Rhoads et al. *EPJ Data Science* (2020) 9:5
https://doi.org/10.1140/epjds/s13688-020-00222-1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE

Open Access

Measuring and mitigating behavioural segregation using Call Detail Records

Daniel Rhoads[1], Ivan Serrano[1], Javier Borge-Holthoefer[1] and Albert Solé-Ribalta[1,2*]

*Correspondence:
asolerib@uoc.edu
[1]Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Barcelona, Spain
[2]URPP Social Networks, University of Zurich, Zurich, Switzerland

Abstract

The overwhelming amounts of data we generate in our daily routine and in social networks has been crucial for the understanding of various social and economic factors. The use of this data represents a low-cost alternative source of information in parallel to census data and surveys. Here, we advocate for such an approach to assess and alleviate the segregation of Syrian refugees in Turkey. Using a large dataset of mobile phone records provided by Turkey's largest mobile phone service operator, Türk Telekom, in the frame of the Data 4 Refugees project, we define, analyse and optimise inter-group integration as it relates to the communication patterns of two segregated populations: refugees living in Turkey and the local Turkish population. Our main hypothesis is that making these two communities more similar (in our case, in terms of behaviour) may increase the level of positive exposure between them, due to the well-known sociological principle of homophily. To achieve this, working from the records of call and SMS origins and destinations between and among both populations, we develop an extensible, statistically-solid, and reliable framework to measure the differences between the communication patterns of two groups. In order to show the applicability of our framework, we assess how house mixing strategies, in combination with public and private investment, may help to overcome segregation. We first identify the districts of the Istanbul province where refugees and local population communication patterns differ in order to then utilise our framework to improve the situation. Our results show potential in this regard, as we observe a significant reduction of segregation while limiting, in turn, the consequences in terms of rent increase.

Keywords: Mobile phone data; Segregation; Refugee integration; Residential mixing

1 Introduction

The analysis of segregated communities, due to its important implications for the lives of citizens [1, 2] and for social cohesion [3], has held the attention of policy-makers and academics in the field of social and urban sciences for some time. Segregation can have many dimensions [4], and may come with many different faces: spatial [5], economic [6], occupational [7], gender based [8–10], religious [11, 12], ethnic [13] and etc. Along with the quantification of segregation there is a long-standing debate on how to promote social cohesion in ethnically diverse environments, which often focuses solely on the effects of spatial segregation [14]. However, a broad perspective on the effects of housing and urban

Springer

planning suggests that top-down public policies centred exclusively on spatial mixing are not fully effective at promoting social integration [15]. After the social unrest experienced in the US and Europe in the early 2000s, policies turned to the need to foster social and community cohesion [16] as well. Toward this objective, traditional census data face three main limitations: difficulties in reaching certain segregated groups which are reluctant to participate, the practically inexistence of social data (necessary to assess non-spatial aspects of segregation), and high economic cost for the administration. To overcome this situation, many countries have increasingly incorporated alternative sources of information from administrative registers or sample surveys [17]. In this lines, the consolidation of ICT, mobile phones and social networks offers an outstanding possibility to complement or even provide higher quality data in a variety of public policy areas, even though the use of Big Data also raises a number of methodological and moral concerns [18] related to privacy or biases embedded [19, 20] in the information collected by social networks and machine learning algorithms, e.g. the use of satellite images to quantify poverty [21, 22].

Here, we contribute to the segregation analysis debate by defining, analysing and optimising integration as it relates to the communication patterns across groups. Specifically, we are interested in studying the variation in communication patterns between different segregated communities. It is our claim that, in order to integrate separate communities, it is not enough to simply bridge the gap between individual characteristics (e.g. social, economic and occupational) and spatial distributions, but that communities should also be similar in the way they interact with each other. In this regard, the main assumption we rely on is that, if two communities are equally distributed across a territory (that is, if there is no spatial segregation) and their behavioural patterns are similar, their calling patterns should also be similar, based on a sort of natural equilibrium derived from the geo- and socio-economic situation of the territory. We further assume that an increase in the similarity of the two group's average behaviour could have a positive impact on their level of interaction, following the principle of homophily in social interactions [23]. This approach to behavioural segregation (as opposed to purely static aspects such as residential, spatial segregation) is paralleled in studies of the mobility patterns of segregated groups (*activity-space* segregation) [24], and in other works focusing on levels of contact between groups [25].

