Integration of Syrian refugees: insights from D4R, media events and housing market data*

Simone Bertoli, Paolo Cintia, Fosca Giannotti, Etienne Madinier, Çağlar Özden, Michael Packard, Dino Pedreschi, Hillel Rapoport, Alina Sîrbu, and Biagio Speciale

Abstract We explore various means of quantifying integration using two of the D4R Challenge datasets. We propose various integration indices and discuss their output. We combine the data from the D4R Challenge with data from the GDELT Project and with data on transactions on the housing market in Turkey. We also describe research directions to be undertaken should an extended access to the data be provided.

Simone Bertoli
Université Clermont Auvergne, CNRS, IRD, CERDI, Clermont-Ferrand, France, e-mail: simone.bertoli@uca.fr

Paolo Cintia, Dino Pedreschi, Alina Sîrbu
Department of Computer Science, University of Pisa, Pisa, Italy, e-mail: paolo.cintia@unipi.it, dino.pedreschi@unipi.it, alina.sirbu@unipi.it

Fosca Giannotti
ISTI-CNR, Pisa, Italy , e-mail: fosca.giannotti@isti.cnr.it

Etienne Madinier, Hillel Rapoport, Biagio Speciale
Paris School of Economics, Paris, France, e-mail: etienne.madinier@psemail.eu, hillel.rapoport@univ-paris1.fr, Biagio.Speciale@univ-paris1.fr

Çağlar Özden
The World Bank, Washington DC, United States, e-mail: cozden@worldbank.org

Michael Packard
Georgetown University, Washington DC, United States, e-mail: mmp77@georgetown.edu

* The findings in this paper do not necessarily represent the views of the World Bank's Board of Executive Directors or the governments they represent. Any errors or omissions are the authors' responsibility.

1 Overview

Responding to a sudden arrival of large number of refugees is a daunting task for many host societies and governments. After addressing the immediate humanitarian needs of millions of people fleeing from civil war and violence, destination countries turn their attention to medium and long-term issues since the refugees are unable or unwilling to return home in most cases, in fear of their safety and well-being. At that point, assimilation and integration become the key concern, in order to reduce the burdens on the host society and the refugees. Identifying the extent and determinants of refugee integration will help policymakers mitigate the negative impacts (perceived or real) of refugees and further facilitate more refugee settlement.

The integration of minority groups, whether native- or foreign-born, has been a focus of academics across a variety of fields. Measures of integration will vary based on the dimension of interest and the data available. For example, in economics, wage convergence or occupational placement, obtained via labor force surveys, are the most commonly used measures (see, for instance, [2]). In sociology or political science, commonly used indices are based on social interaction, language acquisition, residential integration or cultural convergence (see [6]). Again, individual or household level surveys are the most commonly used data collection methods. These types of sources, despite their value, have a major shortcoming: They do not provide high frequency data in terms of a time or space dimension due to the cost and complexity of conducting such surveys.

One type of data that addresses this shortcoming is the use of social big data that is generated by phone records, social media, print media or daily economic transactions. This paper aims to contribute to our knowledge in this direction by combining D4R (see [7] for details) datasets with other big data sources to assess the economic, social and physical integration of Syrian refugees in Turkey. In addition to constructing various geographic integration and communication indices based on D4R, we merge the D4R-based data with real-estate market data and media data to explore their interaction.

While we acknowledge the need for further investigation, we identify several interesting patterns in the data regarding refugee integration. First, we observe heterogeneous segregation across provinces, this heterogeneity appears to be correlated with the size of the refugee population. More specifically, areas with higher refugee shares of the population are, on average, more integrated. Potentially indicating that refugees are settling in areas in which they are more accepted. Interestingly, though, spatial integration is not correlated with more inter-group phone calls. Second, segregation appears to be declining over time, this is to be expected as refugees expand their social networks and become more intertwined in the local economy. Third, segregation tends to be lower during the day than at night, indicating that refugees tend to work more closely to native Turks than where they live. Finally, there are clear linkages with events and residential markets but requires further analysis.

There are numerous policy implications of the observations and the results presented in the paper. Naturally, further analysis, more detailed data and a certain level of policy experimentation are needed to design appropriate policy instruments that

can be employed. These policies, especially those that enable faster and smoother economic and social integration of the refugees, will benefit both the refugees and the host communities. The first policy measure is on labor market access. Even though the data do not reveal any direct information in this regard, formal access to labor markets is shown to be a critical policy measure in many different contexts. ([8]). Our analysis in the paper shows that refugees “daytime" integration is higher relative to the “nighttime" which indirectly indicates labor market integration (proxied by the former) is higher than social or residential integration. Legal access to labor markets will reduce “resentment” among hosts who would otherwise view refugees either as stealing their jobs by working under the table or simply free-riding by receiving welfare checks. Furthermore, refugees can enter many higher skilled occupations that require formal employment, instead of being informally employed in low-skilled occupations. Similarly, in order to improve residential integration and to prevent refugees from living in isolated urban slums, it is necessary to impose laws that punish discrimination against refugees by landlords as well by real estate agents. Another option is to encourage refugees or subsidize their rents in areas where there is more housing but relatively low level of refugee presence. This would also provide a boost to the real estate markets in these areas.

The rest of the report is structured as follows: Section 2 presents some basic measures of communication built from the D4R datasets; Section 3 describes how we have extracted dyadic call propensities from Dataset 2; Sections 4 and 5 present two time-varying and spatially dis-aggregated measures, the EI index and the dissimilarity index respectively; Section 6 combines the D4R data with geo-localized events related to refugees extracted from the GDELT database; Section 7 explores data on the evolution of the housing market in Turkey using price and sales data from local real estate markets; finally, Section 8 presents the concluding remarks, and it describes the scope for future research.

2 Basic measures of communication

Datasets within D4R have different strengths and, as a result, are better suited for different purposes. For example, Dataset 1 is a more complete universe of observations while Dataset 2 has more detailed information for a smaller sample of users (see [7]). Dataset 2 is the only dataset in the D4R collection which contains point-to-point communication where both the caller and the callee have a *refugee* (*R*), or *non-refugee*[2] (*N*) label. Because dataset 2 is the only one to identify inter-group calls, much of the trends derived regarding call patterns are done using these data. Dataset 2 contains a series of 26 two-week long panels (which we call “waves"), following roughly between five to eighteen thousand refugees and fifty to sixty-five thousand non-refugees in a given wave. One observation in this dataset is a *communication* between (1) a sampled individual and (2) another individual that may or may not

[2] At times, we refer to this group also as *natives*.

be part of the sample. A communication is either a phone call or a text message, that may be either outgoing, i.e., initiated by the sampled user, or incoming, i.e., received by the sampled user. For each communication, we know in which province the sampled user was, based on the antenna location. Out-of-sample individuals are labeled as *unknown*; for the purpose of our analysis, we drop communications involving unknown users.

