Optimal Mobility in the Context of Refugee Settlement in Jordan

Konhee Chang

May 11, 2021

Abstract

What is the socially optimal level of mobility in the presence of large heterogeneities in mobility frictions? This paper studies this question in the context of Jordan, where more than 10% of the country's population is composed of refugees of the Syrian civil war. First, I study the impact of the refugee influx on the local economy. Then, using high frequency cell phone usage data, I estimate different levels of income elasticity of migration for Jordanians and refugees using a gravity equation. Preliminary analyses suggest that refugees eventually settle in high-income urban areas, but that their mobility is restricted compared to the incumbent population. (These findings motivate the spatial general equilibrium model where the social planner faces a trade-off between subsidizing mobility and places (i.e., refugee camps).)[1]

[1] Data for this paper come from an ongoing joint project with Michael Gechter, Nick Tsivanidis, and Nathaniel Young.

1 Introduction

What is the socially optimal level of mobility in the presence of large heterogeneities in migration elasticities or migration costs? More broadly, should governments subsidize migration or subsidize poor places? To answer these questions, I turn to present day Jordan where more than 10% of the country's population is composed of refugees of the Syrian civil war.

Around half of Syrian refugees transit through refugee camps where their mobility is restricted by government rationed permits. Due to this, there exist large heterogeneities in mobility frictions across the population. The government subsidizes refugee camps and face a trade-off between subsidizing refugee camps and potential externalities from refugee influx. To motivate a spatial general equilibrium framework to study the welfare effects of this trade-off, I first perform the following two sets of empirical exercise.

First, using administrative data, I study the impact of the refugee influx on the local economy. Preliminary findings show that refugees who settle outside of camps do so in dense urban areas. A simple difference-in-differences framework suggests that neighborhoods that experience a large magnitude of the refugee influx have a smaller growth in rent prices and economic activities relative to neighborhoods that experience a smaller magnitude of the refugee influx.

Second, using high frequency call details record (CDR) data that contain rich information about cell phone usage, I estimate different levels of income elasticity of migration for Jordanians and refugees using a gravity equation. Preliminary analyses suggest that refugees have a smaller migration elasticity, suggesting that the mobility restrictions are binding.

(Last, motivated by these empirical findings, I develop and calibrate a spatial general equilibrium model where the social planner faces a trade-off between subsidizing mobility and places (i.e., refugee camps).)

This paper contributes to the trade and development literature on migration elasticities (Allen and Donaldson 2020, Bazzi 2017, and Méndez-Chacón and Patten 2021) as well as literature on the impacts of forced migration and refugee settlement (Alhawarin et al. 2020, Beine et al. 2021, and Rozo and Sviatschi 2021). More broadly, this paper is related to the public and urban economics literature on place-based policies (Gaubert et al. 2021). Empirically, this paper connects to recent literature that exploits CDR data for economic research (Beine et al. 2021, Björkegren 2019 and Blumenstock et al. 2019).