For the experimental part, we work with Call Detail Records (CDRs) of Syrian refugees in Turkey, currently the largest refugee population of any country in the world [26]. The situation in Syria remains unstable, and with policies in force to prevent refugee out-migration to the European Union, most analysts agree that actions must rely on the assumption that refugees will remain in the country for the long-term. Under this problematic situation and in the frame of the Data 4 Refugees project, Turkey's largest mobile phone service operator, Türk Telekom, has released a large collection of CDRs with information about the nationality of the citizen who makes the call (Turkish or Syrian refugee), as well as the origin and destination of the call. After a strong data pre-processing effort (Sect. 3.1), we first perform an analysis to quantify the current behavioural segregation in Istanbul on the basis of communication patterns. This analysis provides us with a picture of the current situation, and presents the base scenario on which to improve. Subsequently, we develop a methodology to mitigate segregation based on house mixing strategies. Results show that behavioural segregation can be largely reduced. However, such a change could have side-

effects on other social aspects, such as rent prices. This is analysed in the final sections of the paper. We conclude the paper with a summary and a discussion of the obtained results.

2 Quantifying existing levels of spatial segregation

The development and presence of enclaves is a common phenomenon within immigrant communities. An immigrant enclave is an expression of spatial segregation as defined by the *Dissimilarity Index* [28, 29] or by Louf et al. [30], who quantify segregation in terms of the deviation from the random distribution of populations in an area. As expected, refugees are not distributed equally across space in Turkey. Figure 1A plots the ratio of refugee to local population in Istanbul's 39 districts, from census data. Along with other plots in the paper, the x-axis indices correspond to Istanbul's districts according to Table 1. Refugees are over-represented in districts with ratios above the horizontal red line (representing the average ratio of refugee-local of the Istanbul province), and under-represented in districts with ratios below the same line. Additionally, we calculate the Dissimilarity Index from the same data. The value obtained (around 30%) is not small, but nor is it as large as one might expect in spatially segregated community [31] (e.g. values between 0.50 and 0.6 were found for the geographic segregation of the black and white populations in U.S. cities in 2000). Panels B and D of the same figure map the distribution of the Turkish and refugee populations, respectively. The maps confirm that most of refugee enclaves are

