0248)	0.000276** (0.0001)	0.000507 (0.00119)	-0.00948 (0.0356)	-0.00727 (0.00215)	-0.00466** (0.000363)
ln(Stay)	-0.105*** (0.0824)	-0.237*** (0.0685)	-0.262 (0.0307)	0.354** (0.0591)	0.213** (0.0869)	0.108** (0.0433)
ln(Urb)	-0.362** (0.244)	-0.239*** (0.0322)	-0.113 (0.322)	0.102*** (0.0017)	0.899*** (0.0656)	-0.0859 (0.341)
ln(Pov)	0.158 (0.133)	-0.295*** (0.0636)	-0.24** (0.0342)	0.256 (0.674)	0.363*** (0.0913)	0.31 (0.400)
ln(Educ)	0.13*** (0.0180)	0.342 (0.334)	0.154*** (0.0196)	-0.176** (0.00184)	-0.117*** (0.0151)	-0.156*** (0.0351)
ln(GDP)	0.000377*** (0.000335)	0.000173*** (0.000278)	0.000339* (0.000233)	-0.00142*** (0.000198)	0.00915 (0.813)	-0.00242** (0.000923)
Constant	38.2*** (3.99)	37.4*** (3.52)	37.97*** (3.97)	12.6*** (0.632)	15.8*** (0.6557)	16.2*** (2.59)
Observations	975	975	479	771	771	462
R-squared	0.717	.	.	0.738	.	.
Notes	Robust standard errors are reported in parentheses. This is deemed appropriate due to the results of White test on a preliminary regression, reported in Table 5. *, **, *** represent 10%, 5%, and 1% significance levels respectively. In (4) and (5), *ln(Educ)* is instrumented by *ln(CompEd)* and *ln(GDP)* is instrumented by *ln(Tour)*. Our constant term for (6) is the conditional mean value of *ln(LE)* in the non-parametric regression.					

Table 5: Results of Diagnostic Tests.	Dependent Variable: *ln(LE)*				
Issue	Test	EU nations	Comments	Non-EU nations	Comments
Hetero-scedasticity	Breusch-Pagan/Cook-Weisberg	χ_1^2 = 9.56 p-value = 0.0000	H_0: Constant variance is rejected at all significance levels and we assume heteroscedastic errors.	χ_1^2 = 12.16 p-value = 0.0000	H_0: Constant variance is rejected at all significance levels and we assume heteroscedastic errors.
	White	χ_{152}^2 = 662.51 p-value = 0.0000	H_0: Constant variance is rejected at all significance levels and we assume heteroscedastic errors.	χ_{152}^2 = 344.23 p- value = 0.0000	H_0: Constant variance is rejected at all significance levels and we assume heteroscedastic errors.
Functional Form Misspecification	RAMSEY Reset	F_{956}^3 = 30.9 p-value = 0.0000	H_0: Model has no omitted variables is rejected and higher-order terms of some variables may be appropriate.	F_{354}^3 = 15.48 p-value = 0.000	H_0: Model has no omitted variables is rejected and higher-order terms of some variables may be appropriate.
Collinearity	Diagnostics	Highest VIF = 4.17, R-squared = 0.857	As no variables exhibit VIF values which exceed 5, this indicates that collinearity is no longer a problem in the model following the additional data collection that was conducted to rectify this issue.	Highest VIF = 18.84 R-squared = 0.906	A high R-squared value along with few significant regressors, and VIF values for some variables which exceed 10 suggests that collinearity is an issue in this model.
Normality	Test for skewness and kurtosis in the one-way error-components model	For e: χ_2^2 = 7.15 p-value =0.0281 For u χ_2^2 = 0.63 p-value = 0.7293	We fail to reject the null hypothesis that the unobserved effect is normally distributed. However, we do reject the hypothesis that the idiosyncratic error is normally distributed.	For e: χ_2^2 = 2.83 p-value =0.24251 For u χ_2^2 = 33.25 p-value = 0.000	We reject the null hypothesis that the unobserved effect is normally distributed. However, we fail to reject the hypothesis that the idiosyncratic error is normally distributed.
Serial Correlation	Wooldridge	F_{27}^1 = 26.19 p-value = 0.0000	H_0: No first-order autocorrelation is rejected, and we assume autocorrelation is an issue.	F_{12}^1 = 32.6 p-value = 0.000	H_0: No first-order autocorrelation is rejected, and we assume autocorrelation is an issue.
Endogeneity	Durbin-Wu-Hausman	χ_2^2 = 47.83 p-value = 0.0000	H_0: Variables are exogenous is rejected and thus, we assume endogeneity is an issue.	χ_2^2 = 47.83 p-value = 0.0000	H_0: Variables are exogenous is rejected and thus, we assume endogeneity is an issue.
Overidentifying restrictions	Test for orthogonality of the instruments with Hansen's J Statistic	Equation exactly identified and a p-value of 0.2068	We fail to reject the joint null hypothesis that the instruments are orthogonal to the error term. This suggests that the instruments are appropriate.	Equation exactly identified and a p-value of 0.000	We reject the joint null hypothesis that the instruments are orthogonal to the error term. This casts doubt on the validity of the instruments used.
Weak identification	Test for relevance of the instruments using the Kleibergen-Paap Wald statistic	Kleibergen-Paap Wald F statistic = 39.76 Stock-Yogo weak ID test for 10% maximal IV size given as 19.93	This indicates that the excluded instruments are strongly correlated with the endogenous regressors. The test statistic far exceeds the Stock-Yogo rule of thumb of a value of 10 and the 10% maximal IV value. In this regard, our instruments are valid.	Kleibergen-Paap Wald F statistic = 20.414 Stock-Yogo weak ID test critical value for 10% maximal IV size given as 19.93	This indicates that the excluded instruments are strongly correlated with the endogenous regressors. The test statistic exceeds the Stock-Yogo rule of thumb of a value of 10 and the 10% maximal IV value. In this regard they are valid instruments.

3.1 *Interpretation*

For the EU model, considering life expectancy as our proxy for health status, it is notable that ln(HEx) has significant explanatory power at the 5% level in all model specifications, with the estimated coefficient lying in the range (0.0000226 , 0.000498). Thus, we can interpret this as an elasticity using the approximation:

$$\%\Delta LE \approx\approx (100 * \beta_9)\Delta HEx$$

This implies that, ceteris paribus, an increase in health expenditure per capita of $100 (PPP) would increase life expectancy in that country by between 0.226% and 0.498%. From this, we can see that health expenditure has a significant positive effect on life expectancy, as we would intuitively expect, given that increased health expenditure can improve both the quality and availability of care. Our other variables that are proxies for healthcare resources also have the expected positive signs and are significant at a 1% level once endogeneity and heterogeneity are taken into account in our later models.

Our risk factor variables are shown to have significant explanatory power at the 10% level in most of our models. In particular, ln(Drink) is significant in all models, with a coefficient lying in the interval (-0.01 , -0.0325). This implies that, holding all other factors constant, increasing the litres of alcohol consumed per capita by one litre would lead to a decrease in life expectancy of between 1% and 3.25%. Given that life expectancy is a variable which does not vary greatly over time, issues with our model may mean that this is an overestimate of the true effect of alcohol consumption. In the case of the European Union, we can see that the coefficients on all risk factor variables have the expected signs in all models and, with the exception of ln(Pol), are significant in later models, suggesting that these risk factors have a negative impact on life expectancy.

ln(Stay) is the only quality of care variable that is significant in our parametric model specifications. The coefficient is negative as expected and, in all models except our initial pooled OLS, is significant at a 5% level. We therefore conclude that if a shorter average stay indicates better quality of care, then quality of care has a positive effect on longevity.

ln(OOP) is one of our proxy variables for access to care, with the idea that the financial burden of out-of-pocket expenditure on healthcare can limit access to care and suggest limited healthcare coverage. In all models where out-of-pocket expenditure is significant, we have the expected negative coefficient, suggesting that reduced access to healthcare decreases life expectancy. Our other proxy for access to care, ln(CT) exhibits the expected positive sign also, suggesting that countries which are able to provide greater access to diagnostic exams have longer life expectancy.

Of particular interest amongst our control variables is ln(Educ) and ln(GDP), however we can make no inferences here except in the case of the later models using an instrumental variable approach, as in earlier models the coefficients on these variables will be biased due to endogeneity. We can see that ln(GDP) is significant in both models once we have corrected for this endogeneity, and ln(Educ) is significant in (5), both at a 1% significance level. Furthermore, we can see that once we have corrected for endogeneity, the coefficients on both variables are revised upwards, suggesting that the previous estimates of the effects of education and income on life expectancy were underestimated. Our other control variables representing socio-economic factors all exhibit the expected coefficient signs when we have significant results, and in particular are significant in the later models.

Given the violations of the CLRM assumptions even in our later models, the results of our non-parametric regression are of particular interest. Although non-parametric regression will logically be less efficient than parametric regression, non-parametric regression yields estimate that are consistent even when the CLRM assumptions do not hold. Our results indicate the ceteris paribus average marginal effect of each variable on the mean value of life expectancy conditional on our regressors. Thus, from this we can see that the average marginal effect of each of our main regressors has the expected sign and is significant at a 5% level based on bootstrap standard errors. These ceteris paribus average marginal effects inform us as to the effect of infinitesimal changes in our regressors on life expectancy. It is encouraging to see that our results from the non-parametric model are not dissimilar to those from our parametric regressions; this speaks well to the validity of our estimates. An additional advantage of non-parametric regression is that it is robust to functional form misspecification, which we previously saw to be an issue with our parametric models.

3.2 Further Analysis and Discussion

The aim of our analysis was to consider particularly the role of risk factors, quality of care, access to care, and healthcare resources in determining health-status outcomes. We thus perform F-tests for the joint significance of of the variables in these categories (Table 2).

We find that our risk factor variables are jointly significant at a 5% level in all parametric model specifications above. Our access to care variables are jointly significant at a 10% level in all parametric specifications above, as are our proxies for healthcare resources. However, in (2), (3), and (4), our quality of care variables are not jointly significant at a 10% significance level. This surprisingly indicates that potentially quality of care does not play a significant role in determining health status outcomes, or at the very least that it is less important than risk factors, access to care and healthcare resources.

We now turn the focus of our analysis towards the determinants of health status inequalities across the European Union, using Or's (2000) equation to estimate the effect of individual variables on cross-country differentials in health status outcomes. Using the following equation, we obtain the results in Table 7 using Germany (as arguably the most developed economy in the European Union) as our reference country throughout:

$$D_{i,v} = \alpha_v * (\ln(V_{i,t}) - \ln(V_{Ger,t})) * 100$$

where $D_{i,v}$ is the percentage-point contribution of variable V to the log percentage difference in health status for country i and Germany in period t.

Table 6 demonstrates the relative contribution of different variables to cross-country health status differentials, showing risk factors (ln(Smok), ln(Drink), and ln(Obes) to be of particular importance in this. This is consistent with Or (2000) and our own earlier results. We can also see that ln(Stay) plays an important role in determining cross-country health status differentials, whilst our access to care variables and ln(Phys) appear to play a comparatively smaller role. Interestingly, ln(Grad) seems to make the greatest contribution here, however, this is likely skewed since several southern and eastern European nations are known to produce many medical graduates but export them elsewhere. Thus, we take this result with caution. It is also worth noting that the residual component of our analysis is relatively large, and as such we can infer that other factors are important in determining health status inequalities across the European Union. It is likely that socioeconomic factors would be of particular importance in determining this residual, suggesting that although risk factors and both quality and access to care are notable in determining health inequalities, policymakers might also need to turn their attention elsewhere in the aim of reducing health inequalities in the European Union.

The results of our analysis for non-European countries, presented in Table 3, may at first glance seem not dissimilar to the results of our analysis with European Union member states (Table 2). Indeed, the significant coefficients can be interpreted in much the same way and the results do not seem to be notably different, particularly for the later models. However, the Chow test (robust to heteroscedasticity) for differences in population regression functions across groups yields a test statistic of $F(17, 1378) = 70.14$ (p-value = 0.000). Thus, we interpret this as evidence that our non-European comparator countries follow a different population regression function to European countries. This is corroborated when we include a dummy variable for EU membership and the according interaction terms in our regression and find the coefficients to be jointly significant. This indicates that further separate research is needed on the determinants of health status outcomes in non-European OECD countries.

Table 7: Determinants of Cross-National Variations in Health Status. Differentials between Germany and other EU nations, 2017.

Variable	*ln(LE)*	*ln(Smok)*	*ln(Drink)*	*ln(Obes)*	*ln(OOP)*	*ln(CT)*	*ln(Phys)*	*ln(Grad)*	*ln(HEx)*	*ln(Stay)*	Residual
Coefficient	.	-0.00823	-0.0178	-0.727	-0.0037	0.0029	0.00254	0.0652	0.00038	-0.131	.
Austria	0.084	-0.13	-0.20	-0.66	-0.15	-0.09	0.05	1.19	0.00027	0.91	-0.85
Belgium	0.056	-0.24	-0.17	0.03	-0.07	0.09	-0.08	1.36	0.00022	0.60	-3.46
Bulgaria	-1.92	-0.23	-0.27	0.93	-0.20	-0.23	-0.14	1.39	-0.00232	-0.79	-7.95
Croatia	-1.02	-0.23	-0.29	0.75	-0.06	-0.14	-0.06	1.01	-0.00114	0.60	-2.60
Cyprus	-0.2	-0.17	-0.14	-0.07	-0.19	-0.01	-0.02	-1.77	-0.00027	-0.95	-2.79
Czech Republic	-0.65	0.02	-0.11	1.49	0.22	-0.10	-0.03	2.32	-0.00063	-0.58	-3.87
Denmark	-0.06	0.09	0.32	-0.80	0.05	0.04	0.10	1.48	-0.00014	-0.59	-0.75
Estonia	-0.91	0.03	0.19	0.00	0.20	0.04	-0.05	-0.31	-0.00036	-1.07	-2.96
Finland	0.084	0.30	0.46	-0.07	-0.04	0.04	-0.05	0.90	0.00049	0.72	-2.53
France	0.33	-0.13	-0.13	-0.13	0.04	0.07	-0.08	0.90	0.00098	-0.57	0.35
Greece	-0.08	-0.13	-0.02	0.90	0.02	0.00	0.05	-0.37	0.00022	0.72	-4.26
Hungary	-1.6	-0.13	-0.23	1.33	0.15	-0.06	-0.06	1.19	-0.00189	-0.99	-2.79
Ireland	0.22	0.04	-0.02	0.32	-0.11	-0.38	-0.08	3.75	0.00027	0.95	-2.25
Italy	0.44	-0.05	0.27	-0.69	0.01	-0.14	-0.02	0.79	0.00080	1.73	-1.46
Latvia	-1.9	-0.23	-0.27	0.57	0.10	0.04	-0.07	2.40	-0.00170	0.76	-5.19
Lithuania	-1.7	-0.05	-0.22	1.35	0.08	-0.11	0.02	3.08	-0.00133	0.60	-8.46
Luxembourg	0.22	0.13	0.18	0.20	0.09	0.10	-0.09	-0.99	0.00045	0.45	0.15
Malta	0.25	-0.06	-0.16	2.01	-1.59	-0.15	-0.02	2.57	0.00054	0.09	-8.45
Netherlands	0.11	0.09	0.49	-0.55	-0.01	-0.13	-0.04	1.85	0.00295	-1.53	-0.06
Poland	-0.99	-0.17	0.05	0.48	0.14	-0.12	-0.15	-0.60	-0.00073	-1.15	-3.77
Portugal	0.028	0.13	-0.15	0.00	-0.01	-0.28	-0.01	1.91	0.00054	-0.29	-1.27
Romania	-1.77	-0.01	-0.24	0.32	-0.22	-0.26	-0.14	4.39	-0.00199	0.42	-8.03
Slovak Republic	-1.2	-0.01	0.21	-0.34	0.27	0.01	-0.05	2.22	-0.00123	-1.60	-6.11
Slovenia	-0.08	-0.02	0.14	0.63	0.06	-0.21	-0.08	2.44	0.00027	-1.15	-6.19
Spain	0.55	-0.13	0.03	0.48	-0.01	-0.07	-0.02	1.22	0.00119	1.60	-3.53
Sweden	0.31	0.49	0.76	-0.48	0.02	0.11	-0.01	-0.31	0.00031	1.84	-6.11
United Kingdom	0.028	0.07	0.21	1.93	0.03	-0.03	-0.11	0.45	-0.00014	1.33	-5.87
Notes	Coefficients used here are from our estimation of (5) and we only consider variables that were statistically significant in (5) for our analysis here. Values here were computed using Or's (2000) equation (3) for 2017, with Germany (as one of the more developed nations in the EU) as a reference country.										

3.3 *Limitations and Further Research*

Although our results are pleasingly robust to different dependent variables, which incorporate both quantitative and qualitative aspects of health status, it is worth noting that there are, of course, several limitations to our analysis that must be accounted for and considered when interpreting the results. Although our results conform well with the expected intuition and the general consensus in the literature, there are several potential sources of bias in our model that must be noted. Thus, our results must be interpreted with some degree of caution, particularly when drawing inferences for policymakers.

Firstly, we have the issue of missing data, which was particularly felt for the non-European model and models with ln(PYLL) and ln(Good) as the dependent variable. However, it posed a problem for our European Union model too. The use of multiple imputation mitigated this by randomly drawing imputations from a distribution of imputations constructed from regression analysis and introducing error variances to each imputation. However, this was based on the assumption that data is missing at random, and as we previously discussed, this may not be the case here. Thus, it is an imperfect solution at best. Further research might be able to take advantage of patient registries collated by the European Medicines Agency, which is currently an on-going process, as well as making greater use of individual countries' national statistics databases.

There are also potential issues with our methodology that might be explored by further research, namely the choice to use a fixed effects approach in our panel data regression. Serial correlation in the error term indicates that we should perhaps use the first-differences estimator rather than fixed effects estimation, as does the fact that t is large. Inference from fixed effects estimation is very sensitive to violations of our fixed effects assumptions, particularly nonnormality, heteroscedasticity and serial correlation in the idiosyncratic errors. However, there could potentially be feedback between our error term and future outcomes of our explanatory variables, causing the fixed effects estimator to have substantially less bias than the first-differences estimator. Although we chose to prioritise the unbiasedness of our estimator and therefore went with a fixed effects approach, further research might want to conduct a sensitivity analysis using both of these methods.

It is worth noting that our instrumental variable approach also faced some limitations. Our results in these regressions were vastly sensitive to our choice of instrument. As such, the results of our sensitivity analysis using alternative instruments varied. This was also noted to be an issue by Joumard et al. (2008). Further research might therefore wish to spend more time on this, in particular focusing on finding an instrument that would be valid both for the European Union model and for comparator non-EU countries.

