I looked over the impact assessment paper and glanced over some other sources to get a sense of how we would analyze the effect of FDI on Ukraine. Below is a rough sketch of how this would work:

<https://docs.google.com/document/d/1--nHVC6rz86FWayrrEah6sRh_aSbm00OwslJeQ_BUho/edit?usp=sharing>

TLDR: I think the multiplier effect method is great. IMPLAN doesn’t seem to have Ukraine coverage, so asking for government input-output models is a good idea.

Data sources needed:

1. Input-output models or some type of CGE that covers Ukraine (to estimate regional economic effects)
   1. or use a typical CGE model like IMPLAN and scale according to the specifics of the Ukrainian economy (not sure how this would work)
2. Ukraine airport data: breakdown of yearly expenditure including payroll for employees
   1. or general numbers for airports that we can scale according to Lviv/Kyiv’s specifications

Things to figure out:

1. Which input-output models we will use
2. What methodology is typically used to net out non-effects / calculate opportunity cost of money that would be used otherwise

Super interesting thoughts—thank you for sending!

There’s a lot here, but to address the tax revenues—collection has changed a lot since the war started, which makes things even more interesting. Will have to research more & feed this into our CBA.

Quick articles for tax collection details:

<https://www.reuters.com/world/europe/ukraine-plans-restore-pre-war-tax-system-by-july-finance-minister-2023-06-22/>

<https://taxfoundation.org/blog/ukraine-tax-reform-economy/>

I will research more into economic impacts of airports closer to the end of this week.

Hello!

Did a little bit of looking into the literature about economic impacts on airports.

<https://docs.google.com/document/d/1V4WZtYE92lipy6-YRHasS8vanoUwpaAufjpuLT8aVYI/edit?usp=sharing>

Not conclusive, just exploratory, but I think it provides a little more insight into what we might want to consider doing.

Looked at Marquise McGraw’s work. It looks like there are only two key papers, which are both semi-relevantl.

The role of airports in city employment growth, 1950–2010 [link](https://www.researchgate.net/publication/339609242_The_role_of_airports_in_city_employment_growth_1950-2010)

* Synthetic control
* Data = newly digitized aviation records
* Finds effects on employment and income

Causal Effect of Airport Hubs on Urban Growth [link](http://www.marroninstitute.nyu.edu/uploads/files/McGraw_hubs_paper.pdf)

* Panel regressions and event study techniques
* Less relevant, more about hub airports (which ours will not be)
* Finds industry-based growth effects

I can keep digging or we can ask him ourselves if he has any advice on methodology.

For the synthetic control paper, if we want to go this direction, I can look further into the exact synthetic control he did. Chances are that I can do it too, given we have appropriate data.

Contents of this email:

1. Important question
2. Notes on McGraw SCM paper
3. **PLEASE READ** — proposed SCM procedure

1.

**QUESTION:** Would the SCM be used for just modeling FDI or other dependent variables in our study (military investment, tax revenue, direct services, morale boost, etc?)

2.

Below are some thoughts on Marquise McGraw’s SCM paper:

* **Training set:** Controls were “alternative” and “proposed” landing fields as identified by the newly constructed dataset he used.  – we have to decide what our training set will be (Ukrainian regions? Or world?)
* **Independent variables:**Synthetic control was trained on population, employment (multisector: tradable and untradable, from PUMS, City and County Data Book, NHGIS, County Business Patterns database), geography, climate, and human capital characteristics
  + We can probably use whatever data we have at hand but make sure that we cover all these bases
  + Things to consider: presence of a college in the area, port, river access, tourism likelihood
  + **We need to add a “safety of flight” or “safety of destination”component,**which will be the critical component to model FDI outcomes. Maybe this means we need to include in the synthetic control places around the world in various states of conflict… or maybe just Ukrainian regions at different times (how did airport usage change regionally in Ukraine after 2014?)
* **Concern about small sample size:**In Marquise’s study the data was very numerous and there were so many different airfields and airports included. Here, we’ll have to be careful to make the data as strong as possible.

3.