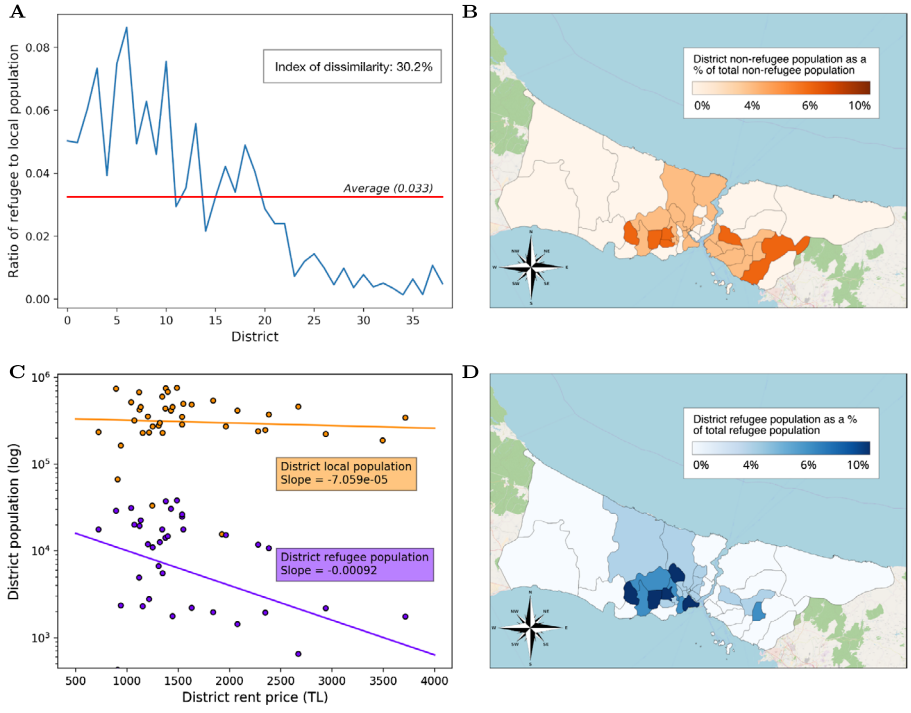

Figure 1 Segregation analysis of the Istanbul province. Panel **A** illustrates the ratio of refugees to the local Turkish population across the 39 districts of the Istanbul province. The horizontal red line indicates the average ratio for the entire city. Deviations from this line indicate district-level segregation. Panel **C** shows the relationship between rent price paid for residence between Turkish and Syrian refugees. For Turkish, we obtained a slope of –7.059e–05 and with 95% confidence bounds of –0.0004516 and 0.0003104. For refugees, we obtained a negative slope of –0.0009222 with 95% confidence bounds of –0.001607 and –0.0002374. Data has been obtained from www.endeksa.com [27]. Panels **B** and **D** show, illustrated in map form, the distribution of the local Turkish and refugees populations, respectively, across the Istanbul province

Table 1 Correspondence between district names and id's

Id	District name	Id	District name	Id	District name
1	Adalar	14	Büyükçekmece	27	Maltepe
2	Arnavutköy	15	Çatalca	28	Pendik
3	Ataşehir	16	Çekmeköy	29	Sancaktepe
4	Avcılar	17	Esenler	30	Sarıyer
5	Bağcılar	18	Esenyurt	31	Şile
6	Bahçelievler	19	Eyüp	32	Şişli
7	Bakırköy	20	Fatih	33	Sultanbeyli
8	Başakşehir	21	Gaziosmanpaşa	34	Sultangazi
9	Bayrampaşa	22	Güngören	35	Tuzla
10	Beşiktaş	23	Kadıköy	36	Ümraniye
11	Beykoz	24	Kağıthane	37	Üsküdar
12	Beylikdüzü	25	Kartal	38	Zeytinburnu
13	Beyoğlu	26	Küçükçekmece	39	Silivri

concentrated in the West-Center part of the province. On the contrary, we observe in the Eastern part of the province a comparatively smaller proportion of the refugee population. Panel C of Fig. 1 provides complementary information to the spatial segregation analysis and shows that indeed Syrian refugees tend to live in cheaper, thus less favourable, neighbourhoods. We see that while there is practically no relation between district rent prices and Turkish population, there is a negative relation (slope p-value < 0.01) between rent prices and population of refugees.

The analysis shown in this section presents an initial picture of the extent to which Turkish and refugee citizens are segregated spatially within Istanbul province. The unexpectedly moderate results of the Dissimilarity Index open up the possibility of exploring other possible, non-spatial measures of segregation, such as one sensitive to behavioural differences. The development and implementation of such a measure of behavioural segregation, through the analysis and comparison of group communication patterns, will be the subject of the rest of the paper.

3 Measuring behavioral segregation through communication pattern analysis

Segregation is usually, with a few exceptions [32, 33], assessed in terms of the local demographic or socio-economic characteristics of each geographic area of interest. However, segregation does not only regard the physical or spatial distribution of communities around an area, but also the relative level of harmonisation between groups [34, 35]. Keeping in mind that behavioural and cultural adoption is not easily quantifiable, here we develop a framework based on mobile phone data records to assess the extent to which communities differ in their behaviour and cultural habits [36].

3.1 Communication network generation

The framework we propose is based on the analysis of communication patterns between various collectives of people in terms of their communication networks (CN). The CN will be represented, as is usual, as an adjacency matrix [37], O, where its entries o_{ij} correspond to the number of communication events originated at location i and with destination j. In particular, we analyse three different CNs, each one representing the communication patterns between pairs of our two study groups: O^{TT}, O^{RR} and O^{RT}. For each dataset, the first letter of the superscript (T for Turkish and R for refugees) is the originating group and the second is the receiving group. We omit the O^{TR} network due to scarcity of the data.