Figure 1 plots the number of outgoing and incoming calls for R and N users in Dataset 2, showing that the number of calls involving N users is relatively stable over time, with an increase between Wave 21 and Wave 24, while the number of calls involving R users follow a steady increase up to Wave 21. This increase reflects a steady growth (again, up to Wave 21) in the number of R users included in the various 2-week samples, rather than in increase in call activity.

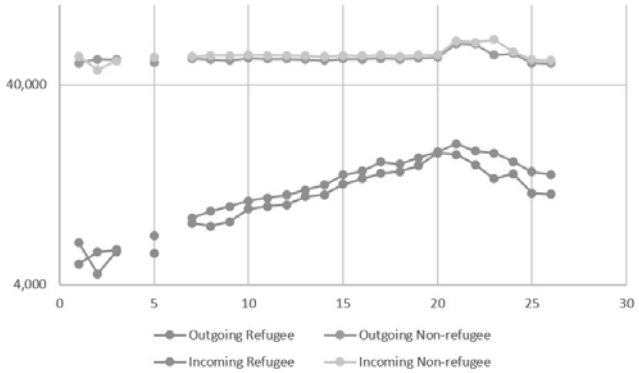

Fig. 1 Number of outgoing and incoming calls for R and N users in Dataset 2; each point corresponds to a different 2-week sample of users; no data are available for Wave 4 and Wave 6; the $y-$axis is in logarithmic scale.

For a high-level overview of Dataset 2, we show in Figure 2 the geographical distribution of calls. We observe a higher concentration of calls in large urban areas and also in the region close to the Syrian border. This is due to the fact that both R and N users were sampled based on the distribution of refugees from official records. If we only consider calls involving R users (see the map in Figure 3), the number of calls decreases but the spatial distribution appears to be similar, with higher concentrations of calls in the same areas as before, and especially around the Syrian border. At the province level, refugee call volume is highly correlated with the official population numbers, though there are a few outliers. For example, Antalya, which has relatively few refugees according to the official numbers (ranked 66^{th} overall) has the eighth highest call refugee call volume. On the contrary, the provinces of Sirnak and Edirne (ranked 21^{st} and 27^{th} in official numbers) rank only 66^{th} and 57^{th} in refugee call volume, respectively.

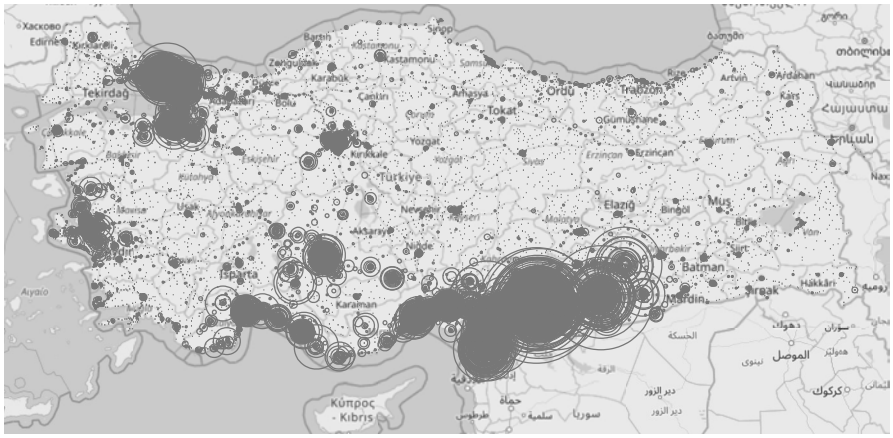

Fig. 2 Geographical distribution of voice calls in Dataset 2. Each circle corresponds to an antenna, and the radius of the circle is proportional to the number of calls made and received by that antenna.

Fig. 3 Geographical distribution of voice calls in Dataset 2, including only calls that involve refugee users. Each circle corresponds to an antenna, and the radius of the circle is proportional to the number of calls made and received by that antenna.

To compare the call activity of refugees and non-refugees, we provide Figure 4, which shows histograms of individual call volumes for two waves, namely Wave 20 (September 25 to October 8) and 23 (November 6 to November 19), which correspond to a high and a low distance between the distributions of the calls for R and N users measured by the Kolmogorov-Smirnov statistic (not reported). A first observation is that, for both groups of users, most of the population is involved in a low number of calls, with a few individuals displaying very large numbers, i.e., heavy-tailed distributions. When comparing R to N, we see that the number of calls made and received are smaller for R, with a smaller fraction of users involved in a very large number of calls. This is difference is larger for incoming calls than

outgoing calls. That is, while refugees and non-refugees make similar numbers of outgoing calls (it is still greater for non-refugees), refugees receive much fewer calls than do non-refugees.

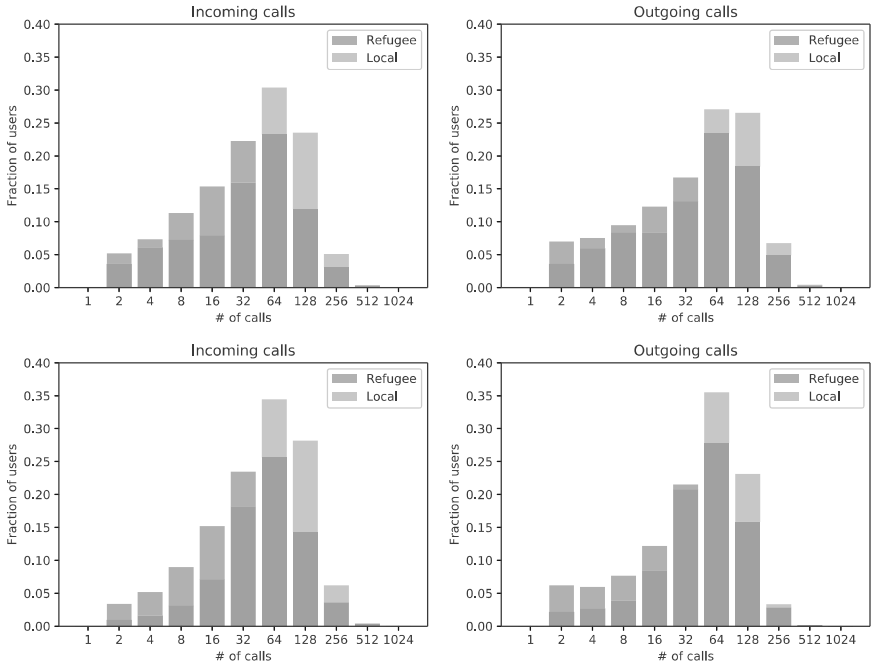

Fig. 4 Activity patterns for Wave 20 (top) and 23 (bottom). The plots show normalised histograms of the number of calls per user for *R* and *N* groups, separated into outgoing and incoming call; please note the logarithmic bin size.

