Refugees in undeclared employment - A case study in Turkey*

Keywords: *Migration, Refugees, Mobile phone metadata, Employment, Agriculture, Construction.*

1 Introduction

The production of timely and precise statistics is critical for effective decision-making and policy implementations by both governments and non-governmental institutions. Nevertheless, many socio-economic measures are based on lagged and imprecise information. Data on employment, for instance, is typically collected by means of surveys which come with statistical challenges and at high costs. Mobile phone metadata has been at the center of recent efforts to find alternative methods to measure socio-economic indicators such as literacy [1], poverty [2], and employment [3], [4], [5].
However, to the best of our knowledge there has not been made any efforts to use mobile phone metadata to measure undeclared employment. By its nature, undeclared employment is especially difficult to measure with traditional approaches like surveys. Instead, mobile phone metadata allow for a fine-grained view on migration and commuting routines, possibly unearthing employment situations which cannot be captured otherwise. We test the methodology in an early proof-of-concept on two distinct cases applying a Difference-in-Differences (DiD) approach. First, we look at the hazelnut harvest in the Turkish province of Ordu, which accounts for a quarter of the world's hazelnut supply, making it a topic of global impact. A hazelnut harvest is seasonal, labour-intensive and regionally limited. This helps to isolate work as the prime reason for temporary migration. Second, the Istanbul Grand Airport construction site - of importance for both the construction sector and, once completed, the Turkish economy - could be associated with daily commutes by refugees. We assume that work patterns among refugees most likely link to undeclared employment, as only 1.3% have a permit to work as of 2017.
The study builds on data provided by the D4R Turkey challenge. It includes a sample of CDRs from a large Turkish mobile network operator for 2017, preprocessed to antenna traffic (SET1). A tag allows to disaggregate SET1 by refugee status (refugee vs. non-refugee). Contextual data is provided as well, including the antenna locations and details on the composition of the sample.[1] Further, we use publicly-available information on the administrative boundaries from the Humanitarian Data Exchange, geographic shapes of refugee camps and of the Istanbul Grand Airport extracted from Google Earth based on location information from the Humanitarian Data Exchange and Google Maps, respectively.

2 Methods

We propose a unified methodological framework to identify employment structures of Syrian refugees in Turkey. SET1 is used to investigate effects of refugee-specific work-related movement using the number of SMS, the number of calls and the volume of calls (network activity). These variables are proxies, as the actual variable of interest - the number of users - is not available. The study uses voronoi tessellation to identify relevant antennas and their respective approximated geographical coverage areas. The

*Contribution to the Data for Refugees (D4R) Challenge

[1] For details on the D4R data and its limitations, see [6].

methodology consists of two parts dealing with specific characteristics of work-related movements: seasonal migration and commuting.

2.1 Seasonal migration

Under the assumptions that a) network activity proxies the number of users, b) mobility is the main driver for user fluctuation, c) the large majority of work performed by refugees is undeclared, and d) refugees are unlikely to go on summer vacations in Turkey, the following procedure gives a descriptive indication whether there is migration of refugees towards undeclared employment for a specific harvest event:

1) Identify harvest events by season start and end, agricultural produce and geographic location
2) Start iteration over harvest events
3) Divide SET1 into three intervals: before, during and after season
4) Compute rate of change Δnet_act between the median values of weekly network activity by interval and harvest location (harvest location vs. non-harvest location) only for the refugee-tagged CDRs, namely:

$$\Delta net\_act = \frac{\tilde{net\_act}_{after}}{\tilde{net\_act}_{before}} - 1 \tag{1}$$

5) End iteration over harvest events

2.2 Commuting

Under assumptions a) - d), this procedure indicates whether there is commuting of refugees due to undeclared employment on the Istanbul Airport construction site:

1) Identify workplaces by sector and geographic location
2) Reduce SET1 to workdays (Mon-Fri)
3) Start iteration over workplaces
4) Divide it into two groups: Work (7-20h) and Home (20-7h)
5) Compute rate of change Δnet_act between median values of network activity by group and workplace (workplace vs. non-workplace) only for the refugee-tagged CDRs following equation (1).
6) End iteration over workplaces

Note, both procedures can be combined, e.g., to test assumption d) for Ordu.