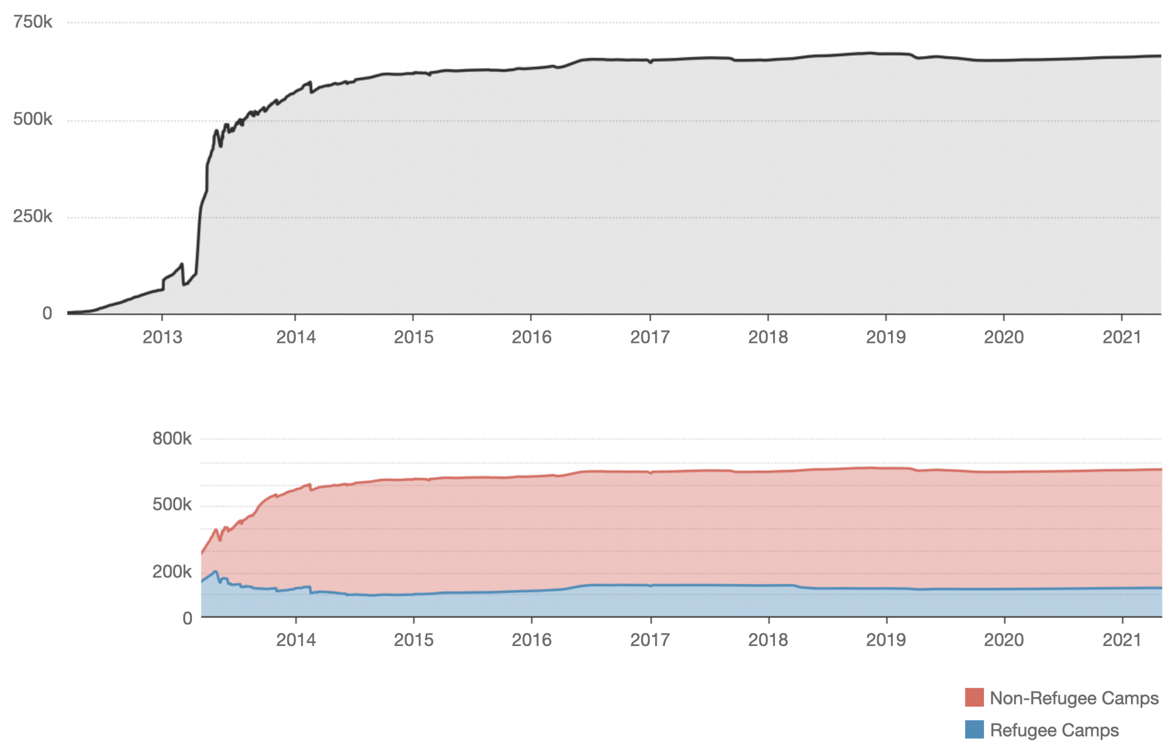
2 Background

2.1 Timeline of the Syrian Refugee Crisis in Jordan

The onset of the Syrian civil war in 2011 resulted in the largest refugee crisis in modern times. Jordan admitted the second largest number of Syrian refugees amounting to an estimated 1.3 million individuals as of 2021. Figure 1 shows the trend of registered Syrian refugee population from 2012 onward. Note that the series only capture registered refugees. Also, note that around 80% of refugees currently reside outside one of three refugee camps.

To accommodate the large refugee population, the Jordanian government established three refugee camps — Za'atari, Azraq, and Emirati — on the Northeastern region of the country. Figure 2 show the geographic location of these camps as well as Amman and Irbid, which are the two largest cities in Jordan. Za'atari, which opened in July 2012, is the largest among the three, accommodating up to 150,000 individuals at one point. Figure 3 shows the growth of the refugee camp over time.

Figure 1: Refugee Trends



2.2 Mobility of Syrian Refugees

While a large fraction of refugees reside outside refugee camps, migration outside camps is officially governed by government issued residential and work permits. Furthermore, movement is further restricted as permits are required for temporary trips as well.

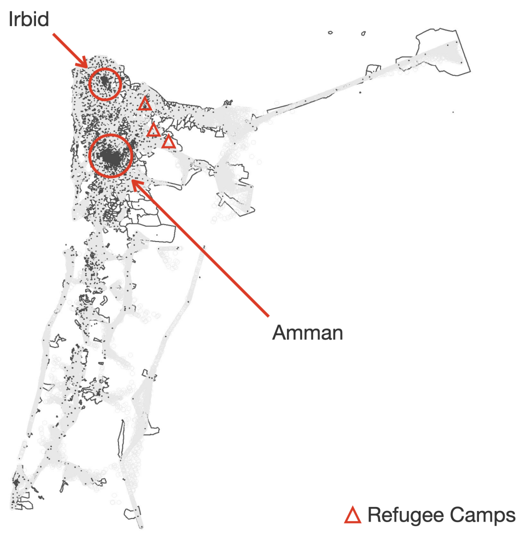
3 Data

3.1 Call Detail Records (CDR) Data

The main micro-data I use for analyses are the Call Detail Records (CDR) data. The dataset comes from one of two largest telecom network companies operating in Jordan, and beginning June 1, 2015 through the present time. The CDR data contain the universe of cell phone usage information from all users in the network. Specifically, each time a user sends or receives a voice call or text message, or uses data service (i.e., referred to as a "transaction"), the information about that transaction is entered into the CDR data.

Figure 2: Maps of Jordan

(a) Modal (Cell Towers) and Hartigan's Locations



(b) Night Light

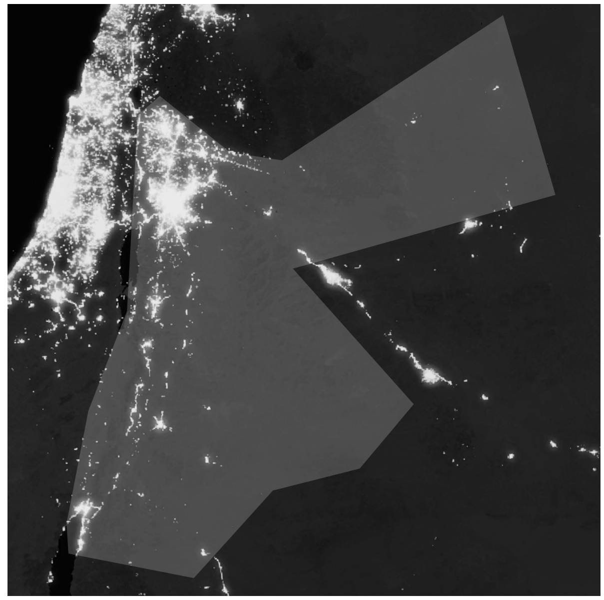
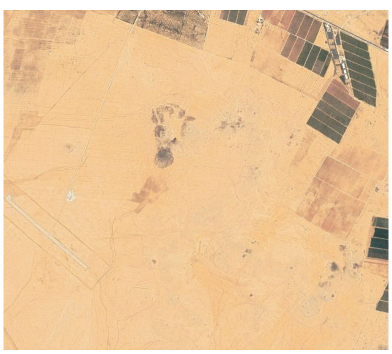


Figure 3: Za'atari Refugee Camp

(a) 2011



(b) 2013

(c) 2021

Table 1 shows a fake snippet of the CDR data. Table 2 shows the number of transactions and the number of unique users observed on June 1, 2015.