Figure 5 presents the smoothed daily number of calls between each of the possible four pairs of caller and callee from Dataset 2, separately for calls recorded in the outgoing and in the incoming portions of the dataset. Daily calls are normalised so that yearly averages become 1 for each of the four possibilities. Figure 5 shows an increase in the number of calls where the caller is a *R* user. Over time, the share of calls going to refugees, from both other refugees as well as non-refugees, is also increasing. It is unclear whether this trend reflects integration of refugees over time, an increase in the overall refugee population, or simply an increase in the refugee population of Turk Telekom users. Finally, for both outgoing and incoming calls, there is a sharp (and persistent) drop in the number of calls around mid-2017. We are not clear about the cause of this sudden change, but the timing is consistent with changes in other measures we present in subsequent sections.

Figure 6 reports the actual and smoothed number of outgoing calls for *R* and *N* users on a daily basis from Dataset 1. The timeline of the call density of *R* users is in line with the one emerging from Dataset 2 (see Figure 5). The consistency of

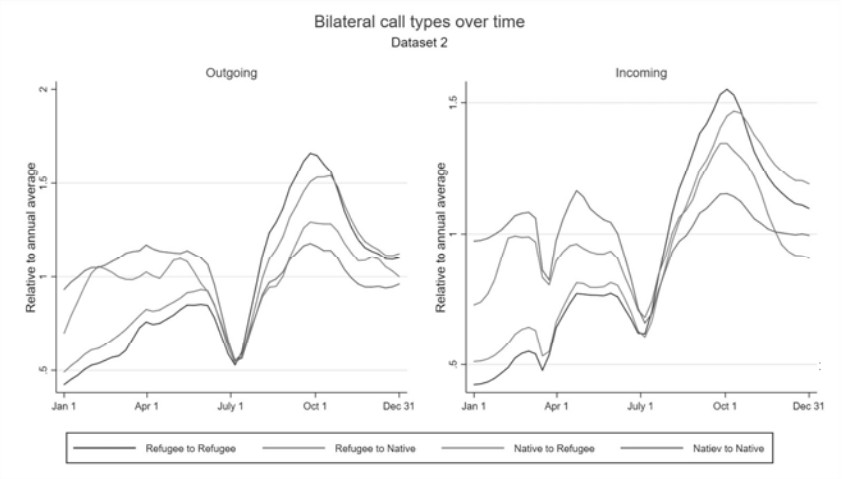

Fig. 5 Smoothed normalized daily number of outgoing (left panel) and incoming (right panel) calls between a caller $g \in \{R, N\}$ and a callee $h \in \{R, N\}$ over 2017; the average number of each type of dyadic calls over the year is normalized to 1; there are 82 days for which the Dataset 2 contains no data

the two datasets is important, as we combine the information coming from Dataset 1 and 2 in Section 4.

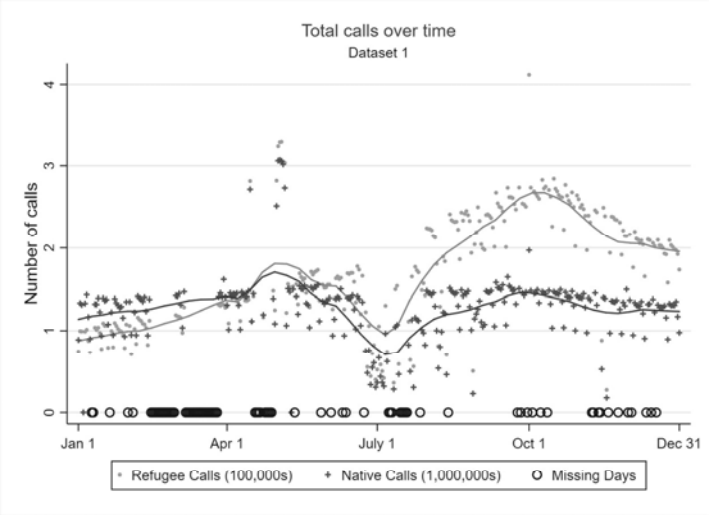

Fig. 6 Actual and the smoothed daily number of outgoing calls for R and N users over 2017; calls for R and N users are reported using a different scale; there are 82 days for which Dataset 1 contains no data.

3 Call propensities

Dataset 2 provides information on the status (R, N or U for unknown) for both the caller and the callee. The Dataset is partitioned into outgoing or incoming call depending on whether the user included in the sample is the caller or the callee. These two portions of Dataset 2 allow us to estimate the propensity of each type of call (R-to-R, R-to-N, N-to-R and N-to-N) for an average user in a given wave, these propensities are estimated separately for both incoming and outgoing calls.[3] This analysis is done at the province level, providing 8 propensities measures for each province-wave pair. A simple analysis of these propensities lead to a few basic facts. First, a majority of calls are made to non-refugees, this is true of both non-refugee and refugee users. Second, non-refugees make and receive calls at higher rates than refugees, making around 60 outgoing calls over a two-week period as opposed to refugees who make around 40, as we also saw in Figure 4. Also, the probability that a refugee calls another refugee is directly related to the number of refugees in a given area; Figure 3 plots the share of R-to-R calls over all the calls made by R users for each province using the data from Wave 22 (October 23 to November 5), provinces where a larger number of refugees are located tend to have higher rates of refugee-to-refugee calls (see Figure 3). These propensities will also be combined with the data on the antenna traffic (separately by type of user) from Dataset 1 to obtain a time-varying estimate of the number of R and N users at various degree of spatial resolution.

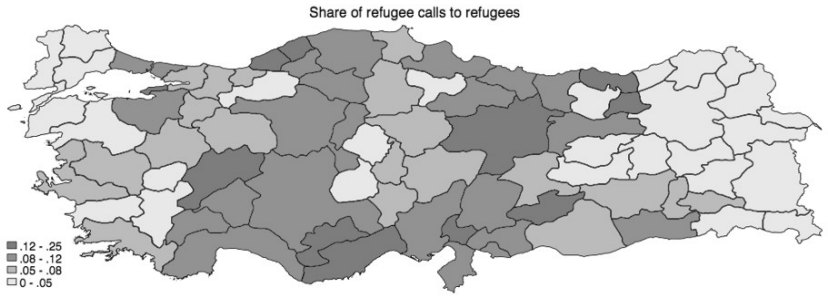

Fig. 7 Share of R-to-R calls among all calls made by R users with data from Dataset 2 (Wave 22).