2.3 Proof-of-Concept

While descriptive statistics are of interest, we also provide an early proof-of-concept of the proposed framework: We apply DiD estimation to account for selection bias arising from differences in antenna characteristics like, e.g., geographic location. DiD builds on two assumptions: 1.) A control group is available, i.e. a set of antennas located such that refugee-specific network activity is not affected by the treatment. 2.) Refugee-specific network activities show a common trend (CT) prior to the treatment. The DiD estimation is based on the following equation:

$$net\_act_{i,t} = \beta_0 + \beta_1 * treat_i + \beta_2 * time_t + \beta_3 * treat_i * time_t + \epsilon_{i,t} \tag{2}$$

where $net\_act_{i,t}$ measures weekly/hourly refugee-tagged network activity at antenna i in week/hour t; $treat_i$ and $time_t$ are dummy variables indicating whether antenna i belongs to the treatment group and week/hour t to the treatment period, respectively; $\epsilon_{i,t}$ denotes the error term.

The DiD estimator is given by β_3 indicating the average change over time in $net\_act_{i,t}$ for the treated compared to the average change over time for the controls.

3 Results

3.1 Hazelnut harvest in Ordu

Using descriptive statistics, we find clear indication of harvest-related seasonal migration to Ordu (see Table 1 and Figure 1). At the start of the harvest, growth in network activity is 77.9%-points higher in Ordu than in the rest of Turkey. At the end of the harvest, network activity in Ordu shrinks by 39.8% vs. an increases by 14.9% elsewhere.

To validate these findings using the DiD approach, we define the beginning of the harvest, i.e. calendar week 31, as treatment. We choose our control group based on the assumption that refugees in areas with high work-to-home-ratio are likely to pursue some kind of working activity and are, thus, less prone towards migrating to Ordu for the harvest. We select all antennas that are comparable to our treatment group with respect to the magnitude of the overall network activity. Figure 2 illustrates the treatment and control group. Figure 3 substantiates the CT assumption, indicating similar growth rates in network activity prior to the treatment. Table 2 gives the treatment effects estimated following equation (2). Confirming the findings based on descriptive statistics, the treatment effects for all network activity measures are positive. The average weekly number of calls, for instance, increased by approx. 85% during to the harvest relative to the number of calls that would have been placed during calendar week 31-36 in Ordu should there not have been the season of hazelnut harvest. Given the relative treatment effects on network activity measures ranges from 60% and 126%, the actual growth of refugee population in Ordu due to the harvest likely lies somewhere within this range. Next, we define the end of the harvest (calendar week 37) as a second treatment. For the validity of the CT assumption, see Figure 4. Table 3 indicates that all DiD estimators are negative, confirming that the spike in network activity by refugees in Ordu is limited to the summer weeks and further substantiating our hypothesis that refugees migrated to Ordu temporarily during the hazelnut harvest. Table 4 compares the network activity growth from the descriptive statistics with the treatment effect from the DiD analysis.

3.2 Construction site of the Istanbul Grand Airport

Using descriptive statistics, we find indication of work-related commuting of refugees to Istanbul Grand Airport. Table 5 compares the growth rate of the average number of refugee-tagged calls between the hours defined as working and home hours for the Istanbul Grand Airport construction site and the rest of Turkey. We observe a network activity growth 966.5%-points higher at the airpot construction site than elsewhere. We define the treatment to be the beginning of the workday (7:00 am). We choose control group antennas from all antennas located near refugee camps and select those which are similar to our treatment group (a set of antennas in the range of the airport) with respect to summary statistics, arguing that refugees living in camps typically do not follow a working routine from 7:00 am - 20:00 pm which would affect their network activity (Figure 5). As illustrated by Figure 6, the CT assumption seems to hold. As expected, the results show a positive treatment effect for all four measures of network activity (Table 6). The relative treatment effect for, e.g., the number of calls suggests that the average number of calls placed by refugees during the day would be 9%-points lower in the area of the airport if there would not be a construction site. The relative treatment effects for alternative measures of network activity suggest this number could be even higher. The positive treatment effects suggest that there might indeed be a commuting pattern towards the construction site during the day, hinting towards undeclared labour by refugees at the construction site of the Istanbul Grand Airport. This interpretation builds on the assumption that a person's work does not affect the network activity. However, phone use may depend on the profession (e.g. manager vs. construction worker). Note that in this case the commuting effects are, therefore, likely underestimated. Table 7 compares the network activity growth from the descriptive statistics with the treatment effects. Even though the results show the same direction, the strength of the effects vastly varies. This is most likely due to the small sample size as it only includes refugee-tagged customers equivalent to 6% of all registered Syrian refugees [6]. In order to evaluate the early proof-of-concept for the case of the Istanbul Grand Airport, a larger sample is necessary.