The `call_day`, `call_time`, and `tower_location` variables contain the most important information for analyses. These provide information about the specific location from which a user initiates each of their transaction Using these high-frequency locations, I infer residential and employment locations of individuals at the weekly- and monthly-level.

I exploit two different methods to identify the residential location of users in the CDR: the modal algorithm and the Hartigan's algorithm. Appendices A and B explain each of these algorithms in detail. In short, the modal algorithm assigns the location of the tower that an individual interacts with the most during residential hours as their residential location. The Hartigan's algorithm improves this approach by clustering adjacent towers and assigning the centroid of an individual's most used cluster of towers as their residential location. This approach performs better in rural regions where cell phone towers are sparsely distributed.

Table 1: CDR Data Snippet

customer_id	call_day	call_time	call_duration	service_type	tower_location	nationality	recipient
1234	2016-01-01	11:53:20	00:07:23	Voice	(31.2567,34,3206)	Jordanian	4321
1234	2016-01-01	18:21:25	00:13:21	Data	(31.3461,34,2231)	Jordanian	
1234	2016-01-01	18:23:25		SMS	(31.3461,34,2231)	Jordanian	9876
...	...	...	...	...	...	...	...
4321	2016-01-01	11:53:20	00:07:23	Voice	(32.1261,35,1121)	Syrian	1234
...	...	...	...	...	...	...	...

Table 2: CDR Data Observations (2015-06-01)

	Obs.	Freq.	Num. Users	p50	Mean	Std. Dev.
Data	44,455,397	0.81				
Voice	9,126,199	0.17				
SMS	1,293,933	0.02				
N/A	144	0.00				
Total	54,875,643	1.00	1,224,175	43	44	270

3.2 Administrative Data

3.2.1 Census

Census data from 2004 (i.e., pre-crisis) and 2015 (i.e., post-crisis) provide aggregate information at the neighborhood-level. Mainly, I use information about the neighborhood-level changes in the share of Syrian refugees and the changes in other neighborhood characteristics, such as changes in rent per bedroom and changes in population size.

3.2.2 Jordan Household Expenditure and Income Survey (HEIS)

Jordan Household Expenditure and Income Survey (HEIS) provides wage information that is not contained in the census data and are available for years 2006, 2008, 2010, 2013, and 2017. I use this data to estimate migration elasticity to wages. Since the CDR data is only available for years beginning 2015, the I use the 2017 HEIS.

3.2.3 Night Light Data

Night light is a proxy for economic activity. The data are processed by NOAA, and cover years 2000 through 2013.[2] I use 2011 data as the pre-crisis level and 2013 data as the post-crisis level.

[2] I am working to gain access to data from more recent years.

3.3 Validity of the CDR

To understand the accuracy of individuals' residential locations inferred from the CDR, I compare the residential densities implied by the CDR with the residential densities in the official 2015 census and find that the relationship is approximately log-linear.

In the preferred specification, I aggregate the CDR from March 2016 at the sub-district-level.[3] Then, we compare the total number of cell phone users whose residential location lies in each sub-district to the total census population in that sub-district.

The preferred specification is at the sub-district-level for a similar reason. While the census data is provided at a more disaggregated neighborhood-level, the comparisons between the CDR and the census are imprecise at this level, again because cell phone towers are sparsely distributed in non-urban regions of Jordan. For instance, a rural neighborhood that happens to have a cell phone tower located within it will mechanically be over-represented in the CDR compared to adjacent neighborhoods that do not, while the population in the census may be more evenly distributed.

In Figure 4, I plot the log of the population in the CDR against the log of the population in the census at the sub-district-level. Because the number of cell phone users who hold a contract with the telecom company is only a fraction of the total population, I normalize both sides to have unit geometric mean such that the levels are balanced. Panel (a) reports results using the modal algorithm and Panel (b) shows results using the Hartigan's algorithm. The black line represents the 45-degree line. For each scatterplot, I report the coefficient and R-squared from the corresponding regression. The slope coefficients from both the modal and Hartigan's regressions are close to one, implying that a one percent increase in the CDR population is associated with an approximately one percent increase in the census population. The R-squared implies that 74.5% and 77.6% of the variation in the census population is explained by the CDR population from the modal and the Hartigan's algorithms, respectively.

Despite their differences, modal and Hartigan's locations perform very similarly. For computational reasons, I use modal locations in the analyses. Furthermore, in the current iteration of the paper, migration analyses using CDR are conducted at the governorate-level. At the governorate-level, the CDR more accurately captures the population density in the census than at the more disaggregated sub-district-level.

4 Impacts of Refugee Migration on the Local Economy

In this section, I show preliminary analyses on the characteristics of neighborhoods where refugees move to (i.e., where do refugees move to?), as well as the impacts of the refugee influx on neighborhood characteristics (i.e., how does the influx of refugee population shape neigbhorhood characteristics?).