Finally, our analysis offers a unique contribution, in that it synthesises the inputs of the health production function used in previous literature to formulate a comprehensive model and runs the analysis on more qualitative measures of health status. However, the lack of consideration of efficiency issues does represent an important gap that further research should endeavour to investigate. Our analysis gives little consideration to the importance of the efficiency of healthcare systems, and the differentials in health status arising from this across countries. The OECD (2019) identifies the efficiency and fiscal sustainability of healthcare systems as one of the important determinants of health status outcomes. Further research might wish to employ a data envelopment analysis approach, assuming an output orientated model and variable returns to scale (VRS) using our regressors as inputs and one of our health status proxies as an output to allow us to assess the healthcare production efficiency of different European Union countries and the value for money of healthcare expenditure across Europe.

4 Conclusions

4.1 Critical Factors

This paper uses various methodologies, including a panel data approach, instrumental variables, and non-parametric regression, to analyse and estimate the impact of causal factors in determining health status outcomes. The conclusions of this paper stand broadly in line with the overall consensus that currently prevails in the literature on this issue. However, this paper brings a unique perspective to the debate by including a more comprehensive model with greater focus on both the more qualitative proxy measure for health status and on analysing the factors determining health status inequalities. With the current Coronavirus pandemic, understanding the determinants of health status outcomes and inequalities has never been more crucial, and as such this paper hopes to have made a valuable contribution to an important area of analysis.

Despite its limitations, our analysis offers many insights. We can see that risk factors, access to care and healthcare resources are all significant determinants of health status outcomes and have explanatory power in terms of cross-country differentials within the European Union. However, quality of care is interestingly not shown to be significant in this context. This is possibly due to the difficulty of finding a proxy variable which can truly capture quality of care, and the data issues this entails. The important role of risk factors and healthcare resources is broadly consistent with the prevailing consensus in the literature. Although our analysis principally focuses on life expectancy, due to the availability of data, these results are also pleasingly robust to different measures of health status being used (namely measures of the more qualitative aspects). This can be seen from our analysis with predicted years of life lost and self-rated health. Furthermore, we see that although the population regression function for non-European Union countries is distinct from that of the European Union nations, many of the same variables are significant across different estimation methods for both models.

4.2 Policy Implications

On the basis of our analysis, we propose the following policy recommendations. At a high level, to generate an overall improvement in health status in the European Union, policies should be mostly directed to alter unhealthy lifestyle choices through a combination of taxation, better education, and information distribution regarding these risk factors, leading to increased health literacy. Furthermore, policies that improve the overall level of education like investment in the education system or introducing new educational innovations, such as the rise of the Degree Apprenticeship, could be an effective manner in which to proceed. Traditional macroeconomic policies of expansionary monetary and fiscal policy might also prove to be successful in affecting health status outcomes via their effect on income per capita. Furthermore, greater expenditure and investment in health care is strongly supported by our analysis as a policy to improve overall health status in the European Union, given that it is significant in increasing life expectancy in all of our model specifications for the EU.

Importantly, however, when considering each of these policies as potential routes for improving health status outcomes, one must also consider the marginal cost and benefit of each. In today's difficult climate, it is essential that policymakers are able to ensure value for money from their policy decisions. As such, this might support the implementation of policies such as greater investment in education, given the likely positive multiplier effects of this throughout the economy as the policy comes to fruition. Equally, given the long-time lag before any benefits to a policy like this is seen, one might alternatively consider the benefits of greater investment and expenditure on health care as a shorter-term solution. Ultimately, each country within the EU is in itself a heterogeneous entity, and as such, the most effective policy response will vary throughout. Our paper hopes to have offered some illumination on which policies have the potential to be most effective, but each country will have unique considerations that must be accounted for, even more so in light of the current Covid-19 situation. There can be no "one size fits all" approach to policymaking in this regard.

As we have heard so often over the past few months, the Covid-19 pandemic has both discriminated based on and further exacerbated existing socioeconomic inequalities, and this is of course particularly the case with health inequalities. Thus, policymaking implications that can be drawn regarding health inequalities are arguably of greater importance than ever before. Our results, and the trends observed more recently throughout the Covid-19 pandemic, suggest that policy responses concerned with population groups that are at high risk of experiencing poorer health is hugely important in reducing cross-country inequalities in healthcare outcomes in the EU. To take just one example, consider the EU's current framework for national Roma integration strategies. Furthermore, increased healthcare expenditure and investment both in healthcare resources and the creation of opportunities for medical professionals in less developed European Union member states is of great importance in order to discourage the exodus of medical graduates from these nations that has been observed in recent years. Regarding policymaking to combat health inequalities in particular, Data Envelopment Analysis would certainly be useful in allowing us to determine which countries are producing at the production frontier. From this, we may deduce which countries might benefit most from policies to improve the efficiency of their healthcare production, i.e., increased innovation and use of technological developments. Going forward, it is hoped that this will be an area further research might choose to focus on.

5 References

Ardent, J., 2005. Does education cause better health? A panel data analysis using school reform for identification, Economic of Education Review 22, 149-160.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., et al., 2016. The association between income and life expectancy in the United States, 2001–2014, Journal of the American Medical Association, 315(16), 1750–1766.

Cutler, D. and Lleras-Muney, A, 2006. Education and Health: Evaluating Theories and Evidence, NBER Working Paper Series No. 12352.

Cylus, J., Papanicolas, I. and Smith, P.C., 2015. Using Data Envelopment Analysis to Address the Challenges of Comparing Health System Efficiency https://doi-org.ezp.lib.cam.ac.uk/10.1111/1758-5899.12212.

Drukker, D, 2003. Testing for Serial Correlation in Linear Panel-Data Models, The Stata Journal, 3, Nr 2: 168-177.

Fayissa, B., Danyal, S. and Butler, J.S.,2011. The Impact of Education on Health Status: Evidence from Longitudinal Survey Data, Department of Economics and Finance Working Paper Series.

Joumard, I., Andre, C. and Chatal, O., 2008 Health Status Determinants: Lifestyle,

Environment, Health Care Resources and Efficiency, Economics Department Working Papers No. 627.

Kiuila, O. and Mieszkowski, P.,2007. The Effects of Income, Education and Age on Health, Health Economics, 16: 781-798.

Mariani, F., Narahona, P. and Raffin, N., 2008. Life Expectancy and the Environment CES Working Papers, No. 48: 1-23

Mankiw, N.G., Romer, D. and Weil, D.N., 1992. A Contribution to the Empirics of Economic Growth, Quarterly Journal of Economics 107(2): 407-37.

Monsef, A., & Mehrjardi, A., 2015. Determinants of Life Expectancy: A Panel Data Approach. Asian Economic and Financial Review, 5(11), 1251–1257.

OECD, 2019. Health at a Glance 2019 – OECD Indicators, OECD Publishing, Paris. https://doi-org.ezp.lib.cam.ac.uk/10.1787/4dd50c09-en

OECD/EU, 2018. Health at a Glance: Europe 2018: State of Health in the EU Cycle, OECD Publishing, Paris. https://doi.org/10.1787/health_glance_eur-2018-en

Or, Z., 2000. Determinants of health outcomes in industrialised countries: A pooled, cross-country time series analysis, OECD

Economic Studies, no. 30, 2000/I.

Or, Z., W ang, J., & Jamison, D., 2005. International differences in the impact of doctors on health: A multi- level analysis of OECD countries., Journal of Health Economics, 24(3), 531–560.

Poças, A., & Soukiazis, E., 2010. Health status determinants in the OECD countries. A panel data approach with endogenous regressors, Estudos do GEMF, 4, FEUC. Grupo de Estudos Monetários e Financeiros. http://hdl.handle.net/10316/13325

Poças, A., Soukiazis, E. and Antunes, M., 2020. Factors Explaining Life Expectancy at Age 65: A Panel Data Approach Applied to European Union Countries, Social Indicators Research, https://doi.org/10.1007/s11205-020-02290-2

Schaffer, M.E., 2007. xtivreg2: Stata module to perform extended IV/2SLS, GMM and Ac/HAC, LIML and k-class r egression for panel data models.

Shaw, J.W., Horrace, W.C., and Vogel, R., 2005. The Determinants of Life Expectancy: An Analysis of OECD Health Data, Southern Economic Journal, 71(4) pp. 668-783.

Sillies, M., 2009. The causal effect of education on health: Evidence from the United Kingdom, Economic of Education Review, 28, 122-128.

Stock, J.J. and Watson, M.W., 2020. Introduction to Econometrics, 4th ed. Harlow-Pearson.

Spinks, J. and Hollingsworth, B., 2005. Health Production and the Socioeconomic Determinants of Health in OECD Countries: The Use of Efficiency models, Centre for Health Economics, Working Paper 151.

Wooldridge, J.M., 2 015. Introductory Econometrics: A Modern Approach, 7th ed. South-Western.

Yavari, K. and Mehrnoosh, M., 2006. Determinants of Life Expectancy: A Cross-Country Analysis, Iranian Economic Review, 11(15): 1 3-142

Variables Impacting Water Demand for Residents of the New York City Housing Authority

Cameron Somerville

University of British Columbia

Abstract

I examine the demand patterns of New York City's water system among individuals living in New York City Housing Authority developments. These are low-income households throughout New York City's five boroughs. Given the financial costs of the water delivery system to residents in New York and the pollution caused by this system, there are social, financial, and environmental incentives to examine the demand that drives it. I, therefore, look at whether building level variables as a whole have a statistically significant impact on water consumption, and which specific variables are significant. Using data from New York Open Data, I conduct a four-layered regression, layering variables by relevant category, to determine whether there is a relationship. At the 1% level I find four variables with a robust and significant relationship: a Seniors home building will use more water; a building with greater density uses more water per person; increased distance from the building to public infrastructure leads to more water consumption; the borough in which the building is located has significant results in differing directions, depending on the borough. From these results I am able to determine that there are significant building level variables, and I find areas for further research to elaborate on the results.

Acknowledgment

I want to thank the individuals who made this research possible. Dr. Jonathan Graves, my research supervisor at the University of British Columbia, provided the necessary guidance for me to create something meaningful. Alessandro Cattaneo provided me with numerous suggestions (such as looking at water tanks) and allowed me to bounce all my ideas off of him. Al Sweigart taught me how to use Python through his free online materials.

Table of Contents

Table of Contents

1 Introduction To Topic

New York City (NYC) is one of the most well-known and impressive cities in the world. With well over 9 million inhabitants, the city is densely populated and, therefore, has tremendous demands on its infrastructure (City Planning, 2013). Perhaps its most important infrastructure, as without it all inhabitants would perish, is the system to deliver clean drinking water to every household. Without its state of the art water supply system New York City as it is today would not be possible (NYT, 2014).

New York has a unique water supply amongst North American Cities for many reasons. Its current system is over a century old, yet it remains an engineering marvel. It relies solely on the power of gravity to get water from the Catskills and Delaware reservoirs nearly 150 kilometres down to the city of New York (Rueb, 2016). This means every building under six stories tall will have water naturally flow up into their faucets without a pump (Runyeon and Schwarzer, 2014). Taller buildings generally use water towers (Runyeon and Schwarzer, 2014). The system is also renowned for its cleanliness. New York is one of only five US cities to have a mostly unfiltered water supply (Hu, 2018). About 90% of NYC's water is unfiltered (Hu, 2018). This feat was considered so impressive that many engineers fought for the acclamation of having been the inventor (Koeppel, 2001, p.285).

This system, while impressive, is also expensive. The city's US$1 billion 2018 pledge to protect "the nation's largest municipal water system" demonstrates the cost borne by the city (Hu, 2018; Hill, 2018). Prior to that investment, the city spent over US$1.7 billion since the 1990s to protect its unfiltered water supply (Hu, 2018). Both these investments would be a drop in the bucket, however, if the worst case scenario occurs: rising demand outpaces what the unfiltered reservoirs can provide, and climate change leads to increased turbidity in water, making filtration necessary. If New York had to rely on more filtered water it would need a new filtration plant at an estimated cost of US$10 billion up front, and US$100 million per year thereafter (Hu, 2018). This would be the single largest capital project ever in New York's history (Hu, 2018). The city was already forced to spend US$3.2 billion on a new filtration plant in 2015, thanks in part to rising demand (Hu, 2018). With all the exorbitant costs of this system there is a financial incentive to examine the demand patterns for New York City's water.

The costs are not purely financial, either. The treatment and delivery of water emits considerable carbon emissions, which contribute to climate change. In the United States, about 19% of the energy delivered to households is used to heat water, and over 8% of all residential sector emissions come from laundry alone (Ro, 2020; Golden et al., 2010). We can use water consumption as a metric to understand the emissions being created; the more water that is being used, the more emissions one produces (Ro, 2020). About 80% of the emissions produced by domestic water use are based upon an individual's decisions (Ro, 2020). A change in the water-use behaviour of American consumers could have an impact equivalent to removing 12.1% of vehicles from US roads or closing 23 of the United States' coal power plants (Golden et al., 2010; Pasion, Oyenuga, and Gouin, 2017). In New York State specifically, water treatment accounts for a significant portion of emissions. About 35% of municipal energy use in New York State goes toward water treatment—about 3 billion kilowatt hours per year (DeIorio, 2008). This, based on US energy averages, is equivalent to 1,067,173 passenger vehicles driven for a year or 570,000 homes' energy use for a year (EIA, 2020; EPA, 2018). These emissions are significant. For the world to avoid multiple degrees Celsius of warming and significant environmental damage, they must be tapered quickly (Commission, 2019).

In contrast, these emissions could continue to grow. New York's population is projected to continue its steady climb (City Planning, 2013). Since every person requires water, it is logical that without any external change, demand for water will continue to rise in line with the population. Therefore, given the financial, social, and environmental risks, it is necessary that we examine what factors are impacting the demand of this system. I propose to do so by looking at those who will be most impacted by future costs.

Ultimately, as demand rises for New York's water and emissions become more expensive[1], it follows that prices should rise, since the supply is fixed. An increase in the price of water would impact low income households disproportionately. In New York City, these households are concentrated within the New York City Housing Authority (See Section 3.1). It is these individuals whose demand is the most elastic; they are more likely to change consumption habits due to a change in price. The water consumption habits of these individuals, therefore, is of vital importance to determine what will happen to New York's water supply in the coming decades. I aim to see if building level variables currently impact the water demand of these individuals and how large these impacts are.

I find that household and building level variables do have a statistically significant impact on water consumption per capita. Furthermore, I find four specific variables that have this significant and robust relationship with consumption. At the 1% level, the following are significant: whether the building is a seniors home, density of the building, proximity to local public infrastructure, and the borough in which the development is located.

The paper is laid out in the following order: Section

1 Emissions have increasing marginal damages. As greenhouse gases in the atmosphere accumulate, adding additional emissions will have an even greater negative impact, therefore, being more costly (Gillingham 2019; Forrest 2019).

2 details the research question, Section 3 contains a LiteratureReview and background information, Section 4 details the data sources and alterations, Section 5 describes my model, Section 6 outlines the results, Section 7 is a discussion surrounding the results, Section 8 is my conclusion, Section 9 contains additional figures, and Sections 10+ are appendices.

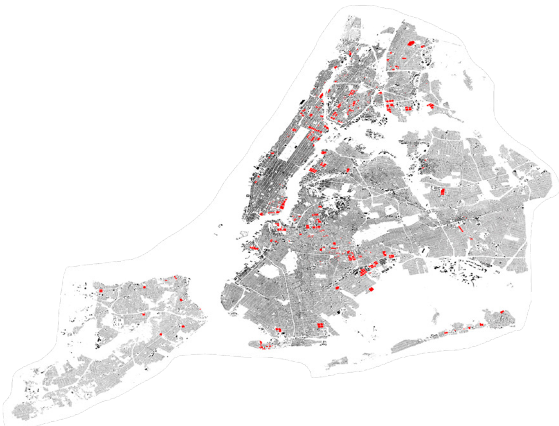
2 Research Question

I want to examine the impact of various economic indicators on water consumption in New York City for residents of the New York City Housing Authority. I aim to determine whether building level variables have a statistically significant effect on water consumption, and, if so, which variables those are. I do this through looking at the impact of development level variables (building age, apartment square footage, percent on fixed income, etc.) of the buildings to discover what causes different NYCHA buildings to consume different amounts of water. Analysis is done through a layered regression model.

3 Literature Review and Background Information

3.1 New York City Housing Authority

The NYCHA focuses on providing housing for low and medium income families. The authority provides homes for over 400,000 people in 326 different public buildings, and subsidizes rent for another 235,000 residents who are in private housing developments throughout New York City's five Boroughs (NYCHA, 2020a; Data, 2019a, 2019b). The NYCHA is North America's largest public housing authority; it has survived while authorities in other cities were demolished (Ferré Sadurní, 2018). This hasn't been due to building quality; the NYCHA currently faces US$25 Billion in needed capital and recently admitted to lying to federal building inspectors (Ferré Sadurní, 2018). A contributing factor to these rising costs and issues is building age. Most of the buildings were built at the same time, so are depreciating together (The Editorial Board, 2019). Mold, rats, and general uncleanliness are prevalent (The Editorial Board, 2019). Multiple buildings have had issues with boilers, meaning they've gone days with no hot water (Ferré Sadurní, 2018). Others have had lead contamination concerns (Ferré Sadurní, 2019a).



In an attempt to modernize the buildings and reduce costs the buildings are changing; many are switching to have their day-to-day operations carried out by privately owned companies (Ferré Sadurní, 2018). The notoriously mismanaged board has been undergoing a personnel change (Ferré Sadurní, 2019c). These changes have led to some units receiving upgraded appliances in their kitchens, impacting water consumption. Upgrades have stretched beyond single units; many repairs have been focused on improving infrastructure within NYCHA buildings (Ferré Sadurní, 2019c, 2019d). Replacing leaking pipes, for example, affects water consumption in the building as a whole. Buildings that have had these renovations are referred to as "rehab" buildings, as they have had their infrastructure rehabilitated (Chen, 2020). All rehabilitated buildings are required to use "ENERGY STAR" certified appliances, including dishwashers and washing machines, if appliances are replaced (NYCHA, 2017). This would reduce the water consumption as a clothes washer that is ENERGY STAR Certified uses about 33% less water than a standard one, and an ENERGY STAR Certified dishwasher saves tens of thousands of litres over its lifetime (Star, 2020b, 2020a). These rehabilitations have been slow, however, and the housing authority remains in a financially precarious situation, which could alter the behaviour of residents (The Editorial Board, 2019).

Rehabilitated buildings aren't the only unique sub-category of NYCHA developments. In the late 1970s and early 1980s the Federal Housing Authority of the United States handed administration of over 700 buildings to the NYCHA (NYCHA, 2020c). These buildings had been repossessed by the Federal Housing Authority from their owners. The NYCHA converted these homes to public housing and has worked, over the last 35 years, to allow the tenants to become the owners of the properties (NYCHA, 2020c). This has been successful for over 300 residents. Currently, the NYCHA is looking to continue this program and to rehabilitate the 16 units that are vacant.