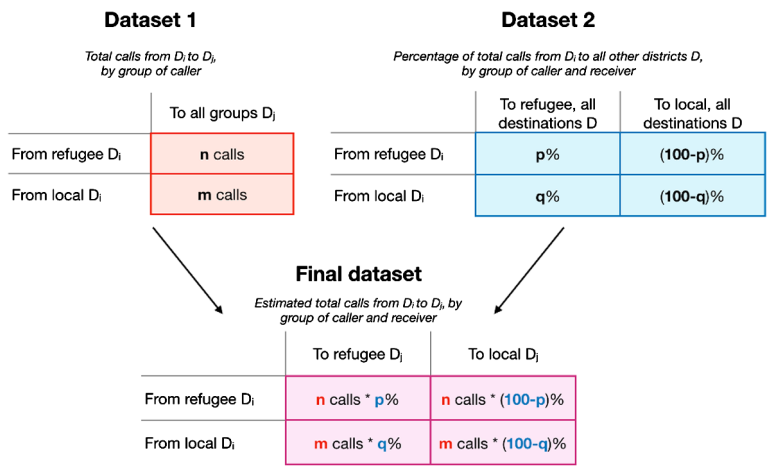

Figure 2 Illustration of the process to merge the information of DS1 and DS2. Top tables indicate the information contained in each individual dataset and bottom table the combined dataset. Colours of entries indicate the dataset from which the information is extracted

To build each CN, we used real communication data provided for the Data 4 Refugees project [38, 39]. The raw data is made up of cellphone calls and SMS (which we will call communications for convenience), and is structured into 3 sub-datasets, DS1, DS2 and DS3, for anonymity purposes. As noted before, the key feature differentiating this dataset from other comparable Call Detail Record datasets [40] is that the users are each assigned a binary tag indicating their status as either refugee or non-refugee (we use the term Turkish to refer to non-refugees). Only DS1 and DS2 were used in this work. The process of constructing the CN is described below, and is illustrated in Fig. 2.

DS1 consists of aggregate communication counts between cell phone antennas on an hourly basis, indicating the total number of calls made by each group (Turkish or refugee) from each antenna i, and directed to each other destination antenna j. However, information about the group receiving the calls was absent. To estimate this, we made use of DS2, which contains information, for each origin antenna, about the destination group but not about the destination antenna. Combining data from DS1 and DS2 we have been able to estimate, for each origin and destination antenna, the total communications made by each group that was directed at each other group over the entire period of study. So, for example, the number of calls refugee-to-refugee from antenna i to j would be the total refugee-originated calls (DS1) multiplied by the proportion of refugee-originated calls from i directed at other refugees (DS2). Communication events originating from and received by the same district are also represented in our Communication Networks as self-loops.

Finally, for convenience, and to reduce data noise, antenna-to-antenna data were aggregated into district-to-district data. We considered districts a better unit of measurement, as they have explicit administrative meaning (as opposed to a Voronoi tesselation of antenna locations, for example). The aggregation of large amounts of antenna data also lessens the risk of uneven geographical distribution of antennas, skewing the interpretation of the data. Turkey is divided into 81 large administrative provinces, which are further subdivided into smaller districts, 923 in total. Our analysis focused on Istanbul, a province of Turkey containing 39 districts.

3.2 Aggregate communication pattern analysis: province scale

Once the CNs have been assembled, we are in position to analyse the communication patterns of both collectives. We start with a macro-analysis of call destination probability in each CN, independent of the individual district. This provides an initial overview of how different the communication patterns are between the two originating groups. A visual analysis of the results, see Fig. 3, suffices to show that both distributions have a similar shape, and it may seem there is not much difference between the communication habits of both collectives on average. However, detailed comparison at the district level evidences a different situation, see Fig. 4. Panel A shows that, while there are districts where differences are small, in many others are they much larger than in the aggregated analysis. Panel B shows the differences in the distributions for three hand-picked districts. The differences are visually evident.