[3] The propensity is defined as the average number of calls a type of user performs towards another type of user.

4 The EI index

We rely on the EI index, introduced by [5] and originally proposed to measure homophily to analyze the frequency of calls across the R and N groups. We compute the index for every month (more precisely, 13 four-week periods). For a given province i in month $t = 1, ..., 13$, and for $g, h \in \{R, N\}$, we define $m_{it}(g, h)$ as the number of communications where the sampled user was located in province i and belongs to group g, while the other end of the communication belongs to group h (his or her location is unknown). We can also define $m_t(g, h)$, equivalently defined at the national level.

The EI index is defined as the ratio between the difference of between-groups (or external) and the within-groups (or internal) calls over the total number of calls:

$$EI_{it} = \frac{m_{it}(R,N) + m_{it}(N,R) - [m_{it}(R,R) + m_{it}(N,N)]}{m_{it}(R,N) + m_{it}(N,R) + m_{it}(R,R) + m_{it}(N,N)} \tag{1}$$

We clearly have that $EI_{it} \in (-1, 1)$, with low values indicating few connections between groups, while high values indicate many connections between groups, i.e., better integration.

In Figure 8, we plot for each t the distribution of EI_{it}.[4] Overall, the distributions are strongly centered around negative values, indicating few between-groups communications. Still, we must remain cautious in interpreting the absolute values of this index. $EI_{it} = 0$ is equivalent to having the same amount of communications between groups and within groups, but given the strong imbalance between group-sizes, we can only expect negative values. However, the distribution seems to evolve toward more integration, or at least to have a larger dispersion overtime.

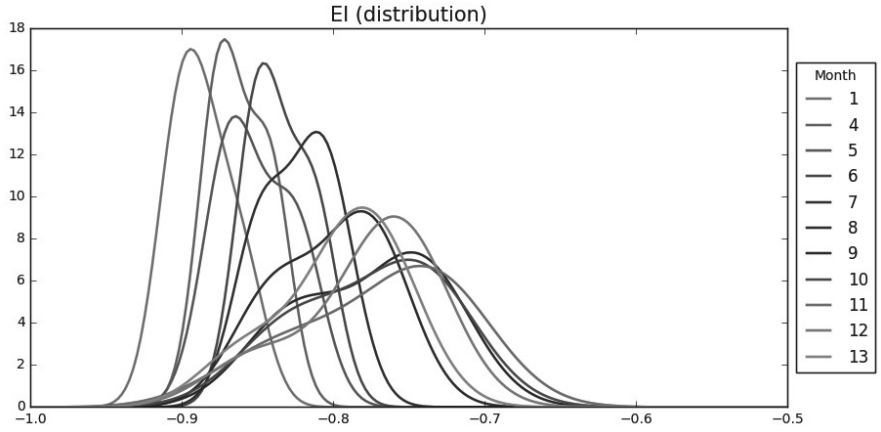

Fig. 8 Histograms showing the distribution of the EI in the 82 provinces. Each line corresponds to one 4 week time period.

[4] Data are mostly missing for "month" 2 and 3 so they are dropped.

In Figure 9, we plot the evolution of the EI index overtime for the five largest cities and for the index computed at the national level. The same general trend appears: integration seems to improve overtime. However, while the time-window is too narrow to make any definitive statement, all series suggest that this increase eventually stabilizes if not reverses at the end of the year.

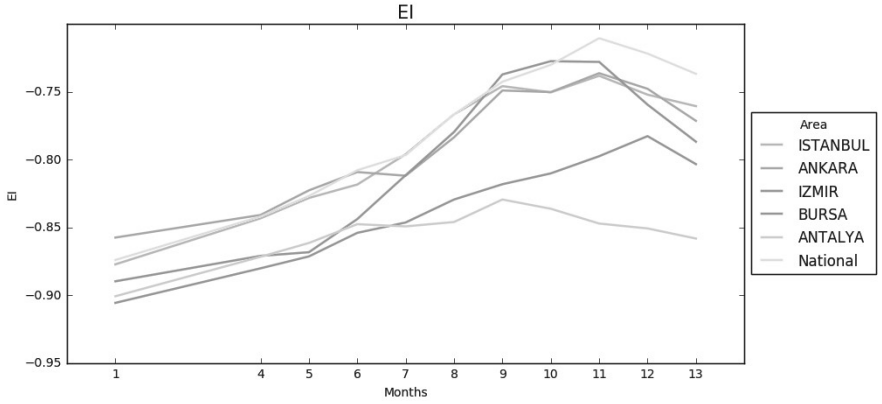

Fig. 9 Time evolution of the EI for the 5 provinces with the largest sample sizes.

It is important to mention, however, that part of this increase in integration could be an artifact of the change in the ratio of R and N users over time. An increase in the (absolute and relative) number of calls involving R users (see Figure 5) entails a reduction in the EI index, as the number of within-group calls for N users remains unchanged, while the three other types of calls increase.[5, 6]

5 Dissimilarity Indices

The most commonly used metric of segregation is the index of dissimilarity (D) originally introduced by [3] and [4]. The basic formula for the index is given by:[7]

$$D_i = \frac{1}{2} \sum_j \left| \frac{r_{ij}}{r} - \frac{n_{ij}}{n} \right| \tag{2}$$

[5] This happens even if we allow the number of within-group calls for R users, i.e., $m_{it}(R, R)$, to be a quadratic function of the number of R users in the sample.

[6] We have computed an Herfindahl-Hirschmann Index of the concentration across Turkish provinces of R users in the sample for each two-week sample in Dataset 2; higher (lower) number of R users in the sample are associated with a lower (higher) value of the Herfindahl-Hirschmann Index, revealing a weaker (stronger) concentration.

[7] Eq. (2) omits the time subscript, but the index D_i is actually time-varying.

where r_{ij} is the population of group R in the $j-$th area of the province i. and r is the total population of the group in the province (n_{ij} and n are similarly defined). In our context, provinces (such as Istanbul or Mardin) are the regions and each area j is the catchment area of each cell-tower within a province. The dissimilarity index for a province i, D_i, is a measure of the evenness of the distributions of the two groups across the area of that province. We can interpret the index as the percentage of a group's population that would have to move to obtain the same percentage of that group within the overall province. The index D_i ranges from 0.0 (complete integration) to 1.0 (complete segregation). Notice that the index D_i is unaffected by a time-varying size of the sample of R users, provided that the spatial distribution of the samples is uncorrelated with their size. Under random assignment, D_i will still be greater than zero as population sizes will vary slightly due to random variation. To test the extent of this factor, we calculate the segregation in call volume for the roughly 1.3 million non-refugee callers that appear in all waves of dataset 2 by randomly assigning the users to two equally sized groups. Under this scenario, the 'random' level of dissimilarity averages .11 across all provinces, ranging from .16 in the most segregated province and .03 in the least segregated. As is shown further down, this level 'random' segregation is significantly lower than the amount observed between refugees and non-refugees.