Housing under the NYCHA is municipal public housing. They also subsidize rent for many individuals who are under federal "Section 8" housing (NYCHA, 2020b). The authority is transitioning some buildings toward an increased number of section 8 renters through having private management take over (Gross, 2018). With section 8 housing, rent is subsidized to be no more than 30% of the tenant's income (Ferré Sadurní, 2018). Any remaining rent beyond that is covered by the NYCHA (NYCHA, 2020b). For one to be eligible for this they must allow housing quality inspections, repairs, and they must obtain the annual certification (NYCHA, 2020b). For both public housing and Section 8 housing it is illegal to sublet these homes, so legally the tenant on the lease should be the one who is consuming the water in that unit (Ferré Sadurní, 2019b). Furthermore, one has incentive to not break the rules as there are approximately 177,000 people on the waiting list for a NYCHA apartment.

For NYCHA's public buildings there are two methods of management. Most buildings use the "conventional" method. For a conventionally managed building, the NYCHA acquires the land and then gives contracts for building maintenance and construction (Authority, 2019a). The other method of management is the "Turnkey" method, in which a private developer purchases the land, constructs a building, and then sells that building to the NYCHA (Authority, 2019a). Additionally, there can be sub-managed buildings within a NYCHA development. There are nonresidential buildings present at multiple NYCHA sites. Usually, they are child care centres (Veiga, 2019).

3.2 *Research on Water Consumption Per Capita*

There has been significant research looking at the most effective means of having one reduce their water use. Reducing water use has been argued to be desirable both due to projected increases in global water demand and the environmental impact of water use discussed earlier. Raymond and Streeter (2013) believe demand will continue to rise as people move from a grain based diet to a protein based one and as the water supply diminishes due to oil and gas extraction (especially fracking). Climate change will cause some cities to have increased access to water, while others have reduced access, and due to water's weight it is difficult to transport from one city to another (Raymond and Streeter, 2013). Therefore, the use of water is something cities must take more seriously. Keohane and Olmstead (2016) argue that pricing strategies are most effective in regulating water use and should be implemented instead of a cap & trade system. New York's water meters make dynamic pricing possible. Olmstead and Mansur (2012) undertook a similar approach, attempting to determine the price elasticity of demand for water in various US cities.

Neither of these projects examined what factors beyond price may have a large influence on consumption behaviour.

Researchers elsewhere have explored variables beyond price, examining the impact of non-economic individual level variables on water consumption. Jorgensen et al. (2013) look at household water consumption in Australian provinces. They determine that individual level variables are significant in creating predictive models for individual or household water consumption. Furthermore, they find that some variables have little impact on household consumption while maintaining an impact on individual consumption. This only occurred, however, in their sample in Southern Australia, indicating that results could vary region to region (Jorgensen et al., 2013). New York is, obviously, a different region than their Australian samples. They find that perception of neighbours' water consumption habits and household size have the greatest impacts on water consumption (Jorgensen et al., 2013). This has implications for New York, in which household size has a much smaller range than rural and urban regions of Australia. Furthermore, their research is focused on individual psychological indicators (attitudes toward water conservation), and only a couple economic indicators (age and income) (Jorgensen et al., 2013). I aim to further examine the impact of economic indicators rather than psychological ones. Lastly, and most importantly, is the researcher's call for further work to be done on multi-occupancy households as most of their sample was single-occupancy (Jorgensen et al., 2013). I will be looking at building level variables, which are all multi-occupancy.

Psychological researchers are not the only ones creating predictive water-use models. Machine learning and neural networks have also been used. These models generally focus on the factors external to the individual in an attempt to determine the water use demands for an entire system (Lee and Derrible, 2020; Firat, Turan, and Yurdusev, 2010). I, instead, want to look at the factors specific to each individual and building to determine what affects water consumption at a more granular level.

3.3 New York's Water Consumption

In 2001, Koeppel wrote a comprehensive history of New York's water system, detailing aspects from before the supply of New York's water by the Croton Water Company. Koeppel highlights the initial use of the water system as a public service, primarily for running fountains and fire hydrants throughout the city (Koeppel, 2001). The motivation for a public water system was largely to ensure better management of fires in the city (NYT, 2014). In the late 1800's, people shifted toward increased in-home water use, thanks to the installation of bathrooms, and the resulting demand for water quickly exceeded the 285 million Litre limit of the Croton aqueduct (Koeppel, 287). Thus, a second Croton aqueduct was built, but this too was soon outpaced by population when New York expanded to encompass the five boroughs that exist today (Koeppel, 289). The current Catskill aqueduct was built from 1907 until 1926 and can deliver 2,100 Litres of water per day. In 1965, the Delaware aqueduct was added to allow for 5,100 Litres maximum between the two reservoirs. These two reservoirs supply 90% of the city's daily water, with the other 10% coming from the old Croton watershed. The Catskill aqueduct provides unfiltered water supply while the Delaware's water is filtered (Hu, 2018). Koeppel's research demonstrates that getting water to the city has been a long and arduous undertaking (building the Delaware aqueduct alone took 28 years). His detailed research of the water supply also shows the need for greater research on the demand of water in New York (Koeppel, 2001; City Planning, 2013).

Currently, there are multiple projects focused on the intricacies of the New York water system. In fact, the whole state is unique. New York uses little water for agriculture: only ~1% as opposed to the national average of ~70% (nys, 2019; Raymond and Streeter, 2013). Fully 25% of New York State's water use is for the "public water supply,"(nys, 2019). The city's use of water is unique as well. Residents of the Greater Vancouver Area, for example, use upwards of 40% of their water outdoors (primarily for gardening) (Schreier, 2020). New York's denser living conditions mean less garden space per capita, greatly reducing this as a factor in water usage (Schreier, 2020). Furthermore, as highlighted earlier, New York's water is unfiltered. Disinfection, therefore, is paramount. New York is home to the world's largest ultraviolet disinfection facility, where harmful microorganisms are killed and a chemical mixture is added to stop the water from corroding the city's pipes (Cohen, 2016). The prevention of corrosion is vital to slowing the seepage of lead pipes and fixtures into the water. A variety of quality tests are then conducted both through automated means and by human scientists travelling throughout the city to collect samples (Rueb, 2016; NYC, 2019). Testing is done for a variety of reasons, but lead contamination tests are of the utmost importance.

Lead contamination in drinking water has had regular air time on North American news stations recently. Multiple Canadian cities, (CBC, 2019) indigenous reserves (Barrera, 2019) and suburban communities like Flint, Michigan (Nelson, 2016) have tested positive for dangerously high lead levels. Lead is removed during treatment of water but can break off from water mains that connect houses to the grid (Griggs, 2019). The older a water main is, the more likely it was built using lead (Rueb, 2016; Nelson, 2016). New York, with its many old lead pipes, has some residences with dangerously high lead levels in their water (Griggs, 2019). US federal regulations permit 10% of buildings in a city to have high lead levels (Lowenstein, 2018). Unfortunately, New York, due to its size, could therefore have an exorbitant number of households exposed to heightened lead levels (Lowenstein, 2018). The high publicity could raise people's awareness of how old the exterior pipes are and could affect how much water individuals use. Furthermore, this could have an impact specifically on residents of the NYCHA. As I detailed in Section 3.1, the NYCHA has many aging buildings, some of which have tested positive for dangerous lead levels. The age of each NYCHA building is therefore important.

While the city's tests for lead are seen as reliable, the same cannot be said of their tests on water tanks. Water tanks in New York create a whole different issue in terms of cleanliness of drinking water. These tanks are the primary drinking source for buildings that are over six stories tall. Runyeon and Schwarzer looked in depth at New York City's water tanks. They found there are between 12 and 17 thousand tanks in New York City. When the *New York Times* tested 14 randomly selected tanks throughout the city they found eight had coliform bacteria and five had E. coli (Runyeon and Schwarzer, 2014). Coliform on its own is not dangerous, but its presence means the water is ripe for bacteria (Rivera, Runyeon, and Buettner, 2014). Presence of E. Coli, however, means the water is not fit for human consumption (Rivera et al., 2014). These results are based on far too small of a sample size for any large conclusions to be drawn, but they proved and publicized that there are some tanks in New York with E. Coli. This is due largely to the tanks being in a state of disrepair and having openings in their tops (Runyeon and Schwarzer, 2014). New York City currently requires that tanks be drained and cleaned once per year, to prevent growth of bacteria and algae, but it is estimated that nearly 60% of buildings do not comply (Runyeon and Schwarzer, 2014). Furthermore, tanks with routine maintenance can still possess E. Coli, and at least two of the three largest NYC water tank maintenance companies use "Sea Goin' Poxy Putty," which is in violation of the city's health code (Rivera et al., 2014). If the research by Runyeon and Schwarzer is indicative of a larger issue, then individuals whose buildings use water tanks may alter their consumption habits. Slightly over 21% of NYCHA buildings whose water is tracked are over 6 stories tall (Authority, 2019a, 2019b).

Multiple sources have reported on the financial details of New York City's water system. As detailed in the introduction the system is exceptionally expensive and could quickly become even more expensive. Notably, the investment to build a third major water main under the city is currently one of New York's largest ever capital projects (Cohen, 2016). This pipe is needed as it would allow for repairs on the other two pipes, which were built in 1917 and 1936, without disrupting the distribution of clean water to New York City (Hu, 2018). One of those old pipes is currently leaking well over 30 million litres of water annually (Hill, 2018). The system is exceptionally expensive, making an examination of the demand for its output valuable.

4 Data

I regress a variety of explanatory variables with my dependent variable being water consumption. I run four regressions, each adding another layer of variables. My initial layer is explanatory variables that are financial in nature. Next, location based variables are added, then billing variables, then building management variables. Through this layered regression strategy, I can see what explanatory variables have an impact on consumption.

My dependent variable is daily water consumption per person in NYCHA buildings. This is, therefore, a vitally important variable in my study. Through New York City's "Open Data" initiative the necessary consumption data for each building are available publicly (Authority, 2019b; of Sustainability, 2019). For this data to be made per capita I have once more relied on New York Open Data, using the NYCHA Data Book figures for this and many control variables (Authority, 2019a). These data should be accurate as it is illegal to sublet one's unit, as described in Section 3.1. To get location variables I used the GIS data from the same source (Data, 2019c, 2019a). Summary statistics for all the variables used are included in Appendix 1 and the summary descriptions for variables of interest are included in Section 4.5. The code for getting those summary statistics is in Appendix 6.

4.1 *NYCHA Water Consumption Data Modifications*

This data includes observations on a subsection of NYCHA buildings and their water consumption. The data include just over 31% of NYCHA buildings. The data span over multiple years and billing periods. First, using Python, I parsed the string-type revenue month variable and created two new variables from it: Revenue Year, and Revenue Month (See Appendix 2). I then cleaned up this data (Appendix 5) to code the remaining string variables and eliminate unneeded ones. Additionally, I created two dummy variables. One shows whether the building is a rehabilitated building and the other shows whether the building was previously repossessed by the FHA. Both of these are as described in Section 3.1. The latter was then used in the Borough variable, as all former FHA buildings have no borough association. This is a comma-separated-values data-set with over 32.5 thousand observations.

4.2 *NYCHA Development Data Book Modifications*

The New York City Housing Authority publishes observations on a variety of variables for almost all their buildings through their "Development Data Book". The data are updated annually and include only the most recent observations. This data-set matches with the NYCHA Water Consumption set using the Tenant Data System numbers (TDS#). This made merging the two data-sets together possible with very few lost observations from the water consumption data set. A few observations were dropped, as they are not tracked in the Data Book. Some work was needed to clean this data and make it usable. Using Python, I created a variable that represents the age of each building in number of days (Appendix 3). As highlighted in Section 3.3, the age of a building is a strong indicator of the likelihood of lead contamination. Age of the building is used therefore as a measure of whether people are concerned about lead contamination in their water. With the data set, I ran into some issues. The methods used for data-entry meant there were multiple string characters within integers, making de-stringing on Stata impossible. I used Python to parse these observations and allow for all desired values to be stored and used as integers (Appendix 4). Basic data cleaning work was then carried out with Stata (Appendix 5). This set only reports once for each building in the NYCHA, not for each billing cycle, so has 323 observations. This data-set is also a comma-separated-values file type.

4.3 Distance from Subway

Using the wealth of GIS data on New City Open Data, I was able to determine the distance from each building to the nearest subway station. This is used as an indicator of the infrastructure in the neighbourhood surrounding the building, rather than the infrastructure in that building. To accomplish this, I used a base map of New York's boroughs (New York City Government, 2020), then added all NYCHA buildings (New York City Housing Authority "Map of NYCHA Developments", 2019a), then added all subway lines and stations (Metropolitan Transport Authority, 2018; Metropolitan Transport Authority, 2019). I used QGIS with the NNJoin Plugin to interpret the map (Arken, 2019). The resulting variable is NEARESTSUBWAY, which measures the distance from each NYCHA building to the nearest subway station in decimal degrees. The variable is used in the Locations layer of the regression.

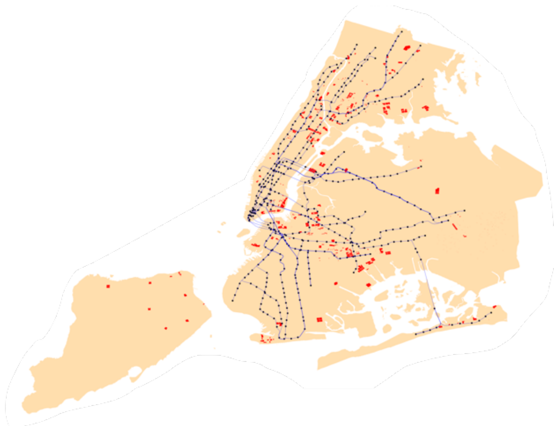


Figure 2: All NYCHA domestic units marked in red with all subway lines and stations.

4.4 *Dropped Observations*

Few observations had such inadequate data that I was forced to drop them. I dropped observations for the buildings in the Data Book that are not in the water consumption data, a little under 70% of the data. Likewise from the water consumption data there are seven buildings that are not tracked in the data book. In the water consumption data there are some observations that are "estimated" rather than actually recorded. The water authority makes an educated guess on the consumption based on the building and season. These observations were all dropped to ensure data accurately reflects reality. Likewise, all bills are analyzed by the New York City Water Board. Any bills that have issues, or illogical data are marked as "exceptions". All of the observations with exceptions on the bill were dropped. That leaves 24,776 unique observations from 94 different NYCHA buildings.

4.5 *Key Variables*

There are many variables in the data set that, based on the issues highlighted in the background section, play a key role in the regressions. The number of stories for buildings determines whether or not they have to use a water tank to store and access their water. Buildings over six stories tall will either need pumps or water tanks, as the gravitational pull of the New York water system reaches its limit just above six stories in most of New York. As highlighted in Section 3.3, water tanks are known for a lack of cleanliness and have been reported to reduce municipal water consumption as some attempt to find clean water elsewhere. To account for this, I created a dummy variable, WATERTANK, that is equal to 1 when a building is taller than six stories and 0 if it is equal to or less than six stories. In other words, it is equal to one when a water tank would be needed. This variable is included in the Building layer (second layer) of the regression.

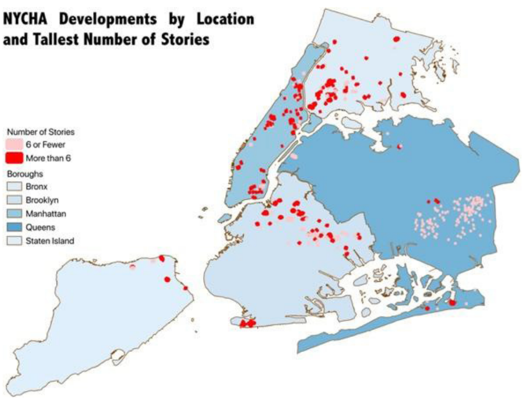
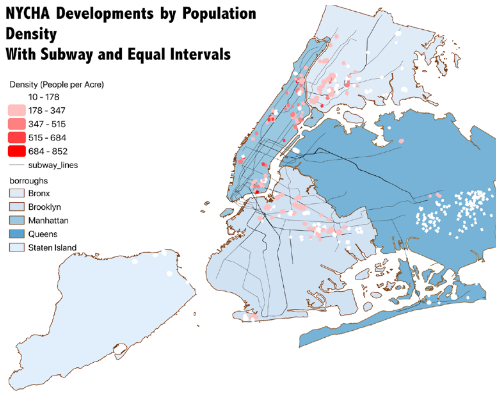


Figure 3: 29% of NYCHA BLDGS. are >6 stories

As highlighted in Section 3.2, research in Australia on the impact of individual level variables found that the square footage of one's dwelling had a large impact on water consumption. Those in houses with more space per person generally use more water. To measure this, I use the density variable, which is the number of people per

Figure 4: Densest BLDGS are in Manhattan acres of buildings in NYCHA developments.



More people per acre of building means each individual would have fewer square feet to themselves. As can be seen in Figure 4, the buildings where population density are the greatest appear to be concentrated in Manhattan. The data back this up, Manhattan's mean density is over 250 people per acre, whereas the Bronx (next closest) has just under 200.

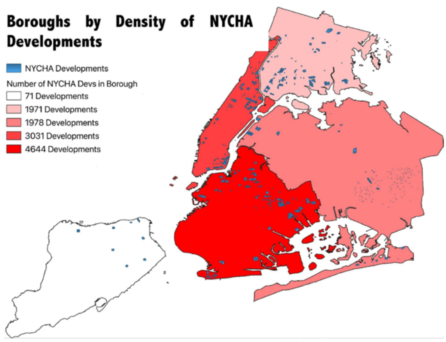


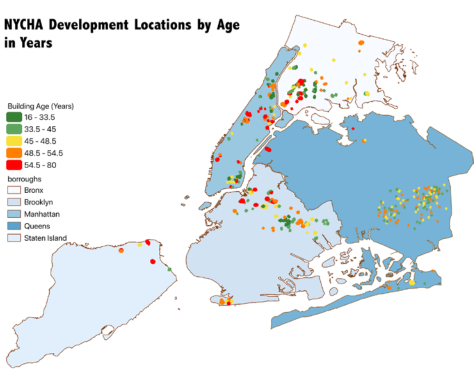
Figure 5: Brooklyn and Manhattan have the most NYCHA buildings

Also highlighted in the findings from Australia was variation in results from different locations. It follows therefore that the borough in which the NYCHA building is located could be of considerable significance. There is variation in the concentration of NYCHA buildings in different Boroughs, as shown in the adjacent figure. The significance of the Borough will be tracked in the location layer of the regression using a qualitative variable that has a category for each of New York's five Boroughs, and a sixth category for FHA buildings (as described in Section 3.1), which do not have a Borough in the data-sets.