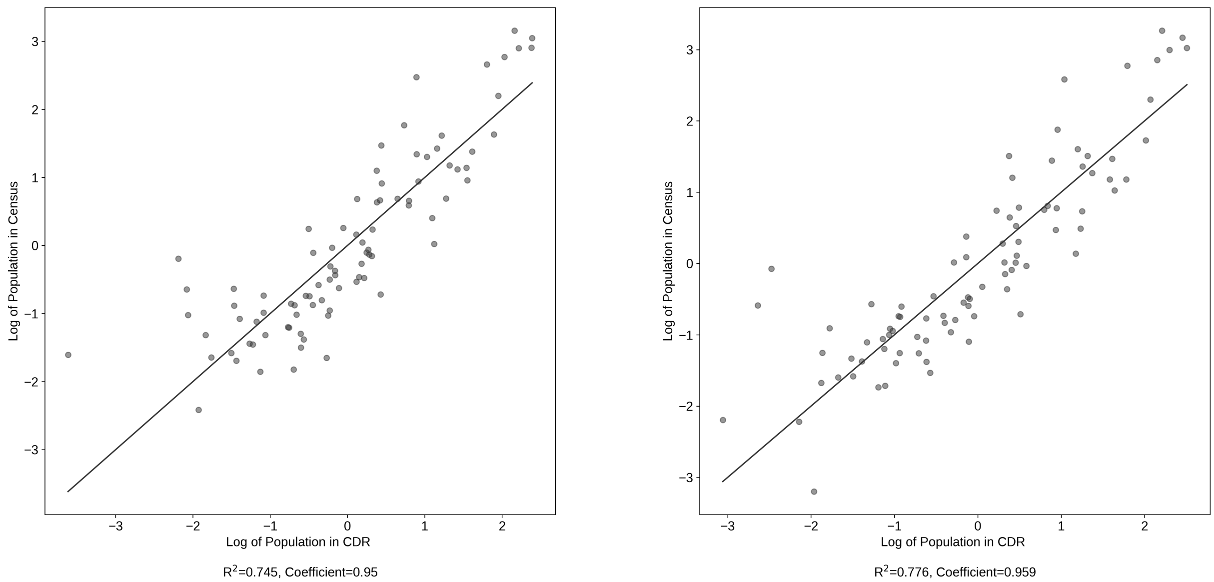
4.1 Initial Characteristics of Neighborhoods

Figure 5 shows binned scatter plots of neighborhoods characteristics pre-crisis (2004 or 2011, depending on data source), against the change in the share of Syrian-born residents between 2004 and 2011. Panels

[3] Geographic units in Jordan in order of descending granularity are neighborhoods (of which there are 1,537 in Jordan), localities (983), sub-districts (89), districts (51), and governorates (12).

Figure 4: Log of Population in CDR V.S. Census

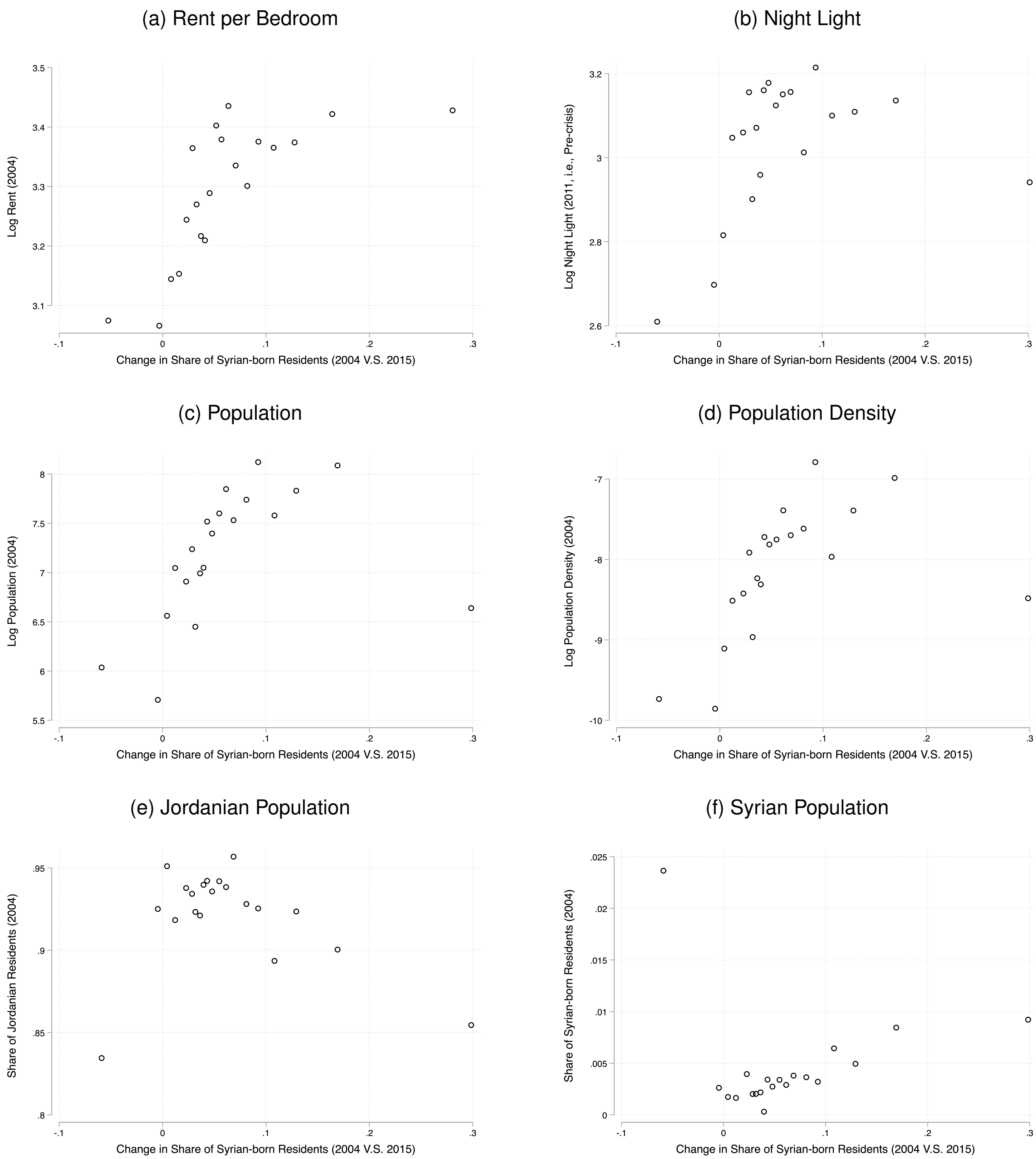
(a) Modal Location (b) Hartigan's Location