3.3 Exploration

Exploratory analysis might help detect areas with a high probability of undeclared employment opportunities for refugees and, thereby, could help to mitigate undeclared work by informing effective intervention planning. Figure 7 shows areas of similar network activity as the airport construction site - low activity during the night, high activity during the day. We identify 60 locations in Turkey with these characteristics and similar magnitude. By comparing the results with information from Google Maps, we are able to identify areas with a high probability of refugee employment such as the industrial park near Çerkezköy (see Figure 7). The approach also produces a high rate of potentially false positives like shopping malls, hospitals and, e.g., the Grand Bazaar in Istanbul where it is impossible to separate refugee-tagged visitors from workers. This shows a fundamental weakness of the exploratory approach: SET1 does not provide information on the duration and regularity of stay derived from individual-level information that could help to differentiate between visitors and workers.

4 Conclusions

This study lays out a framework for identifying potentially undeclared employment among refugees in Turkey. Further, it has provided an early proof-of-concept based on a Difference-in-Differences approach using two case studies: the hazelnut harvest in Ordu in late-summer and the construction site of the Istanbul Grand Airport. We have found clear indication for work-related commuting and seasonal migration among refugees hinting at undeclared employment situations. By informing effective intervention planning, fine-grained information about undeclared employment situations may help to fight undeclared work with all its negative implications: high-risk jobs, pay below minimum wage and lack of access to social security.

However, socio-economic statistics derived from mobile phone metadata carry certain shortcomings: The data generating process for mobile phone metadata is usually beyond the control of the analyzing entity and, thus, might not adhere to relevant statistical concepts such as representativeness. The mismatch between the unit of analysis (the individual) and the unit of observation (the device) adds to this as one individual can have multiple phones and vice versa.

Further, there are additional limitations specific to this case study: 1.) The extraction of the CDR sample does not seem to follow a clear sampling strategy [6]. Consequently, generalizations of findings based on this sample may not be adequate. 2.) The refugee-tag depends on business model decisions of the mobile network operator and thus may prove unreliable in implementation. 3.) The main assumption justifying the link between work-related commuting/migration and undeclared employment is the fact that only 1.3% of refugees have been granted a work permit up to 2017. This might change in the future, which would negatively affect the power of the approach. 4.) SET1 shows unexpected strong growth of interactions over the year. While the reason for this may be found in the CDR extraction, it influences the descriptive statistics on migration by overstating migration growth numbers over time. 5.) This study uses indicators related to network activity as a proxy for the number of users in the network. Thus, changes in those variables are not path-dependent and may only inform on net migration. However, net migration is a result of complex in- and out-flows, which eventually understate the true migration. 6.) While DiD mitigates selection bias, the study does not use additional variables to control for other potential influences on commuting/migration patterns such as climate- or geography-related aspects.

References

[1] Schmid, T., Bruckschen, F., Salvati, N., Zbiranski, T.: Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. In: Journal of the Royal Statistical Society: Series A (Statistics in Society) **180**(4), pp. 1163–1190. Wiley Online Library, (2017)

[2] Pokhriyal, N., Jacques, D. C.: Combining disparate data sources for improved poverty prediction and mapping. In: Proceedings of the National Academy of Sciences **114**(46), pp. 9783–9792, (2017)

[3] Toole, J. L., Lin, Y., Muehlegger, E., Shoag, D., Gonzalez, M. C., Lazer, D.: Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface **12**(107), (2015)

[4] Almaatouq, A., Prieto-Castrillo, F., Pentland, A.: Mobile communication signatures of unemployment. In: International conference on social informatics, pp. 407–418. Springer, (2016)

[5] Sundsøy, P., Bjelland, J., Reme, B.-A., Jahani, E., Wetter, E., Bengtsson, L.: Towards Real-Time Prediction of Unemployment and Profession. In: International Conference on Social Informatics, pp. 14–23. Springer, (2017)

[6] Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö.: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523, (2018)

Tables

Table 1: Average growth of number of refugee-tagged calls

Location	**before → during**	**during → after**
Ordu	241.3%	-39.8%
Rest of Turkey	163.4%	14.9%
Difference	**77.9%**	**-54.7%**