**PROPOSED ORDER OF OPERATIONS — TRAIN ON UKRAINE REGIONS:**

* Collect baseline data on Ukrainian regions at the smallest level possible including geography, climate, population, demographic, employment, industry…
  + Code each Ukrainian region with “airport presence” dummy and/or “airport size” variable
  + Code each Ukrainian region with “safety of flight” or “safety of destination” component (ideas?)
  + Find detailed data for our dependent variable, FDI (ideas? Data might be hard to come about)
    - Typically a study design like this would be focusing on domestic development, but we’re curious about money flowing into the country at large. Maybe this is a reason to widen our training set to the world at large?
* Generate a synthetic 2024 Kyiv and 2024 Lviv
* Simulate the effects of airport vs. no airport

Let me know thoughts and next steps.

**Good news/update:**Found some good global (all countries) data for a start to the SCM. Including:

* ACLED conflict index data – downloaded for years 1994-present but highly disaggregated. might be able to aggregate as a proxy for “danger level”, we’ll see what I can do, might be overly complicated
* Plenty of economic indicators as a basis including: New businesses registered (WB) + FDI (world bank) + some other economic indicators (also WB) + exchange rates inclusive 1994-2024 (IMF)
* List of all global airports and volume, Jan 2023

**Current problem:** Struggling to find good dataset showing flights. Most all-inclusive datasets cost thousands of dollars for just a year of data. Some countries are supposed to have free data (Brazil, India, etc but I haven’t verified yet). **Any ideas for where I can find reliable and free air traffic time series, whether country-specific or global?**

The best I could find is the statistical analysis by month. <https://www.icao.int/sustainability/pages/air-traffic-monitor.aspx>

It seems like the full data of air traffic at airports is not publicly available. Might be useful to ask for the dataset, because until we get the travel advisory info, it’s a little difficult to run an unbiased analysis with a “non-operational airport” treatment. If asking for the data is not really possible then please let me know.

The GT analogue sounds fine, I’ll look into that, but I would obviously rather have the flights or advisory data since they would not be perfectly correlated indicators, and I’m not sure how I would elegantly filter out those sorts of anomalous events. Has there been any progress on the data front for either flights or travel advisories? I’m building an SCM dataset, and it’s almost ready for that variable to be added and to generate the SCM.

All: I have yearly data compiled. SCM framework itself is set up, but I'm having some trouble with NAs. That can be figured out after there's more clarity on which data we're using. I realized I should have also grabbed monthly data. Should I disaggregate to month, day, week, etc? I have ACLED (violence numbers) aggregated to day, month, and year, but not the World Bank econ indicators (but I can fix that today if needed).

<https://www.bts.gov/browse-statistical-products-and-data/bts-publications/data-bank-28im-t-100-and-t-100f-internationa-0>

I spent a few hours trying to make git work, and the issue is that there are some large files that are messing things up, even if I remove them, there is something configured that is preventing the upload. In the interest of time I think I’m going to move on without a working github, but if this is important enough or there are certain files you would like me to share, I can do that.

I popped my R project (without the large datasets) into a google drive. If you download the whole thing, it may work the same as git.

<https://drive.google.com/drive/folders/1aWE8uHWehkzHYkhUEF-UIRIQb_Y1fTOj?usp=sharing>

Just wanted to let y’all know that. Moving forward now.

\_\_

I’m extremely close to getting FDI estimates, but I lack data on 2023 Ukraine economic indicators. WB lacks global data after 2023 but also lacks Ukraine data after 2022. **Does anyone have at least year estimates for any key economic variables for Ukraine in 2023 (GDP, FDI, etc)?**I will input them manually into WB time series and be able to generate something by tomorrow night hopefully.

Update on methodology now: Because 2024 numbers aren’t very viable for any of the data, and I don’t want to generate “fake Ukraine” numbers for 2023 by just copying and pasting 2022/23 numbers, for the SCM to work we would have to generate estimates from a simulated case in which the airports opened on 2023, generating the synthetic control from 2023 data. **(Not a projection into the future, but a “what if” in the past). I’m generating an FDI estimate for the case that Ukraine opened its airport(s) again on January 1 2023, and assessing how FDI, GDP, tax revenue, etc would have changed. Please let me know if this sounds incorrect or not useful.**

It's not a quantile regression, but uses SCM, assuming this works. I can work on quantile later, since it’s a new methodology to me, but this should be simpler and faster.