Note: The figure shows scatter plots of sub-district-level population of Jordan. The y-axis is log of the number of users in the CDR whose residential location lies in a sub-district. The x-axis is log of the sub-district-level population in the 2015 census. In the left (right) panel, residential locations are inferred using the modal (Hartigan's) algorithm. The black line is the 45-degree line.

(a) and (b) suggest that refugees moved into neighborhoods that have more economic activity, as measured by rent per bedroom and night light. Furthermore, panels (c) and (d) provide evidence that refugees moved into neighborhoods that are relatively more populous and dense. Lastly, panels (e) and (f) show weak evidence that neighborhoods that saw greater increase in refugees are neighborhoods that already have relatively high number of non-Jordanian and Syrian population.

Figure 5: Initial Characteristics of Neighborhoods



Note: The figure shows binned scatter plots of initial characteristics of neighborhoods against the magnitude of the influx of Syrian refugees. The y-axis is (a) log of rent per bedroom, (b) night light, (c) log population, (d) log of population density, (e) share of Jordanian population, and (f) share of Syrian-born population. The x-axis is the change in the share of Syrian-born residents between 2004 and 2015.

4.2 Impacts of Refugee Influx on Neighborhoods Characteristics

To examine the effects of the Syrian refugee influx on local economies, I estimate the following neighborhood-level regression:

$$\Delta \ln Y_j = \alpha_j + \beta_j \Delta \ln SyrianShare_{j,2004-2015} + \phi_j + \epsilon_j$$

where $\Delta \ln SyrianShare_{j,2004-2015}$ is the change in the share of residents that are Syrian-born between 2004 and 2015, and ϕ_j is the neighborhood fixed effects. $\Delta \ln Y_j$ is the log change in outcome variables — rent, Jordanian population, population, and night light — between pre- and post-crisis period. Pre- and post-periods correspond to 2004 and 2015 for all variables except night light, for which the associated periods are 2011 and 2013.

Panel (A) show results for a sub-sample of neighborhoods in urban governorates Amman, Irbid, and Zarqa. Interestingly, an increase in the share of Syrian-born residents is associated with a decrease in rent and night light. Panel (B) show results similar patterns, but rent and night light regressions lack statistical significance at the t%-level.

These findings contradict the findings by Rozo and Sviatschi (2021) that implies a modest increase in rent at the governorate-level. The empirical approach differs in several important dimensions. First, the analyses in this paper are at the neighborhood-level whereas Rozo and Sviatschi (2021) performs their analyses at the governorate-level. Second, Rozo and Sviatschi (2021) instrument influx of Syrian refugees with distance to refugee camps. Analyses in this paper are OLS.

Table 3: Impacts of Refugee Influx on Neighborhoods Characteristics

	Δ Log Rent	Δ Log Jordanian Residents	Δ Log Residents	Δ Log Night Light
Panel A: Amman, Irbid, and Zarqa				
Δ Syrian Share	-0.417**	-0.895**	1.288***	-0.086***
	(0.211)	(0.379)	(0.415)	(0.024)
Observations	579	631	632	593
Panel B: All Governorates				
Δ Syrian Share	-0.226	-1.122**	0.839**	-0.044*
	(0.164)	(0.527)	(0.378)	(0.026)
Observations	1,067	1,184	1,193	1,138
Governorate F.E.	Yes	Yes	Yes	Yes

Standard errors are reported in the parentheses. Statistical significance of 10%, 5%, and 1% is represented by *, **, and ***, respectively.