Table 2: Treatment effects of the beginning of harvest on network activity by refugees

Network activity measure	Control group CW 1-30	Control group CW 31-36	Treated group CW 1-30	Treated group CW 31-36	**Absolute TE**	**Relative TE**
No. of SMS	0.46	0.53	0.36	0.70	0.26	59.8%
No. of calls	2.78	5.44	3.75	11.89	5.47	85.3%
Call volume	339	599	327	1327	739	125.6%
No. of interactions	3.24	5.97	4.11	12.59	5.74	83.7%

Table 3: Treatment effects of the end of harvest on network activity by refugees

Network activity measure	Control group CW 1-30	Control group CW 31-36	Treated group CW 1-30	Treated group CW 31-36	**Absolute TE**	**Relative TE**
No. of SMS	0.53	1.55	0.70	0.59	-1.13	-65.7%
No. of calls	5.44	7.44	11.89	7.53	-6.35	-45.8%
Call volume	599	871	1327	783	-815	-51.0%
No. of interactions	5.97	8.99	12.59	8.12	-7.48	-48.0%

Table 4: Hazelnut harvest: Comparing descriptive statistics and treatment effects

	%-change in no. of calls	
	before → during	during → after
Descriptives	77.9%	-54.7%
Treatment effect	85.3%	-45.8%

Table 5: Growth rate of average number of refugee-tagged calls before & during work

Location	**home → work**
Istanbul Grand Airport	1275%
Rest of Turkey	308.5%
Difference	**966.5%**

Table 6: Treatment effects of beginning of the workday on network activity by refugees

Network activity measure	Control group CW 1-30	Control group CW 31-36	Treated group CW 1-30	Treated group CW 31-36	**Absolute TE**	**Relative TE**
No. of SMS	0.00	1.05	0.41	1.67	0.21	14.2%
No. of calls	7.39	20.79	1.39	16.09	1.31	8.8 %
Call volume	844	2,284	282	2042	319	18.5%
No. of interactions	7.39	21.84	1.80	17.76	1.51	9.3%

Table 7: Istanbul Grand Airport construction site: Comparing descriptive statistics and treatment effects