As discussed in section 3, there are implications to living in an older building. The older a building is, the older may be the pipes. The creation of the DAYSOLD variable, as described above, accounts for this. When used in the regression, it demonstrates the effect that an individual living in an older building and, therefore, having concerns of lead contamination in their water might have on consumption.

As highlighted in Section 3.1, the NYCHA subsidizes rent for individuals so one will never pay more than 40% of their income in rent in public housing, and 30% in section 8 housing. Included in the regressions is a variable for the average gross monthly income for the NYCHA from rent payments. This shows the relative wealth in different buildings, as wealth is a possible tributed determinant of consumption.

Figure 6: Older buildings appear evenly distributed



There are multiple important control variables. Data from UBC professor Hans Schrier show the importance of seasonal variation on water consumption, this makes the month in which the bill was recorded an important control variable (Schreier, 2020). With the variable marking the month of the bill, the effects of seasonal variation will not cause issues in the results. Similarly, one particularly hot summer will not throw off results, as the year of the water bill is also recorded. I also control for changes at the building level. As described in Section 3.1, there are buildings attached to NYCHA developments that house services for residents, such as daycares. Developments with daycares likely use more water than those without. To control for this a variable is included to account for the different number of non-residential buildings. Additionally, I control for the quality of the building in which residents are living. Since rent payments will not necessarily reflect the building's quality, I control for this using the development cost of the building divided by the number of apartments in the building. As can be seen in the two figures, rental payments and building development cost per room do not line up perfectly.

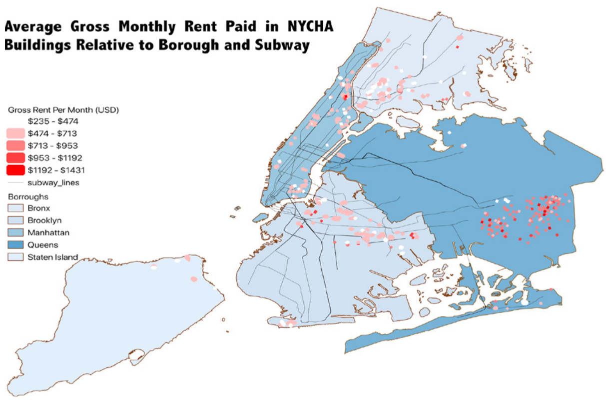


Figure 7: Rent is capped at 40% of income

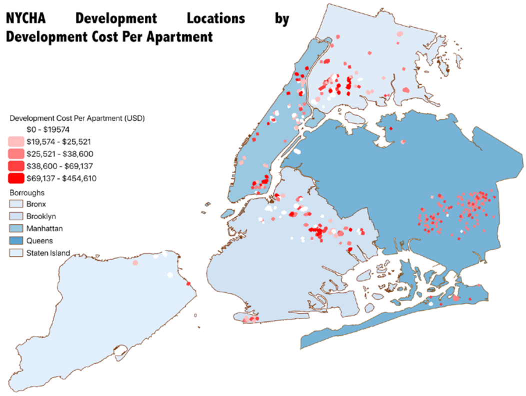


Figure 8: Doesn't align with gross monthly rent

There are further important controls. I control for whether a building has been rehabilitated, as described in Section 3.1. This could impact water consumption by ensuring upgrades to more efficient appliances. Different NYCHA buildings use different management methods and funding sources, outlined in Section 3.1, which could alter behaviour of residents. A variable tracking the funding and a separate variable tracking the management method are used to control for these two points.

The regression overall consists of 19 different variables, all of which play an important role. There are summary statistics in the appendix for all variables as well as additional figures in the appendix.

5 Model

I use four regression equations to see the impact of various independent variables on water consumption in the New York City Housing Authority buildings. With the layered regressions, I can see what is significant with few control variables. Then, by gradually layering on more control variables, I can observe the impact on the explanatory variable. This process allows me to see what layer of control variables has a large impact.

Water consumption per capita per day is my dependent variable in the regression. The four regressions are layered on top of one another with variables sorted into four different groups: financial indicators, building management, location, and then billing specifics (in that order). All upper-case variables are ones that I have either created or edited the labeling and coding of significantly.

5.1 *Financial Indicator Regression*

The first regression will use the following equation and variables:

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon_i$$

- Y_i: Water consumption per capita per day in litres, the dependent variable (varname= CONSUMPTION)

- β_1: The size of the effect of the amount paid in monthly gross rent in that building

  -X_1: avgmonthlygrossrent: The average monthly gross rent, which is meaningful since the NYCHA doesn't charge more than 40% of an individual's income as rent, so it either shows individuals with a high enough income to pay the full cost of rent or shows 40% of their income.
- β_2: Shows the effect of the number of people in section 8 transition housing

  -X_2: PERCENTSECT8TRANS: The percent of the population that is in section 8 transition housing, waiting to be moved to a proper section 8 subsidized building or for the whole building to transition to section 8 housing.

5.2 *Regression Layer: Building Management*

The second regression includes the same variables as the first for betas one and two and for Y. Betas four through twelve are related to building management.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * i.X_5 + \beta_6 * i.X_6 + \beta_7 * X_7 + \beta_8 * X_8 + \beta_9 * X_9 + \beta_{10} * X_{10} + \beta_{11} * X_{11} + \epsilon_i$$

- β_3: Shows the size of the effect of a building having been rehabilitated
    - X_3: REHAB: A dummy variable that is equal to one if a building has been rehabilitated

- β_4: Shows the size of the effect of a one unit increase in the age of the building, in number of days
    - X_4: DAYSOLD: The number of days between March 10, 2020 and the building's construction being 95% completed

- β_5: The effect of the building being home solely to senior citizens
    - X_5: SENIORS (Qualitative): A coded variable that is equal to zero if the building is not a seniors home, one if the building is partially a seniors home, and two if the building is exclusively for seniors

- β_6: Shows the effects individually of each of the qualitative categories for the different funding types the building could receive
    - X_6: FUNDING (Qualitative): The source of funding for the building, four potential options

- β_7: The effect of the building's management method
    - X_7: METHOD: A dummy variable that is equal to one if the building uses the turnkey system and

zero if the building uses the conventional method (see Section 3.1).

- β_8: Will show the effects of having more people in a fixed-size building
  - X_8: density: The number of people per square acre in the building

- β_9: Will show the effects of having more nonresidential buildings in a development
  - X_9: numberofnonresidentialbldgs: The number of buildings that aren't residential, whether they are daycares, storage areas, or other facilities (more information in Section 3.1)

- β_{10}: Will show the effects of a more expensive development
  - X_{10}: COSTPERROOM: The cost to develop the property divided by the number of rental rooms

- β_{11}: Will show the effects of a building currently being managed privately
  - X_{11}: PRIVATE: A dummy variable equal to one if the building is currently being managed privately

- β_{12}: The coefficient for a building of a height where a water tank would be needed
  - X_{12}: WATERTANK: A dummy variable equal to one if the building is greater than six stories tall or zero if the building is less than or equal to six stories tall.

5.3 *Regression Layer: Location*

The third regression includes all the variables used in the previous two plus the variables relating to the location of the NYCHA development. Betas thirteen and fourteen are related to building location.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * i.X_5 + \beta_6 * i.X_6 + \beta_7 * X_7 + \beta_8 * X_8 + \beta_9 * X_9 + \beta_{10} * X_{10} + \beta_{11} * X_{11} + \beta_{12} * X_{12} + \beta_{13} * X_{13} + \beta_{14} * i.X_{14} + \epsilon_i$$

- β_{13}: Shows the size of the effect of distance from public infrastructure
  - X_{13}: NEARESTSUBWAY: Shows the distance, in decimal degrees, of the nearest subway station to the development

- β_{14}: The effect of the New York City borough in which the development is located
  - X_{14}: BOROUGH (Qualitative): A variable that shows which borough the development is located within and also uses the borough FHA if the building was formerly repossessed by the Federal Housing Authority

5.4 *Regression Layer: Billing*

The fourth and final regression once again includes all previous variables. This regression layers on variables relevant to the billing for the building's water consumption. Betas fifteen through twenty-two are related to the building's billing system.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * i.X_5 + \beta_6 * i.X_6 + \beta_7 * X_7 + \beta_8 * X_8 + \beta_9 * X_9 + \beta_{10} * X_{10} + \beta_{11} * X_{11} + \beta_{12} * X_{12} + \beta_{13} * i.X_{13} + \beta_{14} * X_{14} + \beta_{15} * X_{15} + \beta_{16} * X_{16} + \beta_{17} * X_{17} + \beta_{18} * i.X_{18} + \beta_{19} * X_{19} + \epsilon_i$$

- β_{15}: Shows the size of the effect from residents paying directly for their electricity

  - X_{15}: ELECTRICITY: A dummy that is equal to one if residents pay their electrical bill directly, zero if they do so indirectly (through their rent or fees). Note that fixed rental prices could effectively mean one doesn't pay for electricity when paying indirectly

- β_{16}: The effect of the month of year in which residents are using the water
  - X_{16}: REVENUEMONTH: The month in which the bill was recorded, on a numeric twelve month scale

- β_{17}: Shows the size of the effect of the year in which residents consumed the water
  - X_{17}: REVENUEYEAR: The year in which the bill was recorded, ranging from 2013 to 2019

- β_{18}: The effect of the type of rate users are being charged
  - X_{18}: RATECLASS (Qualitative): The rate, out of the five options, that the building is being charged

- β_{19}: The size of the effect of charges or credit being present on the buildings' account
  - X_{19}: othercharges: The charges that are present on the bill that are beyond charges for water consumption. Note this variable can be negative if the building has a credit on their account.

6 Results

The following are the results of the regressions in the model. I find five variables with a significant impact using a 99% confidence interval.

- Whether the building is exclusively a seniors home

- Proximity to the nearest subway station
- The area per person of the building

- Whether the building uses a water tank
- The Borough in which the development is located

Detailed analysis of these results is in the Discussion Section as well as data visualizations.

6.1 Brief Overview of Results

The initial two regressions show the impact of the control variables. Regression one shows both the average monthly rent and the percent of households that are section 8 transition housing have a significant impact on water consumption when holding the other variable constant. As highlighted in the background information, the management and upkeep of NYCHA buildings is complicated, so a variety of other variables could have influenced these results. Moreover, this complication causes the second regression, which controls for building level variables such as rehabilitation and the age of the building, to no longer show these variables as significant. On the other hand, the impact of a building being entirely made up of seniors is clearly significant, as is the building's density. Whether the building uses a water tank does not appear significant with only these two layers of the regression. It appears that this layer, building level variables, has a strong impact on the results.

The final two regressions introduce more control variables level variables to be of importance, as both of them remain significant.

	(1)		(2)	
	Financial		Building	
VARIABLES	Litres per day per person	se	Litres per day per person	se
Litres per day per person				
Avg Monthly Rent	-0.219***	(0.0258)	0.0586	(0.0422)
% Section 8 Transition	-480.3***	(90.38)	76.44	(526.1)
If Dev. is Rehabilitated			7.034	(27.41)
Development Age			-0.00305*	(0.00171)
Building is Partially Seniors			4.870	(73.66)
Building is Exclusively Seniors			291.1***	(33.87)
Funding Source is MHOP			19.89	(53.10)
Funding Source is MIXED/LLC1			-141.7	(98.00)
Funding Source is MIXED/LLC2			-30.50	(179.5)
Building Management Method			-10.45	(18.57)
Density			0.554***	(0.116)
# of Non-Residential BLDGS			-15.66	(15.19)
Development cost per rental room			0.00107*	(0.000577)
If Building is Managed Privately			7.345	(26.26)
Water Tank			-17.83	(17.84)
Constant	244.5***	(19.94)	19.29	(56.07)
Observations	24,772		24,684	
R^2	0.004		0.011	

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

	(3)		(4)	
	Location		Billing	
VARIABLES	Litres per day per person	se	Litres per day per person	se
Litres per day per person				
Avg Monthly Rent	0.0535	(0.0510)	0.0566	(0.0512)
% Section 8 Transition	4,847**	(2,004)	4,327**	(2,068)
If Dev. is Rehabilitated	33.06	(29.25)	32.99	(29.30)
Development Age	-0.00285	(0.00234)	-0.00383	(0.00254)
Building is Partially Seniors	-63.61	(76.41)	-53.23	(76.63)
Building is Exclusively Seniors	268.8***	(36.39)	270.7***	(36.58)
Funding Source is MHOP	-44.92	(69.55)	-32.08	(70.68)
Funding Source is MIXED/LLC1	-960.7***	(356.5)	-851.6**	(373.6)
Funding Source is MIXED/LLC2	-1,380**	(628.4)	-1,214*	(646.9)
Building Management Method	69.08*	(35.58)	86.58**	(39.52)
Density	0.388***	(0.147)	0.389***	(0.147)
# of Non-Residential BLDGS	-11.82	(17.76)	-12.69	(17.99)
Development cost per rental room	9.43e-05	(0.000634)	7.93e-05	(0.000634)
If Building is Managed Privately	3.424	(32.47)	2.939	(32.49)

Water Tank	-96.92***	(21.98)	-102.5***	(23.03)
Nearest Subway	3,475***	(1,266)	3,540***	(1,266)
Borough is BROOKLYN	-23.16	(27.74)	-23.65	(27.78)
Borough is FHA	30.20	(49.77)	15.33	(54.44)
Borough is MANHATTEN	167.6***	(33.09)	167.5***	(33.09)
Borough is QUEENS	142.3***	(41.09)	139.3***	(41.94)
Borough is STATEN ISLAND	-883.3**	(355.6)	-802.2**	(364.4)
If Electricity Billed Directly			-34.60	(40.35)
Month of Bill			1.367	(1.502)
Year of Bill			4.750*	(2.720)
Rate Class			34.36	(40.49)
OtherCharges			-0.00543	(0.0105)
Constant	-25.49	(75.19)	-9,582*	(5,486)

Observations 24,407

R^2 0.014

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

7 Discussion

The five variables that are significant once the final layer of the regression has been added are all significant at the one percent level. First, I will briefly discuss the robustness of the results as a whole, and then I will look at each of these five significant variables individually.

7.1 *Robustness*

Using months as a continuously distributed variable is misleading, as it is difficult to tell whether any particular month has an impact on water use. Since the bill month was used solely as a control variable, I initially conducted the regression in this way in order to see the results of each explanatory variable while holding the month constant. By doing this, we could be missing results within the billing month. I re-run the regressions using REVENUEMONTH as a qualitative rather than quantitative variable. In doing so, we can see that no one month is significant with a 99% confidence interval; only September has a significant impact with a 98% confidence interval.

Furthermore, I use robust standard errors to ensure the robustness of results. Using a Cook-Weisberg test for heteroskedasticity, I get Prob > chi2 = 0.0000. With this p-value, I can reject the null hypothesis of constant variance. This further means the results are heteroskedastic. However, robust standard errors can account for heteroskedasticity at the expense of precision, meaning there could be generally larger standard errors and less likelihood of a variable remaining significant. Running the regression with these standard errors shows that four of the five meaningful explanatory variables remain significant at the one percent level. Only "water tank" becomes insignificant.

The following regression table shows a regression run with all

	(1)		(2)	
	Month		Robust	
VARIABLES	*H_2O Consumption*	se	*H_2O Consumption*	se
Seniors Building	270.7***	(36.58)	270.7***	(23.40)
Density	0.389***	(0.147)	0.389***	(0.0539)
Water Tank	-102.7***	(23.03)	-102.5*	(57.17)
Nearest Subway	3,544***	(1,266)	3,540***	(1,085)
Borough is MANHATTEN	167.7***	(33.10)	167.5	(121.3)
Borough is QUEENS	139.3***	(41.94)	139.3**	(55.83)
Borough is STATEN ISLAND	-800.7**	(364.4)	-802.2***	(208.8)
Month of Bill is FEBRUARY	0.998	(24.63)		
Month of Bill is MARCH	-3.076	(24.81)		
Month of Bill is APRIL	-4.467	(24.58)		
Month of Bill is MAY	-3.635	(24.12)		
Month of Bill is JUNE	-2.564	(23.74)		
Month of Bill is JULY	0.998	(23.89)		
Month of Bill is AUGUST	2.503	(24.45)		
Month of Bill is SEPTEMBER	58.57**	(24.33)		
Month of Bill is OCTOBER	-2.822	(24.90)		
Month of Bill is NOVEMBER	-1.785	(26.11)		
Month of Bill is DECEMBER	2.832	(25.11)		
Constant	-9,498*	(5,495)	-9,582	(10,664)
Observations	24,407		24,407	
R2	0.015		0.014	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

previously cited variables, but I have excluded all insignificant ones for space. The full regression results can be found in the Regressions Appendix as the first table.

7.2 Water Tank

The non-robust regression shows that, with a 99% confidence interval, a building having a water tank leads to individuals using less water. Specifically, a building with a water tank would reduce daily water consumption by about 102.5 litres per person, which is a relatively enormous decrease. This is an imperfect measure, however, as the WATERTANK variable is a dummy that is only equal to one if the building is greater than six stories tall. Consequently, the building most likely has a water tank, or it would need some other system to ensure access to water in an emergency.

To cast further uncertainty on this result, it does not hold up with robust standard errors. Of the five significant variables, it is the only one that does not remain significant with the heteroskedasticity adjustment. Ultimately, this variable would be more meaningful if there were a definitive tracking of water tanks for each NYCHA building and if the water quality inspection results of each of those tanks were made public. With this information, I could isolate not only the effect of having a water tank, but the effect of having a water tank that has failed water quality inspections.

This result remains significant but has some glaring issues in terms of robustness. I would put the least stock in this explanatory variable, and I believe further data is necessary to provide a more definitive answer.

7.3 Public Infrastructure

The Nearest Subway variable was used as an indicator of a building's proximity to public infrastructure. I found that the closer one is to a subway station (i.e. they are in an area of increased public investment in infrastructure projects), the less water they use per day.

This result remained significant with robust standard errors. Overall, I found that if someone is one decimal degree further from a subway station, then they use 3,540 more litres of water per day. This amount of water seems illogically large because of the units of decimal degrees. Decimal degrees are indexed on a global scale, so when used at a city level, they are extremely small decimal numbers. Hence, there is a large standard error when increasing distance by one full decimal degree. Furthermore, decimal degrees are a three dimensional unit, so they cannot be converted to metres. The coefficient, therefore, can't tell us about the water consumption change in Litres per additional metre of distance, but it does identify the important relationship between distance and consumption. Using the subway stations as indicators for public infrastructure, we find that individuals in a more urban setting use less water, holding all else constant. This means that those in more suburban areas use greater quantities of water. This could have policy implications if New York City is forced to implement restrictions to curtail water use, since their policies could be aimed at those in less urban settings as they use more water. This result appears to hold up to robustness checks and as such is included as one of my key significant variables.