5 Income Elasticity of Migration

The panel dimension of the CDR provides information about the migratory patterns of individuals. For instance, between 2016 and 2017, 7.4% of 1,847,307 individuals migrated to another governorate.[4] Table 4

[4] I define migration as the change in the governorate of a user's primary residential location between 2015 and 2016. The primary location is the most common within-year modal residential location. For instance, if an individual's modal residential location is

summarizes migration rates for different sub-populations. While the rates are similar for different nationals, residents of the Za'atari camp have far lower migration rate at 3.4%.

Table 4: Year-to-Year Migration Between 2015 and 2016

Sample	Total Num. Users	Total Num. Migrants	Migration Rate
All	1,847,307	137,345	0.074
Jordanians	1,502,015	109,214	0.073
Syrians	71,643	6,007	0.084
Zaatari Residents	9,706	328	0.34

5.1 Gravity Model of Migration

While raw migration rates are informative of broad patterns of migration, it does not help us understand the determinants of migration. Furthermore, raw migration rates are agnostic to the different patterns of migration. To remedy these shortcomings, I turn to estimating income elasticity of migration using a gravity equation. The model follows the standard set up in the trade literature (Méndez-Chacón and Patten 2021 and Beine et al. 2021):

Set Up There are $j \in \{1, \cdots, N\}$ locations. Individuals live two period: they are born in $o \in \{1, \cdots, N\}$ and endogenously choose to live in $d \in \{1, \cdots, N\}$ where they work and consume to derive utility.

Household Utility Households born in o and now residing d have constant elasticity of substitution (CES) preference, deriving utility by consuming goods produced in each of j location and from per capita local amenities in d:

$$U(c_{jd}, A_d, L_d) = \frac{A_d}{L_d}^{1-\alpha} \left(\sum_{j}^{N} c_{jd}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\alpha\sigma}{\sigma-1}}$$
$$W_d = \frac{A_d}{L_d}^{1-\alpha} \left(\frac{w_d}{P_d} \right)$$

where σ denotes CES elasticity, A_d total amenities in d, L_d population, w_d wage, c_{jd} consumption of good produced in j, and $P_d = \left(\sum_{j'}^{N-1} (\tau_{j'd} p_{j'})^{1-\sigma} \right)^{\frac{1}{1+\sigma}}$ CES price index.[5] W_d is the equilibrium deterministic utility of worker in d.

Migration Moving from o to d incurs a moving cost $\lambda_{od} \geq 1$ where $\lambda_{oo} = 1$. Then, deterministic utility of worker is given by $W_{od} = \frac{W_d}{\lambda_{od}}$. Lastly, there is a idiosyncratic taste shocks ν_d that follows a Type 1 Extreme Value (Fréchet) distribution with shape parameter κ. By utility maximization, households choose to migrate to $d = argmax_d \frac{W_d \nu_d}{\lambda_{od}}$. Then, the share and magnitude of out-migrants from o to d is given by:

Amman 11 of 12 months in 2015 and Irbid in 9 of 12 months in 2016, they are a Amman to Irbid migrant in 2016. A more robust approach to defining migration is suggested in Blumenstock et al. (2019), where only the changes in governorate locations lasting at least k many months are considered a migration. At the year-to-year level, this approach is consistent with the simple approach. For computational reasons, the analyses in the paper relies on annual migration data computed using the simple approach.

[5] $\tau_{j'd} > 1$ denotes iceberg trade costs.

$$\frac{L_{od}}{L_o} = -\frac{\left(\frac{W_d}{\lambda_{od}}\right)^{\kappa}}{\sum_{d'}^{N}\left[\left(\frac{W_{d'}}{\lambda_{od'}}\right)^{\kappa}\right]}$$
$$L_{od} = \left(\lambda_{od}\Pi_o\right)^{-\kappa} W_d^{\kappa} L_o$$

where $\Pi_o \equiv \left(\sum_{d'}^{N}\left[\left(\frac{W_{d'}}{\lambda_{od'}}\right)^{\kappa}\right]\right)^{\frac{1}{\kappa}}$.

5.2 Estimation

The equation above motivates the following regression specification:

$$\ln L_{od,t} = -\kappa\mu \ln Distance_{od} + \kappa\alpha \ln w_{d,t} + \kappa(1-\alpha)\ln\frac{A_{d,t}}{L_{d,t}} + \phi_o + \psi_d + \epsilon_{od,t}$$

where ϕ_o and ψ_d are origin and destination governorate fixed effects. I estimate this equation for different sub-populations $i \in All, Jordanian, Syrian, CampResidents$ to uncover heterogeneities in the income elasticity of migration. In this iteration of the paper, all analyses are at the governorate-level. There are endogeneity concerns and ideally, I would need to find instruments for wages and per capita amenities. But in this iteration, I proceed with OLS.