Even though Dataset 2 has call volume of the individual towers from which calls are originating, we choose to use Dataset 1 to measure segregation. Dataset 1 has a significantly larger sample size and there simply are not enough (or even any) observations for a majority of the towers on a given day in Dataset 2. This sampling bias would distort our analysis considerably. The downside of Dataset 1 is that we only know the number of calls from a given tower, not the number of people. In order to link the two databases, we take the provincial level propensities that we calculated from Dataset 2 (R-to-R, R-to-N, N-to-R and N-to-N) for each of the 26 waves. Then, we divide the number of calls originating from each tower in each time period for each group (R or N) by these propensities to estimate the number of refugees (R) and non-refugee (N) populations for each tower area for each time slot, i.e., r_{ij} and n_{ij} in Eq. (2) above. Populations are calculated by dividing the total call volume over a two week period by the province level propensities calculated in Section 3 (using dataset 2). The main assumption of this approach is that call propensities are constant across antennae within a province.

One of our key innovations is that we calculate the dissimilarity index for working hours and non-working hours separately[8]. This distinction allows us to compare and comment on residential segregation and employment segregation between the refugee and non-refugee populations in each province. Figure 10 below plots the Dissimilarity Index D_i for each of the 26 waves of 2017 for both the working and non-working hours. The plot is a weighted average of all provinces in the country. There are two immediate observations. The first is that there is a certain degree of dissimilarity (or segregation) between the refugees and non-refugees but it is declining over time. The temporary jump during the waves of 11-16 corresponds to the

[8] Working hours are defined as 8 am to 5 pm, all other hours are assigned as non-working.

time frame where the number of phone calls (in both datasets) decline significantly (see Figures 5 and 6 above). We suspect this is due to the biased sampling issues that need to be explored further. The second observation is that the dissimilarity index for the working (day) hours is always below that of the non-working (evening and nighttime) hours. This pattern indicates that refugees are more integrated in terms of their work and employment relative to residences.

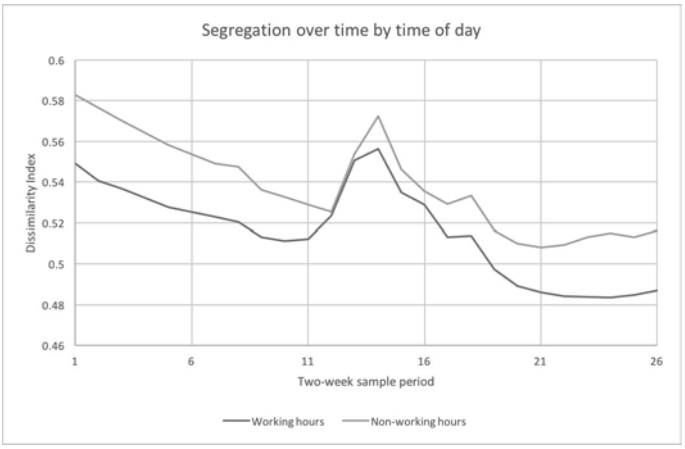

Fig. 10 Evolution of the segregation index D_i; the figure reports the country-level evolution of the segregation index D_i defined in Eq. (2), separately for day (8 am to 5 pm) and night time hours.

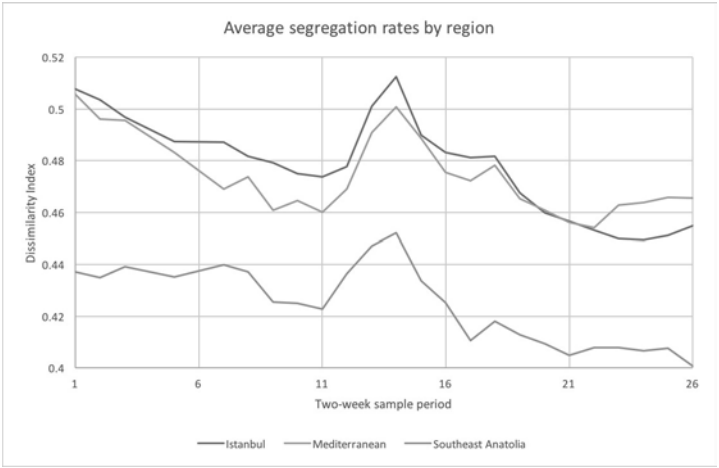

Fig. 11 Evolution of the segregation index D_i for Istanbul, Mediterranean provinces and Southeast Anatolia.

As mentioned above, the dissimilarity index is calculated at the province level over time. Figure 11 presents the index for three different geographic areas: (i) Istanbul, home to over half a million refugees, (ii) Mediterranean coast (provinces of Adana, Antalya, Burdur, Hatay, Isparta, Mersin, Osmaniye) and (iii) Southeast provinces, mostly along the border (Adiyaman, Batman, Diyarbakir, Gaziantep, Kahramanmaras, Kilis, Mardin, Siirt, Sanliurfa, Sirnak) using calls from all hours of the day. The declining segregation index over time for all regions are observed in this figure as well. Another striking observation is that the Southeast region, where the refugees make up the largest share of the total population, has significantly lower degree of segregation than the other regions. This could be because, due to their proximity to the Syrian border, non-refugees in those areas already have a greater familiarity with the Syrian population. Or perhaps the refugee populations along the borders represent the earliest to arrive and thus have had more time to integrate into their host communities.

Figures 12 and 13 show dissimilarity indices for the 40 provinces with the highest share of refugees (according to official numbers) for the first and last time periods of 2017. The maps are color coded so that darker shades show the higher levels of segregation. Figure 12 is for the first wave of the year (Jan 1-15, 2017), while Figure 13 is for the last wave (Dec 18-31, 2017). We can see that the southeast provinces are lighter in color than the rest of the country. Furthermore, the overall map for the last wave is much lighter in color, indicating all provinces became more integrated over time.