Bumping this since I sent it late last night, and I don’t want it to sink to the bottoms of your inboxes!

I could just look up indicators individually — I see that there are estimates for GDP, PPP INT$ ($531.44 billion), tax revenue (UAH 1,098.5 billion), mid-year population (36,744,634), which would make GDP per capita ($14463). Unless there are any official estimates I could use, would it be appropriate to just piece it all together (and take heavy lab notes to account for these?)

Thanks so much for your responses! I will look through and see if they will work. Thank you also Peter for asking around to get more official figures.

To answer abt the WB data: I just queried a few indicators from here:

<https://databank.worldbank.org/source/world-development-indicators>

The ones of interest would likely be GDP (PPP), population, industrial breakdown (service%, tax%, agriculture%, etc), FDI.

If anyone has time to weigh in:

* The IMF numbers 2021-2023 (sent by Peter to cover 2023 in Ukraine) differ significantly to WB numbers. I am just using IMF numbers for 2021-2023 and WB numbers for <2021. Is that ok? (The 2023 numbers were supposed to be used both for synthetic control independent variables and for comparing against the projected increase if an airport were there).
* Differences are like 10% versus 30% for some variables. They are not proportionate to each other.

I tried for no less than 10 hours just today to make the synthetic control work. I specified everything correctly. I just cannot make it happen. I think it’s because for some reason the synthetic control refuses to converge—the numbers are too far from each other. Secondly, it must be something innate to the code that I can’t seem to figure out, that these NaNs are being produced.

If anyone has any clue what is going on here or what else I can do to troubleshoot, that would be awesome. Specifically the “Error in if (m <= tol)” part. I cannot find this anywhere in the underlying function code, and it’s not in my code either. I don’t know what the “In sqrt” parts are either.A white background with red text

Description automatically generated

I think we need to pursue an OLS or quantile. I started an OLS in stata but for some reason it just will not work. I will try this more tomorrow but I have spent literally an entire day doing this today and it is not in my best interest to try to do so again. But of course I will keep working on this until we have *something* by Tuesday*.*

**I can give the dataset to anyone else who wants to start trying to throw some regressions on it, even if they’re SUPER basic.**Everything in the data is clean and tidy! I have a complete version and a version with no NAs, and I can explain them all. The more of us work on this, the more likely we can get an estimate by Tuesday.

Also, if someone else can run the regressions, I can devote my time toward putting more of the Ukraine 2023 numbers into the dataset so that we can regress with them, or toward troubleshooting the SCM more.

I’m so sorry that I couldn’t pull this together. It really wasn’t for lack of trying. I made a LOT of progress with the data, though, and I really gave it my all with the SCM.

I think an OLS would definitely work, just for some reason it didn’t work with the data I was working with when I was doing it. I think that if someone gave it an hour they could probably make a basic OLS work, especially if they’re more experienced than me with running basic regressions with lots of variables. (I will try to do this tomorrow, too, but for a much briefer time).

Updated the contents of the research folder:

<https://drive.google.com/drive/folders/1kGlS3dE9fvfwQEH0p3fUqYB5Gn7jPdX7?usp=sharing>

master\_dataset.csv is the comprehensive dataset including NAs and lot of variables, in long form for easy manipulation

scm\_data.csv is the non-NA, very cleaned data with limited variables, in wide form

SCM.Rmd is the part of the code that I did today trying to make the SCM work. It includes the code used to generate the scm\_data.csv.

I can do my best to get the stats started and try to work on doing ln() for the SCM, but I honestly am very squeezed for time to do everything on my own, so I would prefer if others also help with the stats so I can focus on trying to fix the SCM.

The data set is already in the cloud (see message sent last night -- it's the Google drive link, and I list the datasets below). I assume for basic OLS and descriptive stats, we can just use the full dataset (data\_master.csv), but the tidied dataset (SCM\_data.csv) has completed cases whereby I imputed values for 2023 carried forward from 2022 for all countries, 0s listed for all NAs in ACLED and the DOT flight statistics, and so on.

I can describe more and answer any questions as needed.