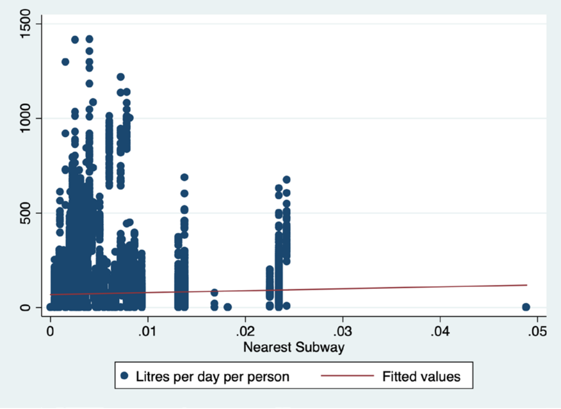


Figure 9: Concentrated at Near to Subway & Low Consumption

7.4 *Seniors*

A building being open exclusively to seniors has a statistically significant impact on the building's water consumption per capita per day. It causes a huge increase in daily consumption per person of almost 271 Litres, relative to a building that is in no capacity a senior centre. This result, when adjusted for the heteroskedasticity of the data, using robust standard errors, remains significant at the 1% level.

This means that senior homes use more water, but this could also mean that age is an important factor in water consumption per capita. If data for the median age of tenants within each building were available, then I would be able to see the effect of this variable. In its current form, this still could have policy implications. If New York City is aiming to implement a policy that reduces water consumption, then the incentives could be aligned with tenants of buildings exclusively open to seniors.

7.5 *Borough*

Some of the boroughs have strong relationships with water consumption. A building being located in Queens or Manhattan has a strong correlation with increased water use - about 139 and 168 additional litres per person per day, respectively. These are significant with a 99% confidence interval. There is a slightly less significant relationship, but significant nonetheless, for buildings in Staten Island. Buildings in Staten Island use significantly less water with a 97% confidence interval.

These results are substantial as they align with the findings of the Australian research, but on a different level. The research, cited in the literature review, found significant interprovincial differences in consumption in Australia. This furthers the notion of location playing a significant role in one's consumption, but on a more micro level. This shows regional variation can be significant even at a city level, not just a provincial one, furthering the existing research in the field.

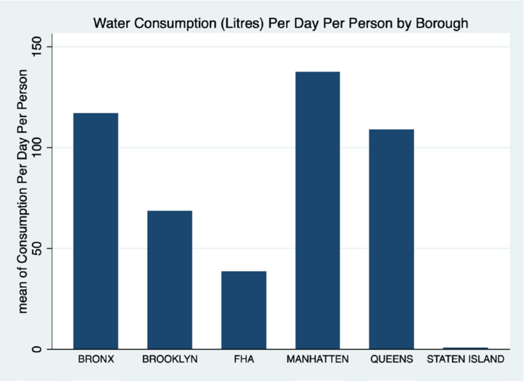


Figure 10: Staten Island uses little water

7.6 Density

The most important variable for an individual's water consumption, according to the previously detailed research, is the area of their home, per person. In other words, the population density of their house. One study's sample had largely single-person dwellings in rural areas, not urban family apartments.(Jorgensen et al., 2013) Their research found that increased area per person (reduced density) resulted in increased water consumption. My research shows the opposite, that reduced area per person (increased density) resulted in increased water consumption per person. I find one additional person per acre adds ~0.389 litres to everyone's daily water bill.

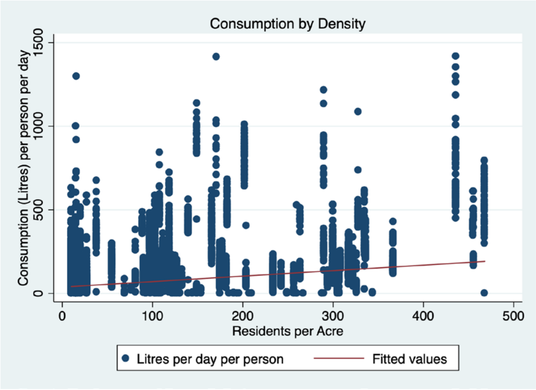


Figure 11: Slightly positive relationship density and consumption

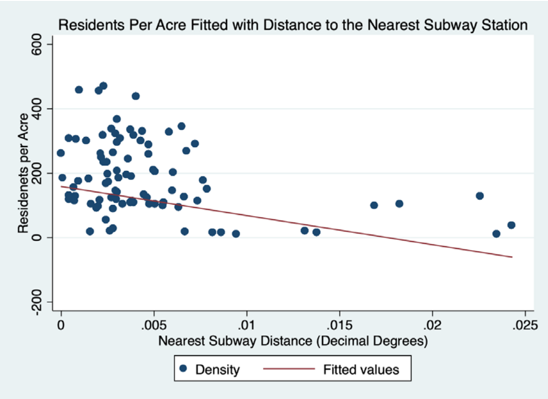


Figure 12: Slightly negative relationship density and subway distance

The conflicting results could be due to the differences in the makeup of our samples. Mine is focused on low-income individuals in a highly urban environment, while theirs is focused on the opposite. Or, as highlighted as a reason for this research, this could simply be regional variation. As there was variation within Australia on the impact of different explanatory variables it could be the case that in a different continent the results of certain variables are flipped. More research should take place on why this change occurs.

7.7 Omitted Variables

There are other variables that could be significant if included in this analysis. Mainly, more data on individuals would be helpful. These would allow me to see what the makeup of the residents in each building is. Common individual level variables such as gender, age, and race could be helpful to control for the effects of individual decision making. It's difficult to know what impact these variables would have on the results if I were to include them. Based on the impact of a building being a senior home, it is possible that the average age of a building's residents could have significant implications on the water consumption patterns. Having access to data with individual level variables would confirm this. In lieu of this data, I have attempted to control for these variables through wages (rent payments locked at 40% of income) and seniors. These variables are not perfect representations of those that have been omitted, but I believe they are the best available.

8 Conclusion

This research shows that building level variables do have a significant impact on water consumption, thus answering my initial question. Furthermore, it displays which variables have a significant and robust relationship with a building's water consumption per capita per day: Borough, Density, Senior Home, and Proximity to Public Infrastructure.

My results indicate a couple areas in which further research could be conducted, allowing us to see the impact of more specific variables and to be more certain of others. The impact of a building using water tanks remains unclear. This could be improved through more research using a dataset where each building's water tank status and cleanliness is recorded. This would allow for us to isolate the impact of a building using a water tank, holding all else constant. Another area of uncertainty within the results is the variation between NYCHA residents and Australian residents. Further research on why population density has opposing impacts on these two areas could prove useful. Lastly, and most importantly, controlling for common individual level variables would add increased robustness to the results, as we could see if the results hold regardless of gender balance, racial profile, or average age of tenants within the building. Specifically, based on the results from senior homes, research into the impact of resident age on water consumption could prove fascinating and meaningful.

9 Figures & Images

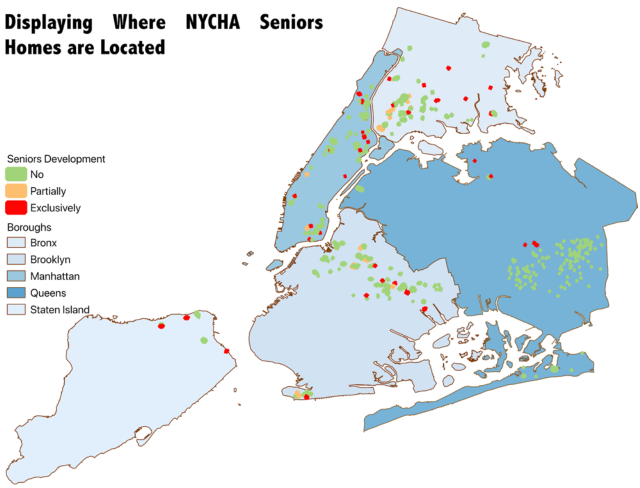


Figure 13: Seniors Homes are in every Borough

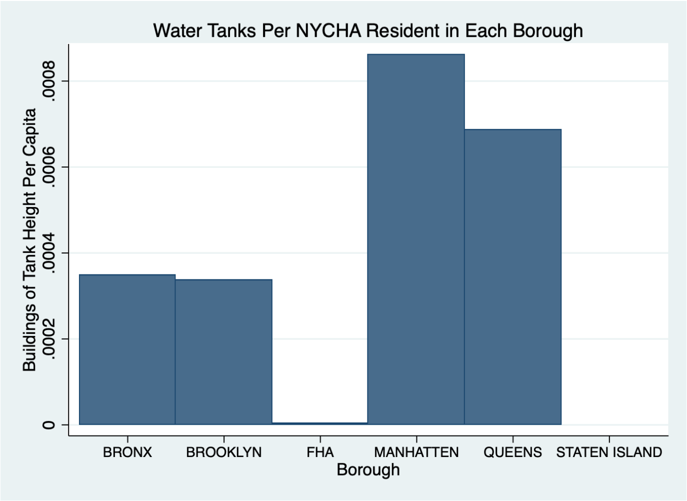


Figure 14: Manhattan has the most water tank height buildings per NYCHA resident

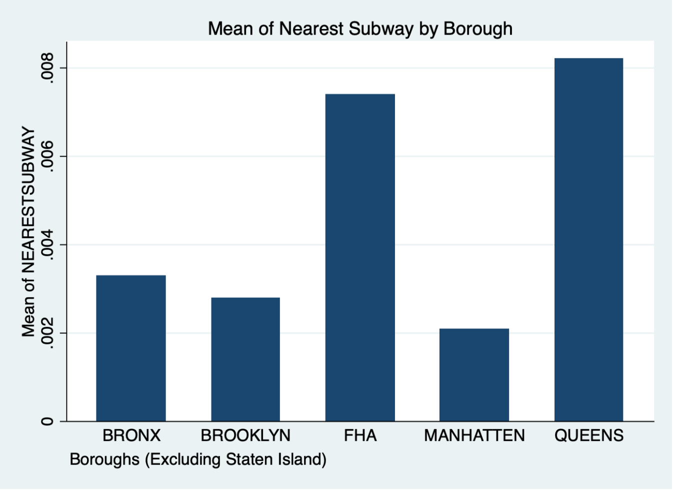


Figure 15: FHA Buildings and residents in Queens are generally furthest from subway stations

10 Appendix 1: Summary Statistics from Data Set

Variable	Mean	Std. Dev.	Min.	Max.	N
CONSUMPTION	74.338	781.553	0	121553.875	24772
avgmonthlygrossrent	734.672	194.588	235	1249	24776
PERCENTSECT8TRANS	0.019	0.056	0	0.314	24776
REHAB	0.10268	0.3035466	0	1	24776
DAYSOLD	15632.446	4718.763	5943	30598	24756
METHOD	0.292	0.455	0	1	24776
density	110.906	111.07	10	468	24776
numberofnonresidentialbldgs	0.308	0.698	0	4	24776
COSTPERROOM	9944.041	11972.41	0	181844.125	24688
PRIVATE	0.064	0.244	0	1	24776
WATERTANK	0.286	0.452	0	1	24776
NEARESTSUBWAY	0.005	0.006	0	0.049	24498
ELECTRICITY	0.399	0.49	0	1	24776
othercharges	-0.064	476.62	-67726.828	17893.84	24776

Figure 16: Density shown in equal Quantiles, rather than equal intervals

10 Appendices

Appendix 1: Summary Statistics from Data

Table 2: SENIORS Tabulated

Whether Building is Seniors Home	Freq.	Percent	Cum.
NO	23,719	95.73	95.73
PARTIALLY	121	0.49	96.22
EXCLUSIVELY	936	3.78	100
Total	24,776	100.00	

Table 3: FUNDING Tabulated

Funding Source	Freq.	Percent	Cum.
FEDERAL	21,877	88.30	88.30
MHOP	357	1.44	89.74
MIXED FINANCE/LLC1	2,429	9.80	99.54
MIXED FINANCE/LLC2	113	0.46	100

Total	24,776	100.00

Table 4: BOROUGH Tabulated

Borough Location	Freq.	Percent	Cum.
BRONX	1,971	7.96	7.96
BROOKLYN	4,778	19.28	27.24
FHA	11,367	45.88	73.12
MANHATTEN	4,611	18.61	91.73
QUEENS	1,978	7.98	99.71
STATEN ISLAND	71	0.29	100
Total	24,776	100.00	

Table 5: REVENUEMONTH Tabulated

Month of Bill	Freq.	Percent	Cum.
JANUARY	2,062	8.32	8.32
FEBRUARY	2,032	8.20	16.52
MARCH	1,974	7.97	24.49
APRIL	2,048	8.27	32.76
MAY	2,221	8.96	41.72
JUNE	2,372	9.57	51.30
JULY	2,308	9.32	60.61
AUGUST	2,101	8.48	69.09
SEPTEMBER	2,148	8.67	77.76
OCTOBER	1,970	7.95	85.71
NOVEMBER	1,631	6.58	92.29
DECEMBER	1,909	7.71	100
Total	24,776	100.00	

Table 6: REVENUEYEAR Tabulated

Year of Bill	Freq.	Percent	Cum.
2013	3,263	13.17	13.17
2014	4,195	16.93	30.10
2015	4,287	17.30	47.40
2016	3,857	15.57	62.97
2017	3,456	13.95	76.92
2018	3,498	14.12	91.04
2019	2,220	8.96	100
Total	24,776	100.00	

Table 7: RATECLASS Tabulated

Rate Class Charge	Freq.	Percent	Cum.
BASIC WATER AND SEWER	24,676	99.60	99.60
COMMERCIAL	9	0.04	99.63
HOT OR COLD WATER IN STORE	87	0.35	99.98
MULTIFAMILY	1	0.00	99.99
WATER-METER	3	0.01	100
Total	24,776	100.00	

Appendix 2: PYTHON FILE - Parse the NYCHAWater Consumption Data to Create date variables

```python

#! /usr/bin/env python3

import csv
import os
import re

os.chdir('../')

ReaderFile = open('Water_Consumption_And_Cost__2013_-_2019_.csv')
Reader = csv.reader(ReaderFile)
ReaderData = list(Reader)

yearFinder = re.compile(r'\d{4}')
monthFinder = re.compile(r'\d{2}')

ReaderData[0].append('REVENUEYEAR')
ReaderData[0].append('REVMONTH')

for row in ReaderData:
    year = yearFinder.search(row[13])
    month = monthFinder.search(yearFinder.sub('', row[13]))
    if year is None or month is None:
        print(row[13])
        continue
    row.append(year.group())
    row.append(month.group())

writerFile = open('Water_Consumption_And_Cost__2013_-_2019_edited.csv', 'w'
    )
outputWriter = csv.writer(writerFile)
for row in ReaderData:
    outputWriter.writerow(row)
writerFile.close()
```

Appendix 3: PYTHON FILE - Get the Age, in Number of Days since March 9, 2020, of Each NYCHA Development

```
#! /usr/bin/env python3

import csv
import os
import datetime

os.chdir('../')

ReaderFile = open('NYCHA_Development_Data_Book.csv')
Reader = csv.reader(ReaderFile)
ReaderData = list(Reader)
allDates = []
for row in ReaderData:
    try:
        allDates.append(datetime.datetime.strptime(row[46], '%m/%d/%Y'))
    except ValueError:
        if row[46] == '':
            allDates.append('.')
        continue  # The only time that it doesn't match is row 1, the title row
dt = datetime.datetime.now()
DaysDiff = ['DAYS_SINCE_COMPLETION']
for item in allDates:
    if type(item) == str:
        DaysDiff.append('')
        continue
    delta = dt - item
    DaysDiff.append(delta.days)

outputFile = open('NYCHA_Development_Data_Book_edited.csv', 'w')
outputWriter = csv.writer(outputFile)
for x in range(len(ReaderData)):
    ReaderData[x].append(DaysDiff[x])
for row in ReaderData:
    outputWriter.writerow(row)
outputFile.close()
```

Appendix 4: PYTHON FILE - Parse the NYCHA Development Data Book Set to Allow for Destringing

```python
#! /usr/bin/env python3

import csv
import os
import re

os.chdir('../')

ReaderFile = open('NYCHA_Development_Data_Book_edited.csv')
Reader = csv.reader(ReaderFile)
ReaderData = list(Reader)

commaFinder = re.compile(r',')
dollarFinder = re.compile(r'\$')
editedData = [ReaderData[0]]
del ReaderData[0]

for row in ReaderData:
    currentApartments = commaFinder.sub('', row[12])
    totalApartments = commaFinder.sub('', row[13])
    rentalRooms = commaFinder.sub('', row[14])
    publicHousing = commaFinder.sub('', row[17])
    totalPopulation = commaFinder.sub('', row[18])
    totalArea = commaFinder.sub('', row[25])
    devArea = commaFinder.sub('', row[27])
    coverageArea = commaFinder.sub('', row[29])
    cubeFeet = commaFinder.sub('', row[30])
    percent = row[31]
    percent = percent[0:(len(percent)-1)]
    devCost = dollarFinder.sub('', commaFinder.sub('', row[33]))
    perRentalCost = dollarFinder.sub('', commaFinder.sub('', row[34]))
    rentCost = dollarFinder.sub('', commaFinder.sub('', row[35]))

    newRow = row[0:12]
    newRow.append(currentApartments)
    newRow.append(totalApartments)
    newRow.append(rentalRooms)
    for item in row[15:17]:
        newRow.append(item)
    newRow.append(publicHousing)
    newRow.append(totalPopulation)
    for item in row[19:25]:
        newRow.append(item)
    newRow.append(totalArea)
    newRow.append(row[26])
    newRow.append(devArea)
    newRow.append(row[28])
    newRow.append(coverageArea)
    newRow.append(cubeFeet)
    newRow.append(percent)
    newRow.append(row[32])
    newRow.append(devCost)
    newRow.append(perRentalCost)
    newRow.append(rentCost)
    for item in row[36:]:
        newRow.append(item)
    editedData.append(newRow)

writerFile = open('NYCHA_Development_Data_Book_edited.csv', 'w')
outputWriter = csv.writer(writerFile)
for row in editedData:
    outputWriter.writerow(row)
writerFile.close()
```