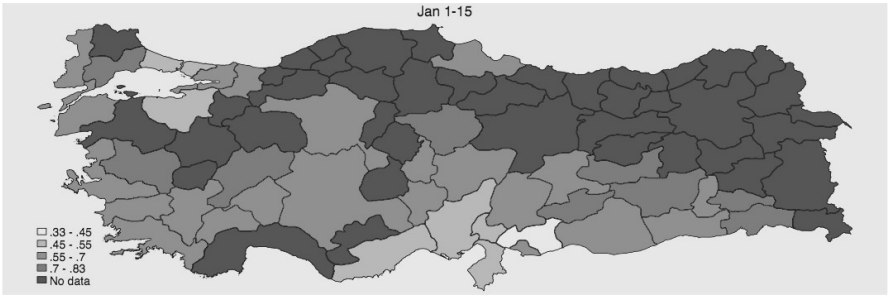

Fig. 12 Province-level measure of the segregation index D_i defined in Eq. (2) computed with the data from Wave 1.

The dissimilarity index D_i in Eq. (2) is the leading index among a large set of indices that have been constructed and analyzed over the last four decades of active research on residential segregation of different communities in many different countries, cities and regions. There are numerous indices that measure other dimensions of communal interaction and integration. Among these are the isolation index (measuring the extent to which minority members are exposed only to one another, see [1]), concentration index (measuring the relative amount of physical space occupied by a minority group in the metropolitan area, see [6]), centralization index (measur-

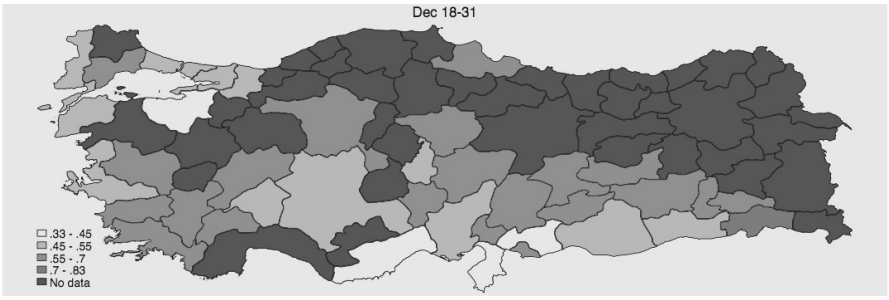

Fig. 13 Province-level measure of the segregation index D_i defined in Eq. (2) computed with the data from Wave 26.

ing the degree to which a group is spatially located near the center of an urban area, see [6]), or clustering index (measuring the extent to which areal units inhabited by minority members adjoin one another, or cluster, in geography, see [6]). We have performed preliminary analysis of these indices with each one providing important insights on the geographic and social distribution of Syrian refugees within Turkey. They are not included in this report due to space constraints but the same trends across space and time are evident across all indices.

We would also like to identify which province characteristics explain high or low segregation of refugees. For example, preliminary results (see Table 1) indicate that high-refugee provinces (as a share of their total population), as well as larger provinces, experience lower levels of segregation as compared to other low population or low-refugee provinces. Interestingly, when we control for refugee share and overall population, there is no relationship between the amount of cross-group calling and segregation levels.

6 Global Database on Events, Language and Tone (GDELT)

The arrival of Syrian refugees in Turkey is a dramatic social and cultural event with important political ramifications, for the Syrian refugees, Turkey as well as the rest of the world. The geographic and time dimension of the phone call data can be exploited to measure the linkages between concentration of refugees, their social interaction and political events. For this purpose, we integrate the D4R data with another unique database - Global Database of Events, Language, and Tone (GDELT), which we use to measure the extent of refugee-related events across both time and space. GDELT collects news media articles from around the globe in over 100 languages, going back to 1979. Each media observation is classified into an event data, a form of data common in political science to study political history in a systematic way. Events are classified by location, a set of actors (e.g., governments, NGOs, refugees, private companies, etc.), a set of actions (e.g., announcements, diplomatic meetings,

Table 1 Simple regression of the segregation index D_i

	D_i
Share of outgoing calls made by R to N	0.015
	(0.012)
Share of outgoing calls made by N to R	-0.014
	(0.012)
Ln(population)	-0.037**
	(0.014)
Refugee share of population	-0.252***
	(0.082)
Observations	40
R^2	.369

Notes: Table displays results from a linear regression on a single cross-section of the 40 largest refugee provinces using data from the first two-week period in January, 2017. ***, ** and * denote significance at the 1, 5 and 10 percent level respectively; standard errors between parentheses; call shares transformed to have a mean equal to 0, and a standard deviation equal to 1.
Source: Authors' elaboration on Datasets 1 and 2, refugee and overall population data comes from Turkish Ministry of Interior.

accidents, etc.) as well as other information that attempts to predict the tone and impact of an event.

For our purposes, we queried all events from January 2016 to June 2018 that were located in Turkey and included refugees as at least one of the actors. This query yielded 119,000 events over the 2.5-year period, although only 22,431 events occur in 2017 (the year which overlaps with the D4R phone data). Of those events in 2017, 9,498 include a specific province in which the event occurred. Other events are either national in nature without specific assignment to a province or the GDELT text-processing algorithm was simply unable to assign a location.

Using the events data, we constructed a daily panel of events across 81 Turkish provinces. Observations include the number of daily events as well as the average tone of events (tone is calculated from a textual analysis of the media article and is done by GDELT.) We also include a weighted measure of events and tones in which the weight of each event is calculated as the square root of the number of news articles that mention an event. Figure 14 presents the distribution of the events that were extracted from GDELT for the whole country. Spatially, events are most prevalent in Istanbul, Ankara (as the political center of Turkey), the southeast region of Turkey which borders Syria, as well as western regions along the Aegean coast (coinciding with common departure points of refugees attempting to enter Europe). There is also significant variation over time; a substantial portion of events occur in the first three months of 2017 and there are important surges in June and September (see Figure 15).

The critical feature of the GDELT dataset is that it has both the time and space dimension and can be matched with the D4R datasets. There are several potential