\_\_

I lacked GDP numbers in real terms because it was in hryvnias, but maybe we can get the exchange rate for each year for hryvnias to generate a “total FDI in real US$” measure and we can run that?

The problem is also that I would have to carry forward the 2022 numbers for GDP in each country to fill in 2023 numbers, except for Ukraine, but I think that would have a minimal effect.

I can try to make this happen if you think it would be worth it.

Got different values than Peter for the quantile regression. In the process of taking the log of the ‘passengers’ variable, all 0 values are coerced to -Inf. I set these to 0 and got a different table:

Mine (with -Inf set to 0):

A white background with black numbers and letters

Description automatically generated

Peter’s (with unknown handling of -Inf):

*A screenshot of a computer

Description automatically generated*

I ran it without setting -Inf to 0 and got an error. Peter, did you use a different method to handle undefined values? I can’t observe any other differences in our code.

Otherwise, do y’all think it’s ok to coerce all undefined measures to 0, when taking the ln?

Thank you so much for the pointers @both. I ended up going with the log1p() approach.

See first knit Rmd for progress [first\_quantiles]. I ran quantile regressions for the original specification with no weights (1) weighing by log of battle fatalities (2), log of fatalities from violence against civilians (3), and log of total number of violent events (4).

The reason I am using weights is because there is no way to specify a control variable in rq(), only in non-linear regression functions which I’m not sure would be as parsimonious here. I’m happy to look into these more if you think the weight parameter is not sufficient.

Looking now in the documentation for how to display significance — for some reason it’s not obvious.

Attached is progress thus far, including regressions with leads/lags and with filtering for >100 battle deaths.

[quantile-analysis-2]

Figured it out. Here are 15 quantile regressions with specifications listed below the outputs. Let me know if you need me to explain anything.

If I have time/energy tonight I’m going to try to transform these percentages into $US total FDI.

 [exported\_tables] – march 3 – IGNORE

Attached are three excel spreadsheets. All three include quantile regressions with specifications listed below the outputs.

* “exported\_tables\_total\_FDI.xlsx” is regressed with net inflows of total FDI in terms of current $US.
* “exported\_tables.xlsx” is regressed with net inflows of FDI as a % of GDP.
* “exported\_tables\_total\_FDI\_w\_controls.xlsx” is regressed with total FDI in terms of current $US, but uses various control variables by running multivariate regressions (I figured it out finally at 1:30am—stupid mistake).

Let me know if you need me to explain anything here.

Attached are also a couple jpegs. Of interest are especially plots 4-12. They show exactly what we wanted to show: that ***no matter if you weigh the results by conflict or control by conflict level*,** **the positive relationship between passengers\_in and FDI increases as passengers\_in increases. (**This is more variable when using FDI in terms of %GDP, but very very straightforward when measuring in $US FDI.)

There are plenty of other interesting relationships which you can poke through as curiosity strikes. Let me know if you need anything explained—I am free sporadically all day.

I have also attached my Rmd to generate these, but it is thousands of lines long and may not be very comprehensible.

Let me know if you also want me to generate more plots (I only did what I thought would be the most important regressions that you may want to present, Prof. Becker).

One more addition about controlling for GDP:

Even when you control for GDP on top of everything else, the relationship holds, yet is less significant. (the coefficients become more positive for countries/years with the highest decile of ln(passengers\_in\_total), too).

Plot and updated “exported\_tables\_total\_FDI\_w\_controls.xlsx” attached.

Updating on use of “percent change” variables.

**TLDR**: Regressing total FDI on ln(percent change in passengers flying into the country) produced significant and positive effects. See plots 20-25 and “exported\_tables\_pct\_changes.xlsx” for coefficients (attached). I tried other combinations (outlined below) and all brought less significant results or dead ends. [Please use this drive folder to access any previous coefficients—I made tweaks](https://drive.google.com/drive/folders/1aWE8uHWehkzHYkhUEF-UIRIQb_Y1fTOj?usp=drive_link). Contact me 11:30-1:15 for anything needed.