Appendix 5: STATA DO FILE - Clean and Merge Open Data Sets

```
/* All work in this do-file is done in the order that the variables are in
    in the dataset*/
/* All original variables are lower case, all variables I create are upper
    case */
clear all

/* First set up and create the subway distance variables */
/*Change the directory to where everything is*/
cd "NYC/NYC_Data"

import delimited "./SubwayDistance/SubwayDistance.csv", delimiter(comma)
    varnames(1) case(upper)

/* Switch tds to be lower case so it matches later data sets*/
rename TDS tds
rename SUBWAYDISTANCE NEARESTSUBWAY
label var NEARESTSUBWAY "Nearest Subway"

save "SubwayDistance.dta", replace

/* Now to set-up and clean the data from NYC Open Data on NYCHA development
     data book*/
clear all

/*Change the directory to where everything is*/
cd "/users/dudliness/Library/Mobile Documents/com~apple~CloudDocs/School/
    UBC/2019W2/ECON490/NYC/NYC_Data"

/* Import the data */
import delimited "NYCHA_Development_Data_Book_edited.csv", delimiter(comma)
     varnames(1)

/* The tds numbers are given as strings rather than integers since some of
    them have typos*/
/* Force the strings to become integers and then delete any observations
    that are missing that value*/
/* This is fine since the four observations that have typos are not
    buildings that are in the water consumption data*/
destring tds, replace force
drop if missing(tds)

/* Label all the tds values that also appear in the NYCHA water consumption
     data set*/
label define tds_label, add
label define tds_label 1 "FIRST HOUSES", add
label define tds_label 2 "WILLIAMSBURG", add
```

```
label define tds_label 4 "RED HOOK EAST", add
label define tds_label 5 "QUEENSBRIDGE SOUTH", add
label define tds_label 8 "SOUTH JAMAICA I", add
label define tds_label 14 "INGERSOLL", add
label define tds_label 27 "SMITH", add
label define tds_label 28 "MELROSE", add
label define tds_label 37 "RANGEL", add
label define tds_label 43 "NOSTRAND", add
label define tds_label 46 "BOULEVARD", add
label define tds_label 51 "OCEAN BAY APARTMENTS (OCEANSIDE)", add
label define tds_label 61 "VAN DYKE", add
label define tds_label 66 "SOUTH JAMAICA II", add
label define tds_label 67 "SOTOMAYOR HOUSES", add
label define tds_label 83 "MARLBORO", add
label define tds_label 91 "BAISLEY PARK", add
label define tds_label 92 "BAY VIEW", add
label define tds_label 97 "TAFT", add
label define tds_label 100 "GOMPERS", add
label define tds_label 114 "STAPLETON", add
label define tds_label 123 "CLINTON", add
label define tds_label 143 "REHAB PROGRAM (COLLEGE POINT)", add
label define tds_label 149 "POLO GROUNDS TOWERS", add
label define tds_label 167 "REID APARTMENTS", add
label define tds_label 205 "FENIMORE-LEFFERTS", add
label define tds_label 209 "FHA REPOSSESSED HOUSES (GROUP I)", add
label define tds_label 210 "ARMSTRONG I", add
label define tds_label 212 "FHA REPOSSESSED HOUSES (GROUP II)", add
label define tds_label 213 "FHA REPOSSESSED HOUSES (GROUP III)", add
label define tds_label 222 "BETANCES III, 9A", add
label define tds_label 226 "FHA REPOSSESSED HOUSES (GROUP IV)", add
label define tds_label 232 "CONLON LIHFE TOWER", add
label define tds_label 235 "BRYANT AVENUE-EAST 174TH STREET", add
label define tds_label 237 "EAST 152ND STREET-COURTLANDT AVENUE", add
label define tds_label 260 "FHA REPOSSESSED HOUSES (GROUP V)", add
label define tds_label 263 "EAST NEW YORK CITY LINE", add
label define tds_label 267 "MORRISANIA AIR RIGHTS", add
label define tds_label 268 "THOMAS APARTMENTS", add
label define tds_label 273 "FHA REPOSSESSED HOUSES (GROUP VI)", add
label define tds_label 274 "FHA REPOSSESSED HOSUES (GROUP VII)", add
label define tds_label 275 "FHA REPOSSESSED HOUSES (GROUP VIII)", add
label define tds_label 279 "SHELTON HOUSE", add
label define tds_label 283 "FHA REPOSSESSED HOUSES (GROUP IX)", add
label define tds_label 284 "FHA REPOSSESSED HOUSES (GROUP X)", add
label define tds_label 285 "BETANCES VI", add
label define tds_label 292 "LOWER EAST SIDE REHAB (GROUP 5)", add
label define tds_label 293 "WASHINGTON HEIGHTS REHAB (GROUPS 1&2)", add
label define tds_label 296 "MANHATTANVILLE REHAB (GROUP 2)", add
label define tds_label 302 "BUSHWICK II (GROUPS A & C)", add
label define tds_label 303 "BUSHWICK II (GROUPS B & D)", add
label define tds_label 304 "EAST 165TH STREET-BRYANT AVENUE", add
```

```
label define tds_label 305 "SOUTH BRONX AREA (SITE 402)", add
label define tds_label 307 "CLAREMONT REHAB (GROUP 2)", add
label define tds_label 309 "FORT WASHINGTON AVENUE REHAB", add
label define tds_label 312 "CROWN HEIGHTS", add
label define tds_label 313 "OCEAN HILL-BROWNSVILLE", add
label define tds_label 316 "INTERNATIONAL TOWER", add
label define tds_label 322 "EAST 004TH STREET REHAB", add
label define tds_label 324 "BUSHWICK II CDA (GROUP E)", add
label define tds_label 326 "LOWER EAST SIDE I INFILL", add
label define tds_label 330 "WASHINGTON HEIGHTS REHAB PHASE IV (C)", add
label define tds_label 331 "WASHIGNTON HEIGHTS REHAB PHASE IV (D)", add
label define tds_label 333 "STUYVESANT GARDENS II", add
label define tds_label 335 "CLAREMONT REHAB (GROUP 4)", add
label define tds_label 337 "LOWER EAST SIDE II", add
label define tds_label 338 "EAST 173RD STREET-VYSE AVENUE", add
label define tds_label 339 "HOWARD AVENUE", add
label define tds_label 343 "UPACA (SITE 5)", add
label define tds_label 345 "BELMONT-SUTTER AREA", add
label define tds_label 346 "BOYNTON AVENUE REHAB", add
label define tds_label 351 "PARK ROCK REHAB", add
label define tds_label 352 "RALPH AVENUE REHAB", add
label define tds_label 353 "STEBBINS AVENUE-HEWITT PLACE", add
label define tds_label 354 "TAPSCOTT STREET REHAB", add
label define tds_label 355 "UPACA (SITE 6)", add
label define tds_label 356 "UNION AVENUE-EAST 166TH STREET", add
label define tds_label 357 "BERRY STREET-SOUTH 9TH STREET", add
label define tds_label 358 "MARCY AVENUE-GREENE AVENUE SITE B", add
label define tds_label 359 "154 WEST 84TH STREET", add
label define tds_label 360 "WEST FARMS ROAD REHAB", add
label define tds_label 362 "LONGFELLOW AVENUE REHAB", add
label define tds_label 364 "LOWER EAST SIDE III", add
label define tds_label 365 "HOWARD AVENUE-PARK PLACE", add
label define tds_label 366 "STERLING PLACE REHABS (SAINT JOHNS-STERLING)",
    add
label define tds_label 367 "HUNTS POINT AVENUE REHAB", add
label define tds_label 368 "STERLING PLACE REHABS (STERLING-BUFFALO)", add
label define tds_label 370 "HIGHBRIDGE REHABS (ANDERSON AVENUE)", add
label define tds_label 377 "SAMUEL (CITY)", add
label define tds_label 389 "SAMUEL (MHOP) I", add
label define tds_label 398 "SAMUEL (MHOP) I", add
label define tds_label 505 "QUEENSBRIDGE NORTH", add
label define tds_label 514 "WHITMAN", add
label define tds_label 516 "REHAB PROGRAM (TAFT REHABS)", add
label define tds_label 523 "WASHINGTON HEIGHTS REHAB PHASE III", add
label define tds_label 524 "FRANKLIN AVENUE III CONVENTIONAL", add
label define tds_label 525 "FRANKLIN AVENUE I CONVENTIONAL", add
label define tds_label 526 "WEST FARMS SQUARE CONVENTIONAL", add
label define tds_label 531 "FRANKLIN AVENUE II CONVENTIONAL", add
label define tds_label 547 "HARRISON AVENUE REHAB (GROUP B)", add
label define tds_label 559 "STANTON STREET", add
```

```
label values tds tds_label

/* All these variables are unnessecary*/
drop dataasof
drop hudamp
drop consolidatedtds
drop developmentedp
drop operatingedp
drop hud
drop program
drop borough
drop completiondate
drop nycitycouncildistrict
drop nystateassemblydistrict
drop nystatesenatedistrict
drop uscongressionaldistrict
drop communitydistirct

/* Code and Label the method variable */
gen METHOD = 0
replace METHOD = 1 if method == "TURNKEY"
label define METHOD_label, add
label define METHOD_label 0 "CONVENTIONAL", add
label define METHOD_label 1 "TURNKEY", add
label values METHOD METHOD_label
label var METHOD "Building Management Method"
order METHOD, before(method)
drop method

/* Code and Label the type variable */
gen TYPE = 1
replace TYPE = 2 if type == "NEW CONST"
replace TYPE = 3 if type == "NEW CONST (ELD)"
replace TYPE = 4 if type == "REHAB"
replace TYPE = 5 if type == "REHAB (ELD)"
label define TYPE_label, add
label define TYPE_label 1 "GUT REHAB", add
label define TYPE_label 2 "NEW CONST", add
label define TYPE_label 3 "NEW CONST (ELD)", add
label define TYPE_label 4 "REHAB", add
label define TYPE_label 5 "REHAB (ELD)", add
label values TYPE TYPE_label
label var TYPE "The Type of Building"
order TYPE, before(type)
drop type

rename days_since_completion DAYSOLD
label var DAYSOLD "Development Age"
label var perrentalroom "Dev Cost per Room"
```

```
/* Code and Label the federalized development variable*/
gen SENIORS = 0
replace SENIORS = 1 if seniordevelopment == "PARTIALLY"
replace SENIORS = 2 if seniordevelopment == "EXCLUSIVELY"
label define SENIORS_label, add
label define SENIORS_label 0 "NO", add
label define SENIORS_label 1 "PARTIALLY", add
label define SENIORS_label 2 "EXCLUSIVELY", add
label values SENIORS SENIORS_label
label var SENIORS "Whether Building is Exclusively Seniors"
order SENIORS, before(seniordevelopment)
drop seniordevelopment

/* Convert bldgcoverage to decimal */
gen BLDGCOV = bldgcoverage/100
order BLDGCOV, before(bldgcoverage)
label var BLDGCOV "Building Coverage (%)"
drop bldgcoverage

/* Convert electricity paid to Dummy variable */
gen ELECTRICITY = 0
replace ELECTRICITY = 1 if electricitypaidbyresidents == "YES"
label define ELECTRICITY_label, add
label define ELECTRICITY_label 0 "NO", add
label define ELECTRICITY_label 1 "YES", add
label values ELECTRICITY ELECTRICITY_label
label var ELECTRICITY "If Electricity Billed Directly"
order ELECTRICITY, before(electricitypaidbyresidents)
drop electricitypaidbyresidents

/* Convert private management to Dummy variable */
gen PRIVATE = 0
replace PRIVATE = 1 if privatemanagement == "YES"
label define PRIVATE_label, add
label define PRIVATE_label 0 "NO", add
label define PRIVATE_label 1 "YES", add
label values PRIVATE PRIVATE_label
label var PRIVATE "If Building is Managed Privately"
order PRIVATE, before(privatemanagement)
drop privatemanagement

/* Make numStories, tallest numStories */
replace numberofstories = "21" in 13
replace numberofstories = "14" in 15
replace numberofstories = "13" in 16
replace numberofstories = "6" in 18
replace numberofstories = "7" in 20
replace numberofstories = "14" in 25
replace numberofstories = "6" in 29
replace numberofstories = "6" in 32
```

```
replace numberofstories = "18" in 38
replace numberofstories = "14" in 39
replace numberofstories = "6" in 40
replace numberofstories = "7" in 42
replace numberofstories = "14" in 45
replace numberofstories = "7" in 47
replace numberofstories = "20" in 49
replace numberofstories = "17" in 54
replace numberofstories = "17" in 55
replace numberofstories = "15" in 57
replace numberofstories = "20" in 59
replace numberofstories = "7" in 62
replace numberofstories = "6" in 63
replace numberofstories = "5" in 65
replace numberofstories = "18" in 68
replace numberofstories = "20" in 82
replace numberofstories = "20" in 83
replace numberofstories = "14" in 87
replace numberofstories = "11" in 92
replace numberofstories = "8" in 93
replace numberofstories = "14" in 94
replace numberofstories = "12" in 95
replace numberofstories = "2" in 98
replace numberofstories = "3" in 99
replace numberofstories = "2" in 100
replace numberofstories = "3" in 101
replace numberofstories = "3" in 102
replace numberofstories = "3" in 103
replace numberofstories = "2.5" in 104
replace numberofstories = "2.5" in 105
replace numberofstories = "2.5" in 106
replace numberofstories = "2.5" in 107
replace numberofstories = "5" in 109
replace numberofstories = "14" in 110
replace numberofstories = "25" in 113
replace numberofstories = "14" in 114
replace numberofstories = "24" in 116
replace numberofstories = "14" in 119
replace numberofstories = "21" in 121
replace numberofstories = "15" in 126
replace numberofstories = "5" in 127
replace numberofstories = "6" in 130
replace numberofstories = "14" in 132
replace numberofstories = "14" in 135
replace numberofstories = "13" in 136
replace numberofstories = "5" in 140
replace numberofstories = "1" in 143
replace numberofstories = "1" in 144
replace numberofstories = "" in 143
replace numberofstories = "14" in 143
```

```
replace numberofstories = "11" in 143
replace numberofstories = "10" in 144
replace numberofstories = "14" in 147
replace numberofstories = "14" in 149
replace numberofstories = "20" in 154
replace numberofstories = "14" in 161
replace numberofstories = "14" in 162
replace numberofstories = "18" in 165
replace numberofstories = "9" in 166
replace numberofstories = "6" in 171
replace numberofstories = "6" in 172
replace numberofstories = "15" in 173
replace numberofstories = "6" in 177
replace numberofstories = "16" in 178
replace numberofstories = "11" in 183
replace numberofstories = "20" in 187
replace numberofstories = "15" in 188
replace numberofstories = "20" in 190
replace numberofstories = "20" in 191
replace numberofstories = "29" in 194
replace numberofstories = "22" in 195
replace numberofstories = "16" in 201
replace numberofstories = "15" in 205
replace numberofstories = "13" in 206
replace numberofstories = "16" in 208
replace numberofstories = "8" in 211
replace numberofstories = "7" in 220
replace numberofstories = "6" in 221
replace numberofstories = "14" in 222
replace numberofstories = "7" in 223
replace numberofstories = "7" in 225
replace numberofstories = "14" in 230
replace numberofstories = "14" in 231
replace numberofstories = "18" in 234
replace numberofstories = "15" in 235
replace numberofstories = "7" in 241
replace numberofstories = "15" in 246
replace numberofstories = "4" in 255
replace numberofstories = "7" in 256
replace numberofstories = "8" in 258
replace numberofstories = "20" in 262
replace numberofstories = "12" in 265
replace numberofstories = "15" in 266
replace numberofstories = "6" in 267
replace numberofstories = "13" in 270
replace numberofstories = "7" in 273
replace numberofstories = "11" in 274
replace numberofstories = "16" in 277
replace numberofstories = "14" in 287
replace numberofstories = "16" in 292
```

```
replace numberofstories = "14" in 293
replace numberofstories = "14" in 294
replace numberofstories = "6" in 295
replace numberofstories = "5" in 301
replace numberofstories = "13" in 308
replace numberofstories = "21" in 309
replace numberofstories = "25" in 314
replace numberofstories = "6" in 315
replace numberofstories = "20" in 81
replace numberofstories = "20" in 84
replace numberofstories = "14" in 88
replace numberofstories = "12" in 96
replace numberofstories = "2.5" in 108
replace numberofstories = "14" in 111
replace numberofstories = "14" in 115
replace numberofstories = "24" in 117
replace numberofstories = "14" in 120
replace numberofstories = "21" in 122
replace numberofstories = "5" in 128
replace numberofstories = "6" in 131
replace numberofstories = "14" in 133
replace numberofstories = "13" in 137
replace numberofstories = "5" in 141
replace numberofstories = "14" in 148
replace numberofstories = "14" in 150
replace numberofstories = "20" in 155
replace numberofstories = "14" in 163
replace numberofstories = "9" in 167
replace numberofstories = "15" in 174
replace numberofstories = "16" in 179
replace numberofstories = "11" in 184
replace numberofstories = "15" in 189
replace numberofstories = "20" in 192
replace numberofstories = "22" in 196
replace numberofstories = "16" in 202
replace numberofstories = "13" in 207
replace numberofstories = "16" in 209
replace numberofstories = "8" in 212
replace numberofstories = "14" in 224
replace numberofstories = "7" in 226
replace numberofstories = "7" in 228
replace numberofstories = "14" in 233
replace numberofstories = "18" in 237
replace numberofstories = "15" in 238
replace numberofstories = "7" in 244
replace numberofstories = "15" in 249
replace numberofstories = "7" in 259
replace numberofstories = "8" in 261
replace numberofstories = "12" in 268
replace numberofstories = "15" in 269
```

```
replace numberofstories = "7" in 276
replace numberofstories = "16" in 280
replace numberofstories = "14" in 290
replace numberofstories = "14" in 296
replace numberofstories = "14" in 297
replace numberofstories = "6" in 298
replace numberofstories = "5" in 305
replace numberofstories = "13" in 312
replace numberofstories = "21" in 313
replace numberofstories = "25" in 318
replace numberofstories = "6" in 319
destring numberofstories, replace

save "NYCHA_Data_Book.dta", replace

/* This do-file is to set-up and clean the data from NYC Open Data on NYCHA
    water consumption*/
/* All work in this do-file is done in the order that the variables are in
    in the dataset*/
/* All original variables are lower case, all variables I create are upper
    case */
clear all

/*Change the directory to where everything is*/
cd "/users/dudliness/Library/Mobile Documents/com~apple~CloudDocs/School/
    UBC/2019W2/ECON490/NYC/NYC_Data"

/* Import the data */
import delimited "Water_Consumption_And_Cost__2013_-_2019_edited.csv",
    delimiter(comma) varnames(1)

/* Rename and re-order the variable created through Python Parsing*/
rename revenueyear REVENUEYEAR
rename revmonth REVENUEMONTH
order REVENUEYEAR, after(revenuemonth)
order REVENUEMONTH, before(REVENUEYEAR)
label var REVENUEYEAR "Year of Bill"
label var REVENUEMONTH "Month of Bill"

/* Create labels for the month Variable */
label define REVENUEMONTH_label, add /* Label all the new values in the
    variable */
label define REVENUEMONTH_label 1 "JANUARY", add
label define REVENUEMONTH_label 2 "FEBRUARY", add
label define REVENUEMONTH_label 3 "MARCH", add
label define REVENUEMONTH_label 4 "APRIL", add
label define REVENUEMONTH_label 5 "MAY", add
label define REVENUEMONTH_label 6 "JUNE", add
label define REVENUEMONTH_label 7 "JULY", add
label define REVENUEMONTH_label 8 "AUGUST", add
```

```
label define REVENUEMONTH_label 9 "SEPTEMBER", add
label define REVENUEMONTH_label 10 "OCTOBER", add
label define REVENUEMONTH_label 11 "NOVEMBER", add
label define REVENUEMONTH_label 12 "DECEMBER", add
label values REVENUEMONTH REVENUEMONTH_label

/* Label all the variables */
label var developmentname "Development Name"
label var borough "Borough"
label var accountname "Account Name"
label var location "Location"
label var meteramr "Automatic Reading Meter"
label var meterscope "Buildings Supplied By Meter"
label var tds "Tenant Data System Number"
label var edp "Electronic Data Processing Number"
label var rccode "Budget Responsibility Code"
label var fundingsource "Funding Source"
label var amp "Asset Management Project Number"
label var vendorname "Vendor Name"
label var umisbillid "UMIS Bill ID"
label var revenuemonth "Year and Month of Bill"
label var servicestartdate "Bill Start Date"
label var serviceenddate "Bill End Date"
label var days "Num of Days"
label var meternumber "Meter Number"
label var estimated "Estimated"
label var currentcharges "Current Charges"
label var rateclass "Rate Class"
label var billanalyzed "Bill Analyzed"
label var consumptionhcf "Consumption (Hundred Cubic Feet)"
label var watersewercharges "Water & Sewer Charges"
label var othercharges "OtherCharges"

/* Create coded variable for Boroughs */
gen BOROUGH = .
replace BOROUGH = 1 if borough == "BRONX" /* Match the new variable to old
    one*/
replace BOROUGH = 2 if borough == "BROOKLYN"
replace BOROUGH = 3 if borough == "FHA"
replace BOROUGH = 4 if borough == "MANHATTAN"
replace BOROUGH = 5 if borough == "QUEENS"
replace BOROUGH = 6 if borough == "STATEN ISLAND"
label define BOROUGH_label, add /* Label all the new values in the variable
     */
label define BOROUGH_label 1 "BRONX", add
label define BOROUGH_label 2 "BROOKLYN", add
label define BOROUGH_label 3 "FHA", add
label define BOROUGH_label 4 "MANHATTEN", add
label define BOROUGH_label 5 "QUEENS", add
label define BOROUGH_label 6 "STATEN ISLAND", add
```

```
label values BOROUGH BOROUGH_label
label var BOROUGH "Borough"
order BOROUGH, before(borough) /* Move it to the correct point in the
    dataset */
drop borough /* Get rid of the old variable */

/* Label all the tds values */
label define tds_label, add
label define tds_label 1 "FIRST HOUSES", add
label define tds_label 2 "WILLIAMSBURG", add
label define tds_label 4 "RED HOOK EAST", add
label define tds_label 5 "QUEENSBRIDGE SOUTH", add
label define tds_label 8 "SOUTH JAMAICA I", add
label define tds_label 14 "INGERSOLL", add
label define tds_label 27 "SMITH", add
label define tds_label 28 "MELROSE", add
label define tds_label 37 "RANGEL", add
label define tds_label 43 "NOSTRAND", add
label define tds_label 46 "BOULEVARD", add
label define tds_label 51 "OCEAN BAY APARTMENTS (OCEANSIDE)", add
label define tds_label 61 "VAN DYKE", add
label define tds_label 66 "SOUTH JAMAICA II", add
label define tds_label 67 "SOTOMAYOR HOUSES", add
label define tds_label 83 "MARLBORO", add
label define tds_label 91 "BAISLEY PARK", add
label define tds_label 92 "BAY VIEW", add
label define tds_label 97 "TAFT", add
label define tds_label 100 "GOMPERS", add
label define tds_label 114 "STAPLETON", add
label define tds_label 123 "CLINTON", add
label define tds_label 143 "REHAB PROGRAM (COLLEGE POINT)", add
label define tds_label 149 "POLO GROUNDS TOWERS", add
label define tds_label 167 "REID APARTMENTS", add
label define tds_label 205 "FENIMORE-LEFFERTS", add
label define tds_label 209 "FHA REPOSSESSED HOUSES (GROUP I)", add
label define tds_label 210 "ARMSTRONG I", add
label define tds_label 212 "FHA REPOSSESSED HOUSES (GROUP II)", add
label define tds_label 213 "FHA REPOSSESSED HOUSES (GROUP III)", add
label define tds_label 222 "BETANCES III, 9A", add
label define tds_label 226 "FHA REPOSSESSED HOUSES (GROUP IV)", add
label define tds_label 232 "CONLON LIHFE TOWER", add
label define tds_label 235 "BRYANT AVENUE-EAST 174TH STREET", add
label define tds_label 237 "EAST 152ND STREET-COURTLANDT AVENUE", add
label define tds_label 260 "FHA REPOSSESSED HOUSES (GROUP V)", add
label define tds_label 263 "EAST NEW YORK CITY LINE", add
label define tds_label 267 "MORRISANIA AIR RIGHTS", add
label define tds_label 268 "THOMAS APARTMENTS", add
label define tds_label 273 "FHA REPOSSESSED HOUSES (GROUP VI)", add
label define tds_label 274 "FHA REPOSSESSED HOSUES (GROUP VII)", add
label define tds_label 275 "FHA REPOSSESSED HOUSES (GROUP VIII)", add
```

```
label define tds_label 279 "SHELTON HOUSE", add
label define tds_label 283 "FHA REPOSSESSED HOUSES (GROUP IX)", add
label define tds_label 284 "FHA REPOSSESSED HOUSES (GROUP X)", add
label define tds_label 285 "BETANCES VI", add
label define tds_label 292 "LOWER EAST SIDE REHAB (GROUP 5)", add
label define tds_label 293 "WASHINGTON HEIGHTS REHAB (GROUPS 1&2)", add
label define tds_label 296 "MANHATTANVILLE REHAB (GROUP 2)", add
label define tds_label 302 "BUSHWICK II (GROUPS A & C)", add
label define tds_label 303 "BUSHWICK II (GROUPS B & D)", add
label define tds_label 304 "EAST 165TH STREET-BRYANT AVENUE", add
label define tds_label 305 "SOUTH BRONX AREA (SITE 402)", add
label define tds_label 307 "CLAREMONT REHAB (GROUP 2)", add
label define tds_label 309 "FORT WASHINGTON AVENUE REHAB", add
label define tds_label 312 "CROWN HEIGHTS", add
label define tds_label 313 "OCEAN HILL-BROWNSVILLE", add
label define tds_label 316 "INTERNATIONAL TOWER", add
label define tds_label 322 "EAST 004TH STREET REHAB", add
label define tds_label 324 "BUSHWICK II CDA (GROUP E)", add
label define tds_label 326 "LOWER EAST SIDE I INFILL", add
label define tds_label 330 "WASHINGTON HEIGHTS REHAB PHASE IV (C)", add
label define tds_label 331 "WASHIGNTON HEIGHTS REHAB PHASE IV (D)", add
label define tds_label 333 "STUYVESANT GARDENS II", add
label define tds_label 335 "CLAREMONT REHAB (GROUP 4)", add
label define tds_label 337 "LOWER EAST SIDE II", add
label define tds_label 338 "EAST 173RD STREET-VYSE AVENUE", add
label define tds_label 339 "HOWARD AVENUE", add
label define tds_label 343 "UPACA (SITE 5)", add
label define tds_label 345 "BELMONT-SUTTER AREA", add
label define tds_label 346 "BOYNTON AVENUE REHAB", add
label define tds_label 351 "PARK ROCK REHAB", add
label define tds_label 352 "RALPH AVENUE REHAB", add
label define tds_label 353 "STEBBINS AVENUE-HEWITT PLACE", add
label define tds_label 354 "TAPSCOTT STREET REHAB", add
label define tds_label 355 "UPACA (SITE 6)", add
label define tds_label 356 "UNION AVENUE-EAST 166TH STREET", add
label define tds_label 357 "BERRY STREET-SOUTH 9TH STREET", add
label define tds_label 358 "MARCY AVENUE-GREENE AVENUE SITE B", add
label define tds_label 359 "154 WEST 84TH STREET", add
label define tds_label 360 "WEST FARMS ROAD REHAB", add
label define tds_label 362 "LONGFELLOW AVENUE REHAB", add
label define tds_label 364 "LOWER EAST SIDE III", add
label define tds_label 365 "HOWARD AVENUE-PARK PLACE", add
label define tds_label 366 "STERLING PLACE REHABS (SAINT JOHNS-STERLING)",
    add
label define tds_label 367 "HUNTS POINT AVENUE REHAB", add
label define tds_label 368 "STERLING PLACE REHABS (STERLING-BUFFALO)", add
label define tds_label 370 "HIGHBRIDGE REHABS (ANDERSON AVENUE)", add
label define tds_label 377 "SAMUEL (CITY)", add
label define tds_label 389 "SAMUEL (MHOP) I", add
label define tds_label 398 "SAMUEL (MHOP) I", add
```

```
label define tds_label 505 "QUEENSBRIDGE NORTH", add
label define tds_label 514 "WHITMAN", add
label define tds_label 516 "REHAB PROGRAM (TAFT REHABS)", add
label define tds_label 523 "WASHINGTON HEIGHTS REHAB PHASE III", add
label define tds_label 524 "FRANKLIN AVENUE III CONVENTIONAL", add
label define tds_label 525 "FRANKLIN AVENUE I CONVENTIONAL", add
label define tds_label 526 "WEST FARMS SQUARE CONVENTIONAL", add
label define tds_label 531 "FRANKLIN AVENUE II CONVENTIONAL", add
label define tds_label 547 "HARRISON AVENUE REHAB (GROUP B)", add
label define tds_label 559 "STANTON STREET", add
label values tds tds_label
/* Lebeling tds values made these variables unnessecary */
drop accountname
drop developmentname

/* Code and Label the meteramr variable */
gen METERAMR = 0
replace METERAMR = 1 if meteramr == "AMR"
label define METERAMR_label, add
label define METERAMR_label 0 "NONE", add
label define METERAMR_label 1 "AMR", add
label values METERAMR METERAMR_label
label var METERAMR "Automatic Reading Meter"
order METERAMR, before(meteramr)
drop meteramr

/* These variables are not useful for my purposes */
drop edp
drop rccode

/* Code and Label the funding variable */
gen FUNDING = 0
replace FUNDING = 1 if fundingsource == "MHOP"
replace FUNDING = 2 if fundingsource == "MIXED FINANCE/LLC1"
replace FUNDING = 3 if fundingsource == "MIXED FINANCE/LLC2"
replace FUNDING = 4 if fundingsource == "SECTION 8"
label define FUNDING_label, add
label define FUNDING_label 0 "FEDERAL", add
label define FUNDING_label 1 "MHOP", add
label define FUNDING_label 2 "MIXED FINANCE/LLC1", add
label define FUNDING_label 3 "MIXED FINANCE/LLC2", add
label define FUNDING_label 4 "SECTION 8", add
label values FUNDING FUNDING_label
label var FUNDING "Funding Source"
order FUNDING, before(fundingsource)
drop fundingsource

/* Code and Label the vendorname variable */
gen VENDOR = 0
replace VENDOR = 1 if vendorname != "NEW YORK CITY WATER BOARD"
```

```
label define VENDOR_label, add
label define VENDOR_label 0 "NEW YORK CITY WATER BOARD", add
label define VENDOR_label 1 "OTHER", add
label values VENDOR VENDOR_label
label var VENDOR "Vendor Name"
order VENDOR, before(vendorname)
drop vendorname

/* Code and Label the estimated variable */
gen ESTIMATED = 0
replace ESTIMATED = 1 if estimated == "Y"
label define ESTIMATED_label, add
label define ESTIMATED_label 0 "NO", add
label define ESTIMATED_label 1 "YES", add
label values ESTIMATED ESTIMATED_label
label var ESTIMATED "If Bill Was Estimated"
order ESTIMATED, before(estimated)
drop estimated

/* Code and Label the rateclass variable */
gen RATECLASS = 0
replace RATECLASS = 1 if rateclass == "COMMERCIAL"
replace RATECLASS = 2 if rateclass == "HOT OR COLD WATER IN STORE"
replace RATECLASS = 3 if rateclass == "MULTIFAMILY"
replace RATECLASS = 4 if rateclass == "WATER-METER"
label define RATECLASS_label, add
label define RATECLASS_label 0 "BASIC WATER AND SEWER", add
label define RATECLASS_label 1 "COMMERCIAL", add
label define RATECLASS_label 2 "HOT OR COLD WATER IN STORE", add
label define RATECLASS_label 3 "MULTIFAMILY", add
label define RATECLASS_label 4 "WATER-METER", add
label values RATECLASS RATECLASS_label
label var RATECLASS "Rate Class"
order RATECLASS, before(rateclass)
drop rateclass

/* Code and Label the billanalyzed variable */
gen BILLANALYZED = 0
replace BILLANALYZED = 1 if billanalyzed == "Exception"
label define BILLANALYZED_label, add
label define BILLANALYZED_label 0 "YES", add
label define BILLANALYZED_label 1 "EXCEPTION", add
label values BILLANALYZED BILLANALYZED_label
label var BILLANALYZED "If Bill Was Analyzed for Errors"
order BILLANALYZED, before(billanalyzed)
drop billanalyzed

/* Create a new dummy variable to show if a building is a rehab centre */
gen REHAB = 0
replace REHAB = 1 if tds == 143 | tds == 292 | tds == 293 | tds == 296 |
```

```
    tds == 307 | tds == 309 | tds == 322 | tds == 330 | tds == 331 | tds ==
    335 | tds == 346 | tds == 351 | tds == 352 | tds == 354 | tds == 360 |
    tds == 362 | tds == 366 | tds == 367 | tds == 368 | tds == 370 | tds ==
    516 | tds == 523 | tds == 547
label define REHAB_label, add
label define REHAB_label 0 "NO", add
label define REHAB_label 1 "YES", add
label values REHAB REHAB_label
label var REHAB "If Dev. is Rehabilitated"

/* Create a new dummy variable to show if a building is FHA Repossessed*/
gen FHA_REPO = 0
replace FHA_REPO = 1 if tds == 209 | tds == 212 | tds == 213 | tds == 226 |
     tds == 260 | tds == 273 | tds == 274 | tds == 275 | tds == 283 | tds ==
     284
label define FHA_REPO_label, add
label define FHA_REPO_label 0 "NO", add
label define FHA_REPO_label 1 "YES", add
label values FHA_REPO FHA_REPO_label
label var FHA_REPO "If Building is FHA Repossessed"

/* Merge the two datasets together using the tds number*/
merge m:1 tds using "NYCHA_Data_Book.dta"

/* Drop observations that failed to merge correctly, then drop merge
    variable*/
drop if _merge==1
drop if _merge==2
drop _merge

/* Drop observations that have had exceptions on their bills and those that
     have had their usage estimated */
drop if ESTIMATED==1
drop if BILLANALYZED==1

/* Some variables have missing values instead of zeros */
/* Replace these missing values with zeros */
replace percentfixedincomehousehold = 0 if percentfixedincomehouseholds ==
    .
replace populationsection8transition = 0 if populationsection8transition ==
     .
replace numberofnonresidentialbldgs = 0 if numberofnonresidentialbldgs == .

/* Creating some needed variables for the regressions */
/* Converting nominal numbers to be relative */
gen PERCENTSECT8TRANS = populationsection8transition / totalpopulation
label var PERCENTSECT8TRANS "% Section 8 Transition"
gen COSTPERROOM = developmentcost / numberofrentalrooms
label var COSTPERROOM "Development cost per rental room"