Setting $\ln\lambda_{od} \equiv \ln Distance_{od}$ follows the trade literature. For $\ln\frac{A_{d,t}}{L_{d,t}}$, government budget data on governorate-level expenditure on infrastructure such as schools, hospitals, and transportation normalized by governorate-level population would be ideal. Unfortunately, I do not have access to these data. To remedy this, I use several approaches for amenities. First, I ignore amenities by assuming $\alpha = 1$. Second, I assume counterfactually that amenity values are constant across governorates by setting $A_{d,t} = 1 \forall\ d$. Then, urban governorates such as Amman and Irbid have disamenities compared to rural governorates in the sense that there are only congestion effects from a large population. Alternatively, I use governorate-level "Quality-of-Life" (QOL) index from a 2010 report by Jordan's Department of Statistics. This index is based on proximity to public hospitals and health centers.

For normalization, I set $L_{d,t}$ to equal either the entire governorate-level population inferred from the CDR data or the governorate-level sub-population size (i.e., number of Jordanians, Syrians, or Camp Residents). The latter specification follows from the observation that different sub-population may have access to a different set of amenities. Then, holding fixed the governorate-level variation in sub-population-specific amenity values, individuals compete for amenities with other individuals belonging to the same sub-population, but not with others.

5.3 Results

Tables 5 through 7 show results for the Jordanian, Syrian, and camp resident sub-population, respectively. The preferred specification is the one that ignores utility drawn from local amenities, i.e., $\alpha = 1$. Other specifications lead to counterfactual levels of α, which threatens interpretation!

Despite these findings, coefficients on $\ln w_{d,t}$ is smaller for the refugee population compared to the Jordanian population in all specifications. I interpret this as the refugee population having greater migration frictions compared to Jordanians. If future iterations of the emprical exercise confirms this finding, the heterogeneities in the income elasticity of migration can inform a spatial general equilibrium model.

Plans for the Future Results here are compromised by the following issues that I plan to address in future iterations: I only use data from one cross-section $t = 2017$ due to a lack of wage data in other years for which I have CDR data. Furthermore, as discussed above, OLS can lead to serious bias in the gravity equation estimation. Lastly, the lack of amenity data can also lead to bias in the estimates.

Table 5: Income Elasticity of Migration — Jordanian

	(1)	(2)	(3)	(4)	(5)
Log Distance	-0.596***	-0.593***	-0.596***	-0.592***	-0.595***
	(0.033)	(0.026)	(0.025)	(0.027)	(0.026)
Log Labor Income	0.766	1.396***	2.583***	1.654***	2.964***
	(0.612)	(0.493)	(0.507)	(0.513)	(0.537)
Log 1/Population		-0.812***			
		(0.093)			
Log QOL/Population			-0.908***		
			(0.095)		
Log 1/Group Population				-0.809***	
				(0.100)	
Log QOL/Group Population					-0.946***
					(0.104)
Mean DV	6.283	6.283	6.283	6.283	6.283
R-squared	0.752	0.844	0.855	0.835	0.849
Observations	144	144	144	144	144

Dependent variable is L_{od} for the Jordanian sub-population. All specification includes origin governorate fixed effects. Standard errors are reported in the parentheses. Statistical significance of 10%, 5%, and 1% is represented by *, **, and ***, respectively.

Table 6: Income Elasticity of Migration — Syrian

	(1)	(2)	(3)	(4)	(5)
Log Distance	-0.538***	-0.542***	-0.547***	-0.554***	-0.558***
	(0.039)	(0.028)	(0.026)	(0.024)	(0.023)
Log Labor Income	-1.444*	-0.472	1.269**	1.512***	2.634***
	(0.754)	(0.555)	(0.554)	(0.511)	(0.530)
Log 1/Population		-1.077***			
		(0.104)			
Log QOL/Population			-1.222***		
			(0.101)		
Log 1/Group Population				-0.777***	
				(0.056)	
Log QOL/Group Population					-0.796***
					(0.054)
Mean DV	3.117	3.117	3.117	3.117	3.117
R-squared	0.666	0.825	0.850	0.874	0.881
Observations	133	133	133	133	133

Dependent variable is L_{od} for the Syrian sub-population. All specification includes origin governorate fixed effects. Standard errors are reported in the parentheses. Statistical significance of 10%, 5%, and 1% is represented by *, **, and ***, respectively.

Table 7: Income Elasticity of Migration — Refugee Camp Residents

	(1)	(2)	(3)	(4)	(5)
Log Distance	-0.523***	-0.479***	-0.406***	-0.272***	-0.270***
	(0.114)	(0.090)	(0.088)	(0.068)	(0.081)
Log Labor Income	-0.986	-0.478	1.065	0.465	0.930
	(2.311)	(1.820)	(1.768)	(1.192)	(1.406)
Log 1/Population		-1.060***			
		(0.320)			
Log QOL/Population			-1.299***		
			(0.335)		
Log 1/Group Population				-0.619***	
				(0.090)	
Log QOL/Group Population					-0.556***
					(0.100)
Mean DV	2.278	2.278	2.278	2.278	2.278
R-squared	0.633	0.788	0.817	0.911	0.880
Observations	20	20	20	20	20

Dependent variable is L_{od} for the camp resident sub-population. All specification includes origin governorate fixed effects. Standard errors are reported in the parentheses. Statistical significance of 10%, 5%, and 1% is represented by *, **, and ***, respectively.