	%-change in no. of calls
	home → work
Descriptives	966.5%
Treatment effect	8.8%

Figures

Figure 1: Weekly number of refugee-tagged calls in Ordu and the rest of Turkey

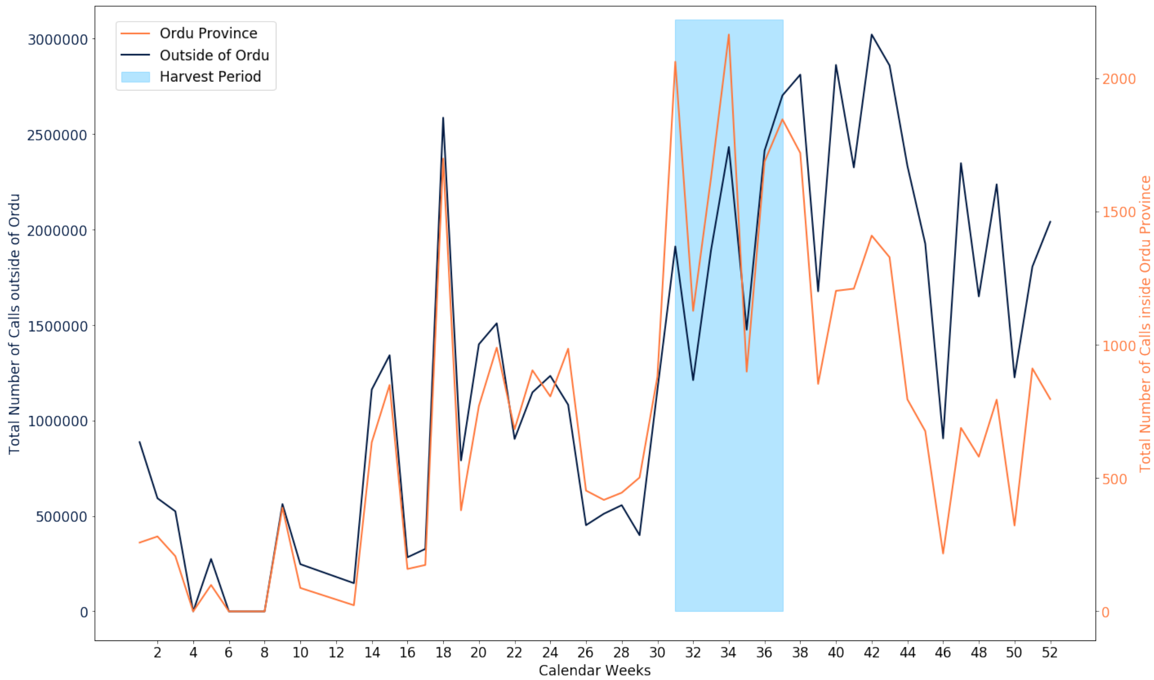
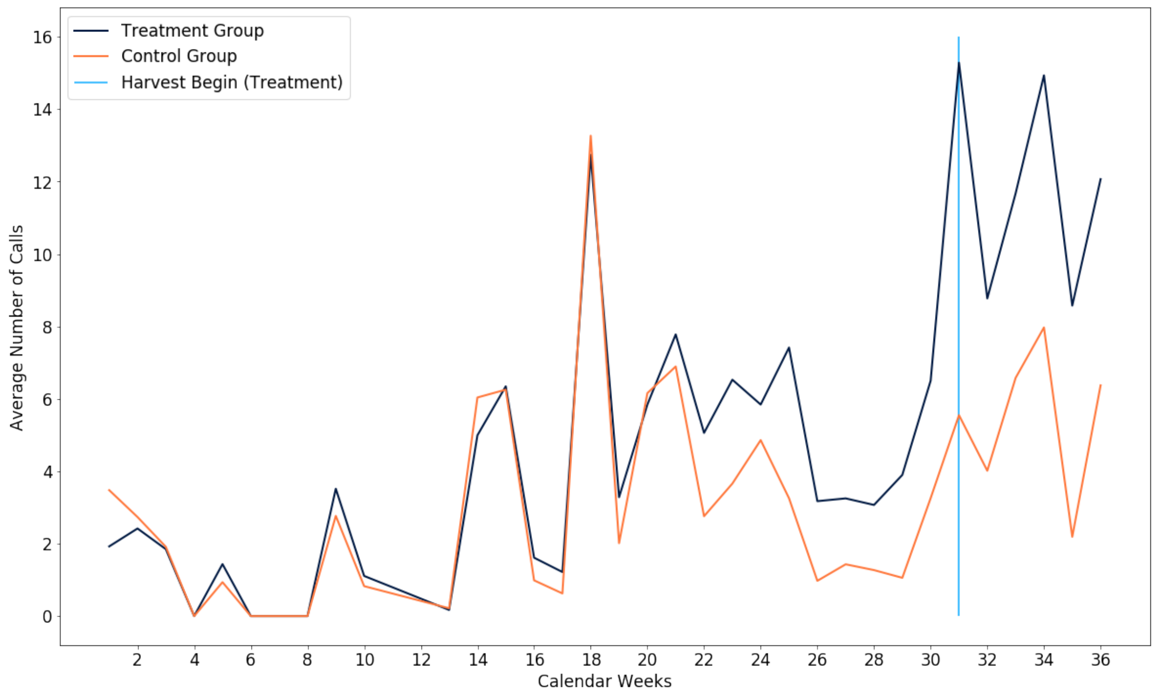


Figure 2: Hazelnut harvest: Treatment group (blue) and control group (orange). Map of Turkey divided into coverage areas of antennas approximated using voronoi tessellation.

Figure 3: Treatment I: Weekly number of calls by treatment and control group



Note, the fluctuations in calendar week 14 & 18 are observable for all Syrian refugees in the sample. The trends for other network activity measures, like number of SMS, call volume, or total number of interactions look similar. The increase in calls made by refugees in Ordu (blue line) starting in week 21 hints towards an anticipation effect, the so-called Ashenfelter's Dip. Refugees seem to have migrated before the harvest started in anticipation of the employment opportunity. Hence, the estimated treatment effect likely underestimates the real effect.