```

```
/* Create the dependent variable per capita per day, and convert to litres
    */
gen CONSUMPTION = (consumptionhcf/totalpopulation)/days * 2831.6846592
label var CONSUMPTION "Litres per day per person"

/* Create a variable that is equal to one if building is greater than 6
    stories*/
gen WATERTANK = .
replace WATERTANK = 0 if numberofstories <= 6
replace WATERTANK = 1 if numberofstories > 6
label define WATERTANK_label, add
label define WATERTANK_label 0 "Less than 6", add
label define WATERTANK_label 1 "More than 6", add
label values WATERTANK WATERTANK_label
label var WATERTANK "Water Tank"

/* Label variables used in regression, so the names are correct there*/
label var percentfixedincomehouseholds "% Fixed Income"
label var avgmonthlygrossrent "Avg Monthly Rent"
label var density "Density"
label var numberofnonresidentialbldgs "# of Non-Residential BLDGS"

/* Now we merge with the subway distances data that was obtained via QGIS
    */
merge m:1 tds using "SubwayDistance.dta"

/* There are some NYCHA buildings not included in the water tracking */
drop if _merge==2
```

Appendix 6: STATA DO FILE - Get NYCHA Water Consumption Summary Statistics

```
/*Outputting Data*/
/*MAKE SURE TO CD to CWD*/
cd "NYC_Data/Tables"

/* To get the table of summary statistics for all quantitative variables*/
ssc install estout
ssc install sutex
sutex CONSUMPTION percentfixedincomehouseholds avgmonthlygrossrent
    PERCENTSECT8TRANS DAYSOLD SENIORS FHA_REPO METHOD density
    numberofnonresidentialbldgs COSTPERROOM PRIVATE NEARESTSUBWAY
    ELECTRICITY REVENUEMONTH REVENUEYEAR numberofstories othercharges,
    minmax

/* To get the table data to summarize qualitative variables*/
estpost tab FUNDING
eststo fundingTable
esttab fundingTable using fundingTable.tex, tex replace

estpost tab BOROUGH
eststo boroughTable
esttab boroughTable using boroughTable.tex, tex replace

estpost tab RATECLASS
eststo rateclassTable
esttab rateclassTable using rateclassTable.tex, tex replace
```

Appendix 7: Regression Tables in full

First the Table from the robust regressions, followed by the four-layered standard regression. The latter two tables appear as they do in the body of the paper, the rst regression is an extended version of that which appears in the paper's discussion section.