6 Spatial Equilibrium Model of Optimal Migration

to be added

7 Conclusion

to be added

References

Ibrahim Alhawarin, Ragui Assaad, and Ahmed Elsayed. Migration shocks and housing: Short-run impact of the syrian refugee crisis in jordan. *mimeo*, 2020.

Treb Allen and Dave Donaldson. Persistence and path dependence in the spatial economy. *mimeo*, 2020.

Samuel Bazzi. Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, 2017.

Michel Beine, Luisito Bertinelli, Rana Cömertpay, Anastasia Litina, and Jean-François Maystadt. A gravity analysis of refugee mobility using mobile phone data. *Journal of Development Economics*, 2021.

Daniel Björkegren. The adoption of network goods: Evidence from the spread of mobile phones in rwanda. *Review of Economic Studies*, 2019.

Joshua Blumenstock, Guanghua Chi, and Xu Tan. Migration and the value of social networks. *mimeo*, 2019.

Cecile Gaubert, Patrick Kline, and Danny Yagan. Place-based redistribution. *mimeo*, 2021.

Esteban Méndez-Chacón and Diana Van Patten. Multinationals, monopsony, and local development: Evidence from the united fruit company. *mimeo*, 2021.

Sandra V. Rozo and Micaela Sviatschi. Is a refugee crisis a housing crisis? only if housing supply is unresponsive. *Journal of Development Economics*, 2021.

Appendix

A Modal Algorithm for Identifying Residential and Employment Locations

To find the residential location of i in month m, the modal algorithm proceeds as follows:

1. For individual i, find $X_{i,m}$, the list of all towers from which they have a transaction in month m.
2. Sort $X_{i,m}$ in descending order of $t_{x,i,m}$, the *number of transactions from tower x during the residential hours in month m*. The residential hours are 7 pm to 9 am during weekdays.

   For simplicity, let $X_{i,m} = x_1, x_2, x_3, ...$ be the sorted list where x_1 denotes the first tower in the list, i.e. the tower that has the largest the number of transactions.
3. Tower x_1 is the residential location. This is referred to as the modal tower.

The same process is repeated to find the employment location of i by replacing residential hours with employment hours (10 am to 3 pm during weekdays). Similarly, weekly residential and employment locations can be found by replacing month m with week w.

B Hartigan's Algorithm for Identifying Residential and Employment Locations

The Hartigan's leader algorithm clusters adjacent towers together, and ranks clusters by usage, instead of ranking individual towers. To find the residential location of i in month m, the Hartigan's leader algorithm proceeds as follows:

1. For individual i, find $X_{i,m}$, the list of all towers from which they have a transaction in month m.
2. Sort $X_{i,m}$ in descending order of $d_{x,i,m}$, the *number of days that i has a transaction from tower x in month m* during the residential hours.[6] The residential hours are 7pm to 9 am during weekdays.

   For simplicity, let $X_{i,m} = x_1, x_2, x_3, ...$ be the sorted list where x_1 denotes the first tower in the list, i.e. the tower that has the largest number of days used.
3. x_1 forms the first cluster. Naturally, x_1 is the centroid of that cluster.
4. If x_2 is within r radius of x_1, it is added to the first cluster. The new centroid of the cluster is defined as the centroid of x_1 and x_2, weighted by $d_{x_1,i,m,r}$ and $d_{x_2,i,m,r}$. If x_2 is more than r distance away from x_1, it forms a new cluster.[7]
5. Repeat the previous step for all subsequent towers in the sorted list. That is, if a tower is within r distance to the centroid of the nearest cluster, it is added to that cluster. Otherwise, it forms a new cluster. This returns a list of clusters, $C_{i,m}$.

[6] Note that unlike in the simple ranking algorithm, I use the number of days that i uses a tower x, instead of the number of transactions from the tower. This approach makes the Hartigan's algorithm more robust to outliers. For instance, if an individual has a burst of activity on their weekend trip, this is less likely to affect the Hartigan's locations than the modal location.

[7] Following Björkegren 2019, I set r equal to the distance from the centroid to the ninth closest tower.

6. Sort $C_{i,m}$ in descending order of the aggregate number of days from all towers in cluster c during the residential hours in m.

   For simplicity, let $C_{i,m} = \{c_1, c_2, c_3, ...\}$ be the sorted list where c_1 denotes the first cluster in the list, i.e. the cluster of towers that has the most number of days used in aggregate.

7. The centroid of cluster c_1 is the residential location.

The same process is repeated to find the employment location of i by replacing residential hours with employment hours (10 am to 3 pm during weekdays). Similarly, weekly residential and employment locations can be found by replacing month m with week w.