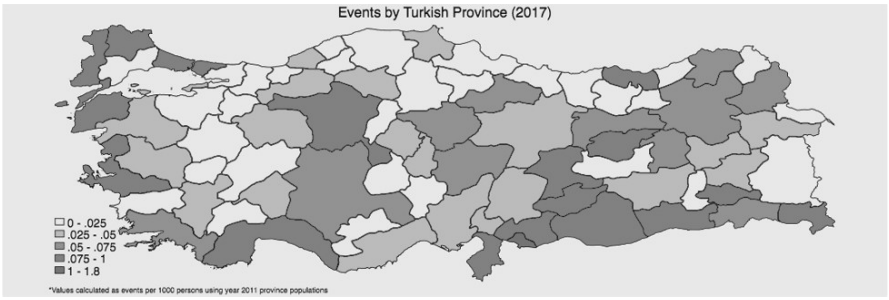

Fig. 14 Events by Turkish province in 2017

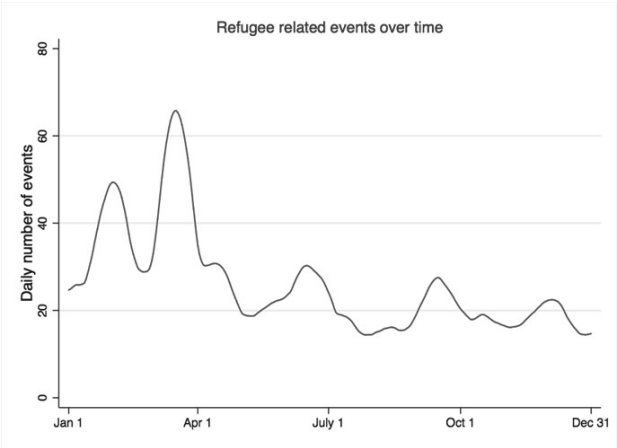

Fig. 15 Daily number of events from GDELT

paths we can follow, linking the two datasets. For example, Figure 16 below plots (natural log of) number of calls to the (natural log) number of events in GDELT where each dot represents (binned) province-day level of observations. The plot shows that, controlling for province-level effects, refugee-related events are correlated with increased call volume. Further analysis (not shown) implies that this increase is driven by native, rather than refugee, call volume. These results are also robust to removing Ankara from the analysis (because of it's political importance, Ankara is an area associated with roughtly 40% of refugee related events). We can go further and link the call propensities (R-to-R, R-to-N, N-to-R or N-to-N) or the dissimilarity/segregation indices with the GDELT indices we constructed. In addition to the number of events in GDELT database, another valuable measure is the emotional tone of the events. This feature is especially informative on a topic such as refugees and their social and economic integration in the host community. The whole issue is highly charged in terms of politics and emotions and this dimension is one of the key issues we intend to explore further.

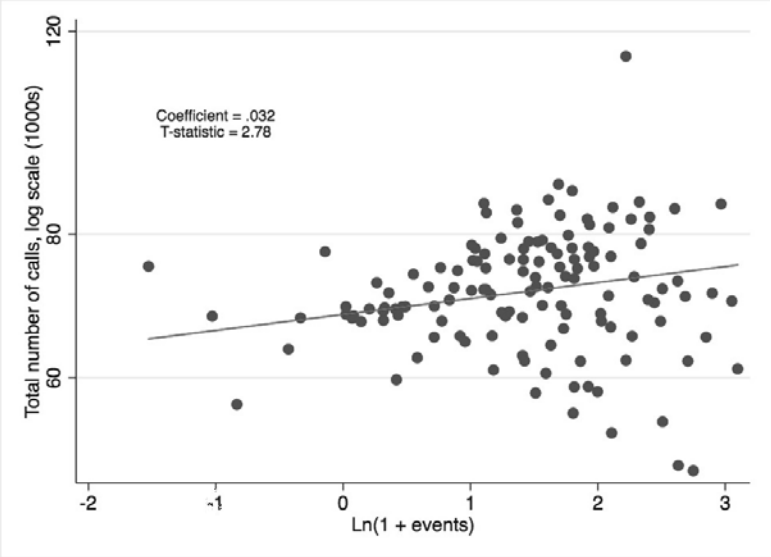

Fig. 16 Events from GDELT and calls from Dataset 1; the figure shows a binned scatterplot of the log number of events the against the log number of phone calls at the province-day level and the corresponding linear best fit line, only including observations that experience more than 10,000 calls per day. Plot controls for province-level fixed-effects

7 Housing data

Economic and cultural assimilation of refugees depends critically on where they live and work. Section 5 showed the existence of segregation between the refuges and the non-refugees, with considerable variation across provinces. Furthermore, we saw that segregation was declining over time across all provinces in the period covered by the D4R datasets.

In order to further explore the determinants of these integration/segregation patterns, we turn to data on Turkish real estate markets.[9] The data includes monthly indices for both rental and sales prices for close to 1,000 distinct real estate markets across the country. Some of these markets are at the provincial level (for smaller provinces) and others are at the neighborhood level for big cities like Istanbul. For the time being, we aggregated their real estate sales and rental price data to the provincial level but the data would allow us to conduct quite disaggregated analysis taking advantage of the geographic distribution derived from the D4R dataset. For 62 of 81 provinces, our indices begin in 2012 or earlier (before the largest inflows of Syrians began), while the remaining 19 indices do not begin until 2015. In addition to price data, there is also data on residential sales volume, again, at the provincial

[9] The data come from REIDIN Data and Analytics, a leading provider of real estate data and information for emerging markets, under a confidentiality agreement.

level. The sales data include the monthly number of sales disaggregated by primary and secondary sales, which represent new construction and resale of existing houses, respectively. These indices begin in 2013 for all provinces and are based on government registration records.

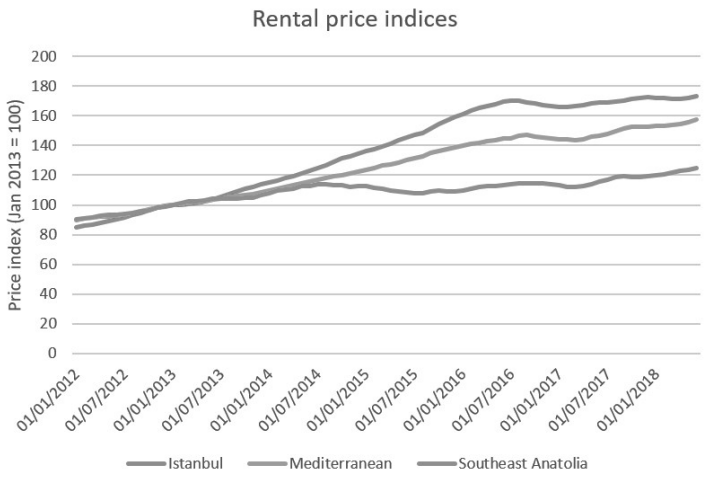

Fig. 17 Evolution of rental price indices for Istanbul, Mediterranean provinces and Southeast Anatolia

A cursory look at the data indicates a distinct break in trend between high and low refugee areas beginning in 2014 among both prices and volume. Figure 17 presents the rental price indices for three regions of the country–Istanbul, Mediterranean coast and the Southeast Anatolia along the border. The surprising observation is that prices in the Southeast, the region with the largest relative number of refugee inflows, have trended below the other regions since 2014. The price difference between Istanbul and the Southeast increased by more than 50 percentage points between 2014 and 2018, even though they were following a nearly identical trend prior to 2014. Given the sharp increase in demand due to the refugees, we would expect the opposite trend and is not consistent with a sharp housing demand shock.