Figure 4: Treatment II: Weekly number of calls by treatment and control group

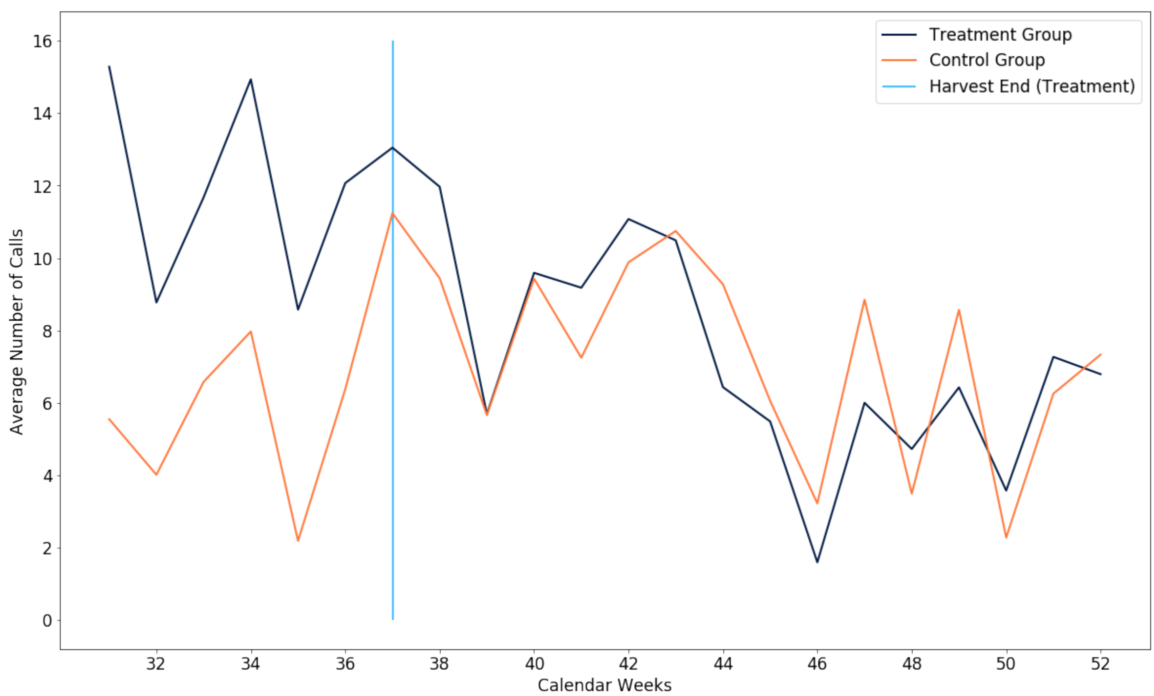


Figure 5: Istanbul Grand Airport construction site: Treatment group (blue) and control group (orange)