VARIABLES	(3) Location Litres per day per person	se	(4) Billing Litres per day per person	se
Litres per day per person				
Avg Monthly Rent	0.0535	(0.0510)	0.0566	(0.0512)
% Section 8 Transition	4,847**	(2,004)	4,327**	(2,068)
If Dev. is Rehabilitated	33.06	(29.25)	32.99	(29.30)
Development Age	-0.00285	(0.00234)	-0.00383	(0.00254)
Building is Partially Seniors	-63.61	(76.41)	-53.23	(76.63)
Building is Exclusively Seniors	268.8***	(36.39)	270.7***	(36.58)
Funding Source is MHOP	-44.92	(69.55)	-32.08	(70.68)
Funding Source is MIXED/LLC1	-960.7***	(356.5)	-851.6**	(373.6)
Funding Source is MIXED/LLC2	-1,380**	(628.4)	-1,214*	(646.9)
Building Management Method	69.08*	(35.58)	86.58**	(39.52)
Density	0.388***	(0.147)	0.389***	(0.147)
# of Non-Residential BLDGS	-11.82	(17.76)	-12.69	(17.99)
Development cost per rental room	9.43e-05	(0.000634)	7.93e-05	(0.000634)
If Building is Managed Privately	3.424	(32.47)	2.939	(32.49)
Water Tank	-96.92***	(21.98)	-102.5***	(23.03)
Nearest Subway	3,475***	(1,266)	3,540***	(1,266)
Borough is BROOKLYN	-23.16	(27.74)	-23.65	(27.78)
Borough is FHA	30.20	(49.77)	15.33	(54.44)
Borough is MANHATTEN	167.6***	(33.09)	167.5***	(33.09)
Borough is QUEENS	142.3***	(41.09)	139.3***	(41.94)
Borough is STATEN ISLAND	-883.3**	(355.6)	-802.2**	(364.4)
If Electricity Billed Directly			-34.60	(40.35)
Month of Bill			1.367	(1.502)
Year of Bill			4.750*	(2.720)
Rate Class			34.36	(40.49)
OtherCharges			-0.00543	(0.0105)
Constant	-25.49	(75.19)	-9,582*	(5,486)
Observations	24,407		24,407	
R^2	0.014		0.014	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

VARIABLES	(1) Financial Litres per day per person	se	(2) Building Litres per day per person	se
Litres per day per person				
Avg Monthly Rent	-0.219***	(0.0258)	0.0586	(0.0422)
% Section 8 Transition	-480.3***	(90.38)	76.44	(526.1)
If Dev. is Rehabilitated			7.034	(27.41)
Development Age			-0.00305*	(0.00171)
Building is Partially Seniors			4.870	(73.66)
Building is Exclusively Seniors			291.1***	(33.87)
Funding Source is MHOP			19.89	(53.10)
Funding Source is MIXED/LLC1			-141.7	(98.00)
Funding Source is MIXED/LLC2			-30.50	(179.5)
Building Management Method			-10.45	(18.57)
Density			0.554***	(0.116)
# of Non-Residential BLDGS			-15.66	(15.19)
Development cost per rental room			0.00107*	(0.000577)
If Building is Managed Privately			7.345	(26.26)
Water Tank			-17.83	(17.84)
Constant	244.5***	(19.94)	19.29	(56.07)
Observations	24,772		24,684	
R^2	0.004		0.011	

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

VARIABLES	(3) Location Litres per day per person	se	(4) Billing Litres per day per person	se
Litres per day per person				
Avg Monthly Rent	0.0535	(0.0510)	0.0566	(0.0512)
% Section 8 Transition	4,847**	(2,004)	4,327**	(2,068)
If Dev. is Rehabilitated	33.06	(29.25)	32.99	(29.30)
Development Age	-0.00285	(0.00234)	-0.00383	(0.00254)
Building is Partially Seniors	-63.61	(76.41)	-53.23	(76.63)
Building is Exclusively Seniors	268.8***	(36.39)	270.7***	(36.58)
Funding Source is MHOP	-44.92	(69.55)	-32.08	(70.68)
Funding Source is MIXED/LLC1	-960.7***	(356.5)	-851.6**	(373.6)
Funding Source is MIXED/LLC2	-1,380**	(628.4)	-1,214*	(646.9)
Building Management Method	69.08*	(35.58)	86.58**	(39.52)
Density	0.388***	(0.147)	0.389***	(0.147)
# of Non-Residential BLDGS	-11.82	(17.76)	-12.69	(17.99)
Development cost per rental room	9.43e-05	(0.000634)	7.93e-05	(0.000634)
If Building is Managed Privately	3.424	(32.47)	2.939	(32.49)
Water Tank	-96.92***	(21.98)	-102.5***	(23.03)
Nearest Subway	3,475***	(1,266)	3,540***	(1,266)
Borough is BROOKLYN	-23.16	(27.74)	-23.65	(27.78)
Borough is FHA	30.20	(49.77)	15.33	(54.44)
Borough is MANHATTEN	167.6***	(33.09)	167.5***	(33.09)
Borough is QUEENS	142.3***	(41.09)	139.3***	(41.94)
Borough is STATEN ISLAND	-883.3**	(355.6)	-802.2**	(364.4)
If Electricity Billed Directly			-34.60	(40.35)
Month of Bill			1.367	(1.502)
Year of Bill			4.750*	(2.720)
Rate Class			34.36	(40.49)
OtherCharges			-0.00543	(0.0105)
Constant	-25.49	(75.19)	-9,582*	(5,486)
Observations	24,407		24,407	
R^2	0.014		0.014	

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix 8: STATA DO FILE - Regression Equations

```
/* This is the initial version of the regressions, some variables not
    finalized */
cd "NYC/NYC_Data/Output"
ssc install outreg2
/* Regression 1: Financial Indicators */
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS, level(99)
est store Financial

/* Regression 2: Financial Indicators and Building Management Variables */
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS REHAB DAYSOLD i.
    SENIORS i.FUNDING METHOD density numberofnonresidentialbldgs COSTPERROOM
     PRIVATE WATERTANK, baselevels level(99)
est store Building

/* Regression 3: 1, 2, and location variables */
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS REHAB DAYSOLD i.
    SENIORS i.FUNDING METHOD density numberofnonresidentialbldgs COSTPERROOM
     PRIVATE WATERTANK NEARESTSUBWAY i.BOROUGH, baselevels level(99)
est store Location

/* Regression 4: 1, 2, 3, and Billing Specifics*/
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS REHAB DAYSOLD i.
    SENIORS i.FUNDING METHOD density numberofnonresidentialbldgs COSTPERROOM
     PRIVATE WATERTANK NEARESTSUBWAY i.BOROUGH ELECTRICITY REVENUEMONTH
    REVENUEYEAR RATECLASS othercharges, baselevels level(99)
est store Billing

/* Robustness 1: Using month as qualitative, not quantitative */
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS REHAB DAYSOLD i.
    SENIORS i.FUNDING METHOD density numberofnonresidentialbldgs COSTPERROOM
     PRIVATE WATERTANK NEARESTSUBWAY i.BOROUGH ELECTRICITY i.REVENUEMONTH
    REVENUEYEAR RATECLASS othercharges, baselevels level(99)
est store Month

/* Robustness 2: Using robust standard errors */
reg CONSUMPTION avgmonthlygrossrent PERCENTSECT8TRANS REHAB DAYSOLD i.
    SENIORS i.FUNDING METHOD density numberofnonresidentialbldgs COSTPERROOM
     PRIVATE WATERTANK NEARESTSUBWAY i.BOROUGH ELECTRICITY REVENUEMONTH
    REVENUEYEAR RATECLASS othercharges, baselevels level(99) robust
est store Robust

outreg2 [Financial Building] using Regressions1, replace sideway tex(pretty
     fragment) label
outreg2 [Location Billing] using Regressions2, replace sideway tex(pretty
    fragment) label
outreg2 [Month Robust] using RegressionsRobust, replace sideway tex(pretty
    fragment) label
```

Appendix 9: PYTHON FILE - Using GIS Data to create the maps of NYCHA Developments in New York City

```python
#! /usr/bin/env python3

import geopandas as gpd
import matplotlib.pyplot as plt

DIMENSIONS = 100

nycha = gpd.read_file('./Map_NYCHA/NYCHA.shp')
nyc = gpd.read_file('./boroughs/borroughs.shp')
subwayLines = gpd.read_file('./map_subway_lines/subway_lines.shp')
subwayStations = gpd.read_file('./map_subway_stations/subway_map.shp')

base = nyc.plot(color='navajowhite', linewidth=2.5, figsize=(DIMENSIONS, DIMENSIONS))
base2 = nycha.plot(ax=base, color='red', linewidth=0.2, figsize=(DIMENSIONS, DIMENSIONS))
base3 = subwayLines.plot(ax=base2, color='blue', figsize=(DIMENSIONS, DIMENSIONS))
layered = subwayStations.plot(ax=base3, color='black', linewidth=10, figsize=(DIMENSIONS, DIMENSIONS))
base.set_axis_off()
plt.savefig('./plots/nychaFullSubwayBoroughs.png', format='png', dpi=72)

nyc = gpd.read_file('./base_map/nyc_base.shp')

base = nyc.plot(color='black', figsize=(DIMENSIONS, DIMENSIONS))
nycha.plot(ax=base, color='red', linewidth=0.2, figsize=(DIMENSIONS, DIMENSIONS))
base.set_axis_off()
plt.savefig('./plots/nychaRed.png', format='png', dpi=72)
```

11. References

"The qgis nnjoin plugin". Arken. September 2019. http://arken.nmbu.no/~havatv/gis/ qgisplugins/NNJoin/.

N. Y. C. H. Authority. Nycha development data book: Nyc open data. 2019.https://data.cityofnewyork.us/Housing-Development/NYCHA -Development-Data-Book/evjd-dqpz.

N. Y. C. H. Authority. Water consumption and cost (2013 - 2019): Nyc open data. 2019. https://data.cityofnewyork.us/Housing-Development/Water -Consumption-And-Cost-2013-2019-/66be-66yr.

Barrera, Jorge. “Attawapiskat declares state of emergency over water quality.” CBC News. July 2019. https://www.cbc.ca/news/indigenous/attawapiskat-water-quality -emergency-1.5204652.

“High levels of lead in canadian drinking water.” CBC video, 2:01, November 2019. https://www.cbc.ca/news/thenational/high-levels-of-lead-in -canadian-drinking-water-1.5347924.

Chen, Stefanos. “New york’s millennial homeowners and where to find them”. New York Times. March 2020. https://www.nytimes.com/2020/03/20/realestate/millennial-first -time-home-buyer.html?searchResultPosition=1.

Salvo, Joseph J., Lobo, Arun Peter, Maurer, Erica. New york city population projections by age/sex & borough, 2010-2040. New York: The City of New York, Department of City Planning, 2013. https://www1.nyc.gov/assets/planning/download/pdf/data-maps/ nyc-population/projections _report _ 2010 2040.pdf.

Cohen, Michelle. “What’s so great about new york city water?” 6sqft. March 31, 2016. https:// www.6sqft.com/nyc-water-101-from-the-catskill-aqueduct-and-robotic -measurements-to-your-tap/.

Canada’s Ecofiscal Commission. Bridging the gap: Real options for meeting Canada’s 2030 ghg target. November 2019. https://ecofiscal.ca/reports/bridging-gap-real-options-meeting-canadas-2030-ghg-target/.

Metropolitan Transportation Authority. “Subway lines”, 2018. NYC Open Data. https://data.cityofnewyork.us/Transportation/Subway-Lines/3qz8-muuu.

New York City Housing Authority. “Map of nycha developments”, 2019. NYC Open Data. https://data.cityofnewyork.us/Housing-Development/Map-of-NYCHA-Developments/i9rv-hdr5.

New York City Housing Authority. “Shapefiles and base map”, 2019. NYC Open Data. https://data.cityofnewyork.us/Housing-Development/Shapefiles-and-base-map/2k7f-6s2k.
Metropolitan Transportation Authority. “Subway stations”, 2019. NYC Open Data. https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49.

New York City Government. “Borough boundaries”, 2020. NYC Open Data. https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm.

DeIorio, V. A. Statewide assessment of energy use by the municipal water and wastewater sector. New York State Energy Research and Development Authority November 2008.

EIA. U.s. energy information administration - eia - independent statistics and analysis. February, 2020.

EPA. “Greenhouse gas equivalencies calculator.” October 2018.

Ferré-Sadurní, Luis. “The rise and fall of new york public housing: An oral history”. New York Times. June 2018. https://www.nytimes.com/interactive/2018/06/25/nyregion/new-york -city-public-housing-history.html?module=inline.
Ferré-Sadurní, Luis. “11,168 children tested positive for lead. the city didn’t inspect the homes”. New York Times. September 2019. https://www.nytimes.com/2019/09/26/nyregion/nyc-lead-exposure.html?searchResultPosition=7.

Ferré-Sadurní, Luis. “Chelsea Apartment! Only $90 a night! (Ignore the NYCHA sign)”. New York Times. September 2019. https://www.nytimes.com/2019/11/09/nyregion/airbnb-new-york -nycha.html.

Ferré-Sadurní, Luis. "He's in charge of housing for 11,000 minnesotans. can he handle 400,000 new yorkers?". New York Times. June 2019. https://www.nytimes.com/2019/06/24/nyregion/ nycha-chairman-gregory-russ.html.

Ferré-Sadurní, Luis. "'Lighting money on fire' as mold and rats persist in new york public housing". New York Times. July 2019 https://www.nytimes.com/2019/07/26/nyregion/nycha-rats-roof-repairs.html.

Firat, M., M. E. Turan, and M. A. Yurdusev. "Comparative analysis of neural network techniques for predicting water consumption time series". Journal of Hydrology, 384(1-2), 46–51.

Forrest, Maura. "Canada needs carbon tax of $210 a tonne by 2030 to meet paris targets, report says". National Post. November 27th, 2019. https://nationalpost.com/news/politics/canada-needs-carbon-tax-of-210-a-tonne-by-2030-to-meet-paris-targets-report-says.

Gillingham, Ken. "The true cost of reducing greenhouse gas emissions". International Monetary Fund. December 2019. https://www.imf.org/external/pubs/ft/fandd/2019/12/the-true-cost-of-reducing-greenhouse-gas-emissions-gillingham.htm.

Golden, J. S., V. Subramanian, G. M. A. U. Irizarri, P. White, and F. Meier. "Energy and carbon impact from residential laundry in the united states". Journal of Integrative Environmental Sciences 7(1) (March 2010): 53–73.

Griggs, Mary Beth. "Lead contamination goes far beyond flint, Michigan." The Verge. March 2019. https://www.theverge.com/2019/3/5/18250801/lead-contamination-flint-michigan-water-testing.

Gross, Courtney. "NYC reveals new plan for critical NYCHA repairs" Spectrum News NY1. November 2018. https://www.ny1.com/nyc/all-boroughs/politics/2018/11/19/

nyc-reveals-new-plan-for-critical-nycha-repairs-involving-rad-section-8.

Hill, M.. “Billion dollar fix for massive nyc water tunnel”. Associated Press. June 2018. https://www.youtube.com/watch?v=jpblw6tUdSs

Hu, Winnie. “A billion-dollar investment in new york’s water”. New York Times. January 2018. https://www.nytimes.com/2018/01/18/nyregion/new-york-city-water-filtration.html.

Jorgensen, B. S., J. F. Martin, M. W. Pearce, and E. M. Willis “Predicting household water consumption with individual-level variables”. Environment and Behavior 46(7), February 2013: 872–897.

Keohane, N. O. and S. M. Olmstead. Markets and the Environment. Seattle, USA: Island Press,Center for Resource Economics, 2006.

Koeppel, G. T. “Water for Gotham: a history”. Princeton, USA: Princeton University Press, 2001.

Lee, D. and S. Derrible. “Predicting residential water demand with machine based statistical learning”. Journal of Water Resources Planning and Management 146(1) (January 2020).

Lowenstein, Ronnie. “Lead in tap water remains issue in some city dwellings”. New York City Independent Budget Office. September 2018. https://ibo.nyc.ny.us/iboreports/safe-at-home-lead-in-tap-water-remains-issue-in-some-city-dwellings-2018.pdf.

Mansur, E. T. and S. M. Olmstead. “The value of scarce water: Measuring the inefficiency of municipal regulations”. Journal of Urban Economics 71(3) (2012): 332–346.

Nelson, Libby. “The flint water crisis, explained”. Vox News. February 2016. https://www.vox.com/2016/2/15/10991626/

flint-water-crisis.

“Drinking water monitoring”. New York City Environment Protection, 2019. https://www1.nyc.gov/site/dep/water/water-monitoring.page.

New York City Housing Authority. “Design guidelines - rehab of nycha residential units”. 2017. https://www1.nyc.gov/assets/nycha/downloads/pdf/nycha-design-guidelines.pdf.

N. Y. C. H. Authority. “About NYCHA”. 2020 https://www1.nyc.gov/site/nycha/about/about-nycha.page.

N. Y. C. H. Authority. “About section 8 - NYCHA”. 2020. https://www1.nyc.gov/site/nycha/section-8/about-section-8.page.

N. Y. C. H. Authority. “FHA homes program”. 2020.https://www1.nyc.gov/site/nycha/about/homeownership-program.page.

New York State. “Water use & conservation”. New York State Department of Environmental Conservation. 2019. http://www.dec.ny.gov/lands/67073.html.

New York Times. “A Billion Gallons a Day — Living City — The New York Times”. New York Times. October 2014. https://www.youtube.com/watch?v=vIKFpQhVqI4

Mayor’s Office of Sustainability. “Energy and water data disclosure for local law 84 2019 (data for calendar year 2018)”, November 2019. NYC Open Data. https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/ vdzd-yy49.

Pasion, C., C. Oyenuga, and K. Gouin. “Inventory of new york city greenhouse gas emissions”, New York State Department of Environmental Conservation. April 2017. https://www.dec.ny.gov/docs/administration pdf/ nycghg.pdf

Raymond, T. and N. Streeter. “The economics of water”. Abbot Downing. 2013 https://www.abbotdowning.com/ asset/dp572n/The-Economics-of-Water.pdf.

Rivera, Ray, F. G. Runyeon, and R. Buettner “Inside city’s water tanks, layers of neglect”. New York Times. January 2020. https://www.nytimes.com/2014/01/27/nyregion/inside-citys-water-tanks-layers-of-neglect.html.

Ro, Christine. “The hidden impact of your daily water use”. British Broadcasting Corporation. March 2020. https://www.bbc.com/future/article/20200326-the-hidden-impact-of-your-daily-water-use.

Rueb, Emily S. “How New York gets its water”. New York Times. March 2016. https://www.nytimes.com/interactive/2016/03/24/nyregion/how-nyc-gets-its-water-new-york-101 .html.

Runyeon, F. G. and M.-S. Schwarzer . “Inside new york city’s water towers — the new york times”. New York Times. January 2014. https://www.nytimes.com/2014/01/27/nyregion/inside-citys-water-tanks-layers-of-neglect.html

Hans Schreier (University of British Columbia Professor Emeritus, Land and Water Systems) in discussion with the author, January 2020.

Energy Star. “Dishwashers”. United States Department of Environmental Protection. 2020. https://www.energystar.gov/products/appliances/ dishwashers.

Energy Star. “Make laundry better”. United States Department of Environmental Protection. 2020. https://www.energystar.gov/products/ appliances/clothes washers.

The Editorial Board. “New york’s public housing isn’t getting better”. New York Times. July 2019. https://www.nytimes.com/2019/07/30/opinion/new-yorks-public-housing-nycha.html.

Veiga, Christina. "As nycha public housing crumbles, the child care centers inside go without crucial repairs". Chalkbeat New York. February 2019. https://ny.chalkbeat.org/2019/2/21/21106954/as-new-york-city-s-public-housing-crumbles-pre-k-centers-go-without-crucial-repairs

Berkeley
Economic
Review