1. PASSENGERS PERCENT CHANGE: Not the most promising results. Plots 14-19 on google drive. Not very significant, but also shows a negative relationship between passenger increase and FDI at the highest decile of passenger.in.change.pct.
2. \*LN+1(PASSENGERS PERCENT CHANGE): Very promising(plots 20-25 attached)and also significant. Note that these also drop NAs that are created from logs of negative values, therefore dropping big decreases in passengers flying into the country (essentially chopping off the lower end of the deciles). Inference should remain unaffected.
3. PERCENT CHANGE IN TOTAL FDI: Adding this as the Y in the previous regression with ln(passengers percent change) didn’t produce that many significant effects. Points me toward taking the ln() of this as well.
4. LN+1(PERCENT CHANGE IN TOTAL FDI): produced many NAs, which is a problem because we can’t ignore this change like we can in (2) due to the x axis deciles.

**If you need a hand tomorrow:** I am free sporadically all day, but I have a free block 11:30-1:15. I will also be tuning into the Zoom at 9am and during that block (11:30-1:15).

Small note: found an error in my code where I forgot to divide total FDI estimates by 100. Really stupid of me, I’m glad I caught it. This is why you don’t rush these things at 2am :) [Here is the link to my google drive](https://drive.google.com/drive/folders/1aWE8uHWehkzHYkhUEF-UIRIQb_Y1fTOj?usp=drive_link) with all the updated code and tables—coefficients and significance levels are different, so if you are quoting these, please take a look at the new estimates.

Questions:

1. **For ‘robustness checks’**, this just means the lags/leads and weights that I tried, right? Anything else that you assume that I should be investigating?
2. **For cointegration testing**, I agree that I need some guidance. Any code snippet you can provide would be helpful, just to give me an idea of what you would like me to do, but I can take a stab at this on my own first if you think that it wouldn’t be that hard to figure out.

1. I think I remember there being code examples available for diff-in-diff during the Warzone Data Science class last year, so I could pull that up and put something together based on that. But **I would like more guidance on the specifications that you both have in mind with the diff-in-diff model. What is the treatment/non-treatment?**Happy to call anytime today if it’s difficult to explain in email (I’m on break and working on things until 4pm/5pm today, but can do after as well).

1. **Any news on SCM?** After DID maybe I can look into this from another angle/package (but only after I finish the other work, since troubleshooting this may be in vain).

1. Deadline is Friday, I guess? Or Thursday night?

Attached is a distilled write-up (the page count is inflated because of the figures and tables included). Let me know if anything is unclear.

Good news: fit the model to the data. Bad news: the treatment is super not significant (but the post-treatment is getting there).

A screenshot of a computer

Description automatically generated

Here is how I specified the treatment/post-treatment variables.

A screenshot of a computer program

Description automatically generated

1. I feel like we would be able to generalize fine at that rate because Ukraine had only one major airport, right? It’s still dicey because the airports may not be proportionate to FDI/GDP and so on.
2. There’s the additional issue of the fact that this is like “top 50” or “top 100” airports so sometimes the largest airport in a country (like Slovakia, for example) escapes to the bottom of the list and then resurfaces. So maybe I can just exclude airports like this, and airports like this that are part of larger countries—they would have to be available for the entire time series.
3. We will end up with a small sample size, maybe ~60 countries, mostly in Europe. But that might be fine.

I literally downloaded and cleaned the data for 30-40 WB development indicators from this exact database, I don’t know how I didn’t see this sooner! I guess I didn’t think that aviation data would be considered development-y enough.

Model (based on N=940, with only a few treatment cases) is not significant, but I am troubleshooting code for why a bunch of observations seemed to have been dropped even though I thought I accounted for them. It’s coming up with a cryptic error so I’ll need to spend more time.

A screenshot of a computer

Description automatically generated

Additionally, we made the decision to consider everything “treatment” after the largest spike in the country’s time series, as long as the spike was larger than a 100% increase (doubling). This means that there are “1”s listed for many years that are NOT increases. Is this still what we want to do?

This was to get around the “multiple spikes” phenomenon and keep the post\_treatment variable.

I’ll try with just a basic treatment variable that marks if there is a doubling at all, and then maybe a logged treatment. That will likely be more accurate.

Taking the log of FDI was smart in the first place! I got a significant value for the treatment!