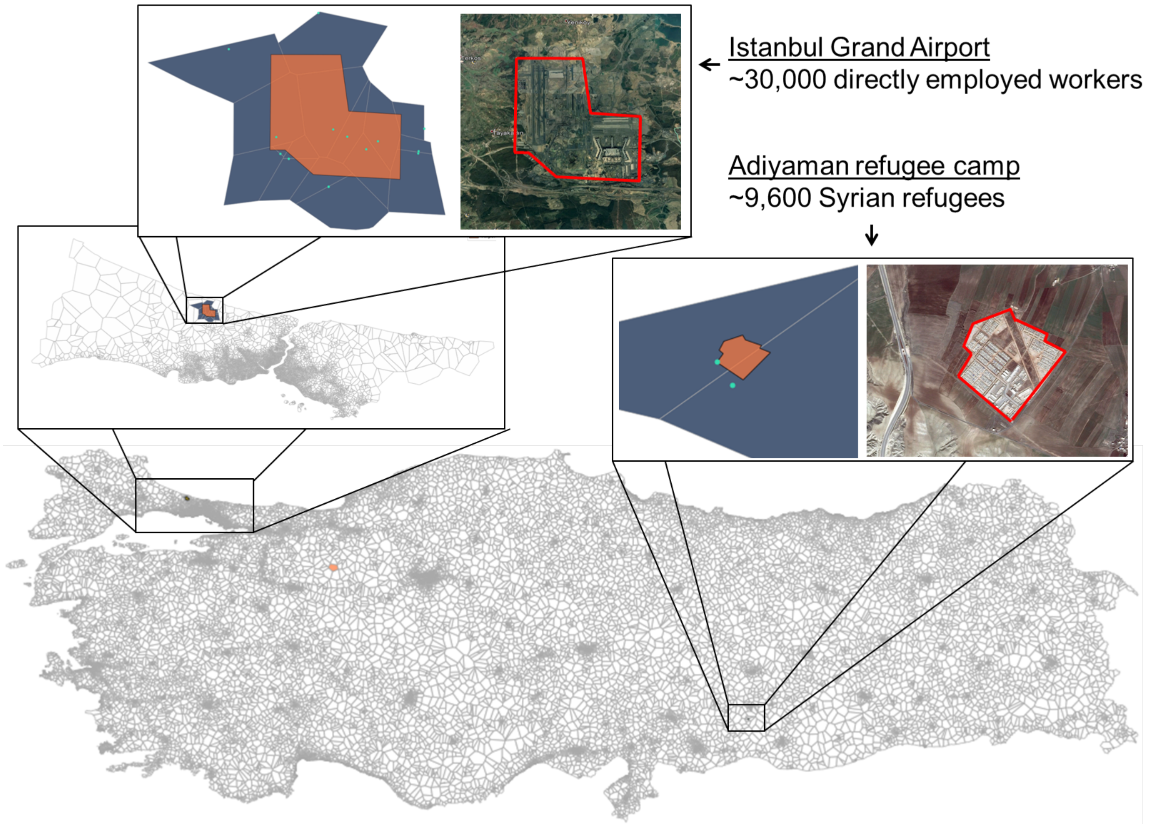
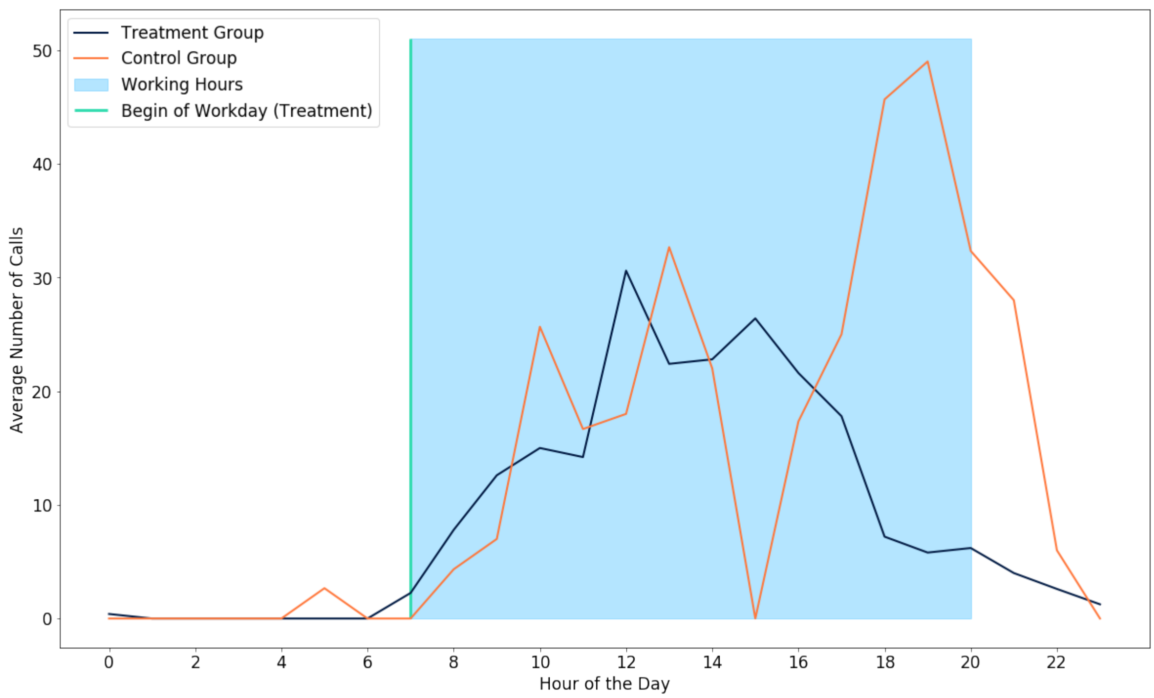


Figure 6: Treatment III: Hourly number of calls by treatment and control group



Note: The figure shows the annual total number of calls in a certain hour averaged over all antennas in the control and treatment group. The trends for other network activity measures, like number of SMS, call volume, or total number of contacts, i.e. SMS and calls, look similar.

Figure 7: Antennas (blue) with similar network activity as the Istanbul Grand Airport area. High probability of refugee employment: industrial park near Çerkezköy.