There are a few forces that can explain this rapid and surprising price divergence between low and high-refugee markets. We believe this phenomenon is explained by rapid supply response and changing composition of housing quality. Figure 18 shows the sales volume of primary housing markets respectively (sales of new construction) where sales in Southeast Anatolia increased drastically compared to other regions. If the Southeast Anatolia region had followed the same path as the comparable Mediterranean region, it would have experienced 16.9 thousand fewer primary market sales. This rapid increase is indicative of a sharp positive supply response of the construction sector. Similarly, Figure 19 shows the sales in the

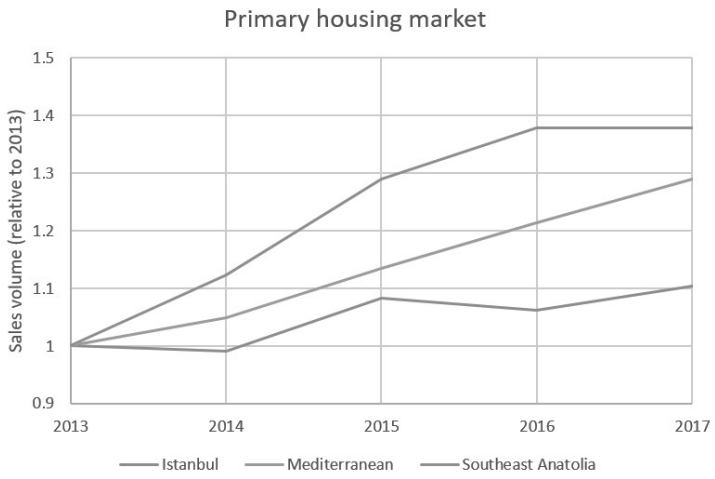

Fig. 18 Primary housing market for Istanbul, Mediterranean provinces and Southeast Anatolia.

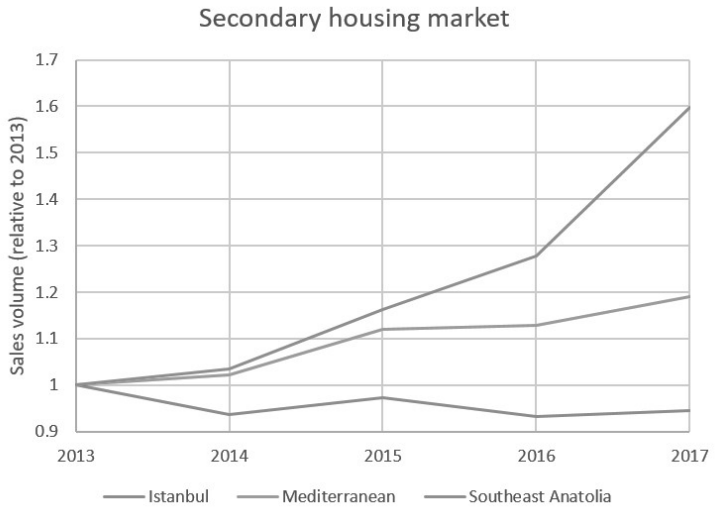

Fig. 19 Secondary housing market for Istanbul, Mediterranean provinces and Southeast Anatolia.

secondary market (of existing homes) where we again see rapid increase in sales. When we look at the prices in the secondary market, we again see a decline, implying increased sales of lower quality homes.

Our next step is to link the segregation indices with rental/sales price data to identify the causal links between real estate markets, integration and social interaction of refugees.

8 Conclusion

The analysis presented in the previous sections reveals that Syrian refugees in Turkey have become more integrated (in terms of communication) and less spatially segregated over the period covered by the D4R Challenge, albeit the various measures of integration (notably, the EI and dissimilarity indices) exhibit a certain degree of spatial variation across provinces. In terms of specific results, we find that the communication between refugees and non-refugees increased over time as indicated by the propensities to call each other. Similarly, spatial segregation of refugees as measured by the dissimilarity index has declined, especially in provinces where refugees make up a higher share of the population. Finally, spatial segregation during the day is lower than at nighttime, implying labor market segregation is lower than residential one. All of these measures indicate improved integration of the refugees into the society.

We performed two additional analyses using GDELT database on events and Reidin database on real estate prices. Both of these analyses were more exploratory in nature, highlighting the possible research avenues while providing preliminary results. GDLET data show there is positive correlation between events and call volume while the housing data reveal that real estate prices did not increase as much as expected, possibly due to increase construction.

The value of D4R dataset for academic research and policy evaluation can be significantly increased by extending the amount of information included in the D4R datasets. For example, a more detailed description of the data collection and sampling procedures would be useful, and possibly by including a larger sample of the non-refugee population. Since the results depend highly on the way natives and refugees were selected to be included in the in the D4R sample, any bias in the sampling procedure will influence results. Furthermore, it would be useful to be able to extract all the calls initiated by R/N users in given province since this is the only dataset that has information on point to point (R to N) communication. We are hopeful that the path paved by this initial D4R dataset will stay open and data from later years will also be made available to explore critical economics, social and cultural integration issues of refugees. The lessons learned will not only be useful for the Syrians in Turkey but for millions of other refugees all over the world.

Acknowledgements This work was supported by the European Commission through the Horizon2020 European project “SoBigData Research Infrastructure — Big Data and Social Mining Ecosystem" (grant agreement 654024).

References

[1] Bell W (1954) A probability model for the measurement of ecological segregation. Social Forces 32(4):357–364
[2] Chiswick BR (1978) The effect of Americanization on the earnings of foreign-born men. Journal of Political Economy 86(5):897–921
[3] Duncan OD, Duncan B (1955) A methodological analysis of segregation indexes. American Sociological Review 20(2):210–217
[4] Duncan OD, Duncan B (1955) Residential distribution and occupational stratification. American Journal of Sociology 60(5):493–503
[5] Krackhardt D, Stern RN (1988) Informal networks and organizational crises: An experimental simulation. Social Psychology Quarterly 51(2):123–140
[6] Massey DS, Denton NA (1988) The dimensions of residential segregation. Social Forces 67(2):281–315
[7] Salah A, Pentland A, Lepri B, Letouzé E, Vinck P, de Montjoye Y, Dong X, Dağdelen O (2018) Data for Refugees: The D4R Challenge on mobility of Syrian refugees in Turkey. arxiv preprint arxiv:1807.00523.
[8] World Bank (2018) Moving for Prosperity: Global Migration and Labor Markets (Policy Research Reports). World Bank Publications, URL http://www.worldbank.org/en/research/publication/moving-for-prosperity