A screenshot of a computer

Description automatically generated

To clarify, this is taking the log of TOTAL FDI which I was not doing in the first place, and it makes total sense to do so, so I should have guessed to do it off the bat.

When I put a log on GDP, it actually removes the significance flag from treatment itself, but the interaction term between treatment and ln\_battle\_fatalities gains a significance marker!!!! Which proves our point I think?

A screenshot of a computer

Description automatically generated

\_\_-

Thanks. I’m not totally satisfied because I’m not sure if the model specification is entirely perfect but this seems to be pretty good stuff. I also couldn’t quite get a scatterplot with the models to plot properly, but that can be figured out later.

For now, attached is the specification I’m using. Curiously, taking out the post\_treatment term makes the treatment itself become significant at the 5% level. Would appreciate if you or others could look through as a sanity check.

Note: don’t quote the screenshot I sent earlier. I removed the treatment/post\_treatment interaction for the word document attached, which slightly changed some coefficients.

I can do quartiles/residuals sometime else this week, but I might also take the opportunity to get ahead on other work for right now.

[Ukr\_mar19]

\_\_

I realized that I asked for advice on the work I sent previously but did not receive any comments. Therefore my write-up pretty much mirrors the email. Feel free to dress it up as needed, and please see the comments and questions on it.

<https://docs.google.com/document/d/1qA3vkUPPJmdINvaQegGAT7YEw-lyShCItlDLYi4W31M/edit?usp=sharing>

Prof. Becker, also wanted to remind you of the report I typed up for Peter which may be useful here as well: <https://docs.google.com/document/d/12pm2a_w1NJxXOc6roihjBqka8E8nHucOhfh4DjjRp1k/edit?usp=sharing>.

Hi,

Hope you are all well. I ended up catching my breath enough to fix the data like I said I wanted to do earlier this week before we submit anything. Some observations changed but inferences thankfully stayed largely (completely?) the same. I would love to discuss these with y’all sometime soon.

Since some big countries weren’t carrying over due to the naming error in my code, the coefficient for the final tau has changed for some plots. It’s difficult to describe through email. But many (all?) of the inferences remain, eerily and thankfully, the exact same. So I just wanted to give a heads up. Example: attached screenshot and plot from X1 = ln\_passengers\_pct (log+1), X2 = ln\_battle\_fatalities (log+1), and Y= total\_FDI\_USD. Coefficients between the middle 30-90% of X1 are significant. I think this is the same as the writeup I initially wrote. I will need to review and update this and resend.

Regarding the DID, it changed slightly and actually became more significant! (At the 1% level!)

Still not sure if this is DID, since I think it just amounts to a linear regression since I had to exclude the interaction term for treatment\*post\_treatment (it was just generating NAs, the way it was specified).

A screenshot of a computer program

Description automatically generated

Also, I was pretty rigorous about making sure it was a complete dataset with no NAs, so I’m unsure about the “missingness” argument here in the screenshot. But this feels right to me.

Pretty sure this is the final edit I will make to the analysis, unless y’all spy anything else I should look into. Apologies for the delay—I meant to do this last weekend but life got away from me.

\_\_

Yes, it didn’t have the term before and I tried to point this out before, so any analysis that considers it diff-in-diff may have incorrect conclusions (but I’m not sure how you wrote it up—I haven’t had a chance to look through what you put together).

Treatment is the maximum spike in a country’s passengers, given that it doubled in that observation (100% increase in passengers flying in).

Checked with my econ major friend in the room and she said that my idea of the post\_treatment is correct. But this is what happens when I put in the interaction term (I feel like I’m missing something…?):

*A screenshot of a computer

Description automatically generated*

To clarify, originally I did not include post\_treatment. The one that is just including treatment is just a linear regression.

Included post\_treatment in the previous email to illustrate why I excluded it.

Sounds good. Peter, attached is some new output, as agreed upon during our last meeting. If you have time, I would love for you to glance it over. My remaining concern is that although slapping logs onto all our variables can help the scale issue (since battle fatalities will have a different scale to FDI, GDP, and passengers), this won’t solve it entirely, and may bias our results.

[4:5 update]

\_\_\_

Hi,

Update:

* After fixing all the models as we decided, they all produce the exact same results:
  + Insignificant coefficients
  + Positive coefficients, varying largely (except for one, which is also the least significant).
  + Joint significance is present (I’m trying to verify that I am reading this correctly… it’s just a significant F-stat right?)
* Why did it become so insignificant?
  + We’re using real instead of nominal FDI, removing the risk of spurious correlations.

TLDR: the evidence is looking weak on real FDI, even when I look at only Europe. I think I need to look into fixed effects, and then we’ll see if any effects pop up when I restrict the dataset to EUR and middle income.

* I should also include the nominal FDI models just to show that we did produce significant results… or use some other way to calculate nominal GDP

Summary of the models I ran:

A white background with black text

Description automatically generated

Best,

Michelle

\_\_\_\_\_

Hi Michelle,

Agree with everything you say.

Bottom line: we report what the data give us. The joint significance is key and we'll run with that. Though including nominal FDI models also makes sense.

best

charlie

Hi,

Dropping a quick note as I work on this (let me know if you have any reservations):

New problem: some FDI is negative (for instance: Austria in many years after 2010). Logs have been removing them and distorting the results. Solution: [IHS transformation](https://worthwhile.typepad.com/worthwhile_canadian_initi/2011/07/a-rant-on-inverse-hyperbolic-sine-transformations.html) instead of log on anything with negatives. IHS-transformed variables should be compatible alongside logged variables, and they have slightly different interpretations (but very similar because of calculus).

Nearing the end of the modeling portion.

Best,

Michelle

Apologies for the double email. Descriptives with the IHS transformation show the limitations to this method: the function bottoms out at ~25 and ~-25 due to the shape of the function itself.

A graph of a function

Description automatically generated

Potentially this means that we remove shades of meaning from variation in FDI by adding the IHS transformation. Any thoughts about whether this is bad or harmful? Or would it just inflate the p-value? The interpretation should technically remain kosher.

Would it be a bad idea to just run everything with the un-transformed FDI variable?

Best,

Michelle

Hi Michelle,

I like the IHS transformation! And its convexity/concavity shape makes intuitive sense – marginal effects at extremes are lower.

Best

Hi,

Awesome to hear. It looks like it doesn’t change much but in some very curious permutations of the model.

Attached is my powerpoint describing my newest model outputs. [Also find on google slides.](https://docs.google.com/presentation/d/1OWXi38K2n6R4JVgxvt-H-9IQ18HnHe3lmmYaoqMr7ac/edit?usp=sharing)

I schedule sent a message to the group including Aaron for 8am EST, but I wanted to send this along to you just in case you wanted to glance it over tonight!

Best,

Michelle

[LOOK AT THE POWERPOINT]

Hi Michelle

Great! I'll get up early and look it over – kind of bleary eyed right now

...but, of course, I had to peek. Quick note; any time you get significant results with a coefficient of 0.000. Then you need to rescale (divide by 1000 or 10,0000).

This is all really helpful. To me, it points to endogeneity of re-opening. We can think over the summer about how to handle it (and discuss in the morning).

All the best

charlie

---  
Charles M. Becker  
Research Professor of Economics

Duke University  
<https://sites.duke.edu/charlesbecker/>

**MSEC homepage:**[**http://econ.duke.edu/masters-programs/degree-programs/msec**](http://econ.duke.edu/masters-programs/degree-programs/msec)

\_\_

Hello,

Case study graphs as promised. Shows how weak the data is. I’m trying to remember what dataset I ended up pulling from after the wild goose chase to find data. I think it was either an IMF or WB set, and not the ascii data that I downloaded from the DOT? I will backtrack and locate exactly what I did.

But either way, the patterns in the data look like they only reflect US carriers. Explains something about the trends, maybe.

Ethiopia looks like the most promising time series. Maybe we can focus on this, if there is an “airport reopening” moment.

Best,

*Michelle Schultze*

Duke ’25 | Economics, Political Science, Russian

\_\_\_

Hi Michelle,

These must be American data. I think you got global data from the World Bank site.

Here's a link to current restrictions: <https://travel.duke.edu/restricted-regions-list/>