The difference between the results obtained from the aggregated analysis (Fig. 3) and the local analysis (Fig. 4) might be indicative of the Simpson's Paradox [41] in the different CNs. Within the aggregated whole of the province, each district has different proportions of refugee and local populations; additionally, social and economic factors vary by dis-

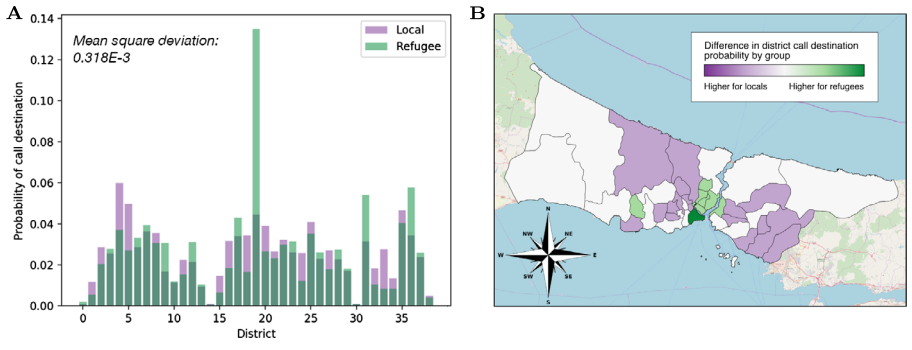

Figure 3 Aggregated analysis of communication patterns of local Turks and refugees. Panel **A** shows the aggregated distribution of call destination (i.e. independent of district). Panel **B** presents the same data on an Istanbul district map

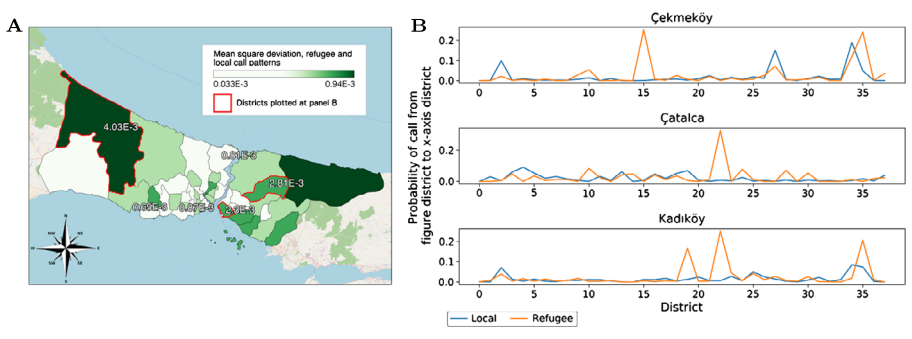

Figure 4 District level analysis of the difference between communication patterns of the Turkish and refugee populations. Panel **A** illustrates the mean squared deviation of the distribution of probabilities of communication (call and SMS) destination originating from each district (e.g., the chances that a call from district i will be directed at district j and not a third district k). That is, for district i this is obtained as $\mathrm{MSD}_i = \frac{1}{39}\sum_j (o_{ij}^{\mathrm{TT}} - o_{ij}^{\mathrm{S}\circ})^2$, where $o_{ij}^{\mathrm{S}\circ} = o_{ij}^{\mathrm{SS}} + o_{ij}^{\mathrm{ST}}, \forall ij$. Panel **B** plots the distributions of the districts where the communication patterns for both communities is largest

trict and population. These considerations indicate the advantage of a local-scale analysis of the CN in the characterisation of the differences between local and refugee communication patterns. In the next section, we address the formal structure of this local-scale Communication Network analysis, which forms the basis of the rest of this work.

3.3 Fine-grained communication pattern analysis: district scale

Given the CNs of the different community pairs, Turkish–Turkish, Refugee–Refugee and Refugee–Turkish, respectively O^{TT}, O^{RR} and O^{RT} we define our behavioural segregation measure in terms of the χ^2 test for homogeneity between the various outgoing communication distributions. Among the many other options we could use to perform this comparison (e.g. cosine similarity, mean square displacement, Pearson correlation, etc), we have chosen the χ^2 since it can work directly on the raw data we have and requires no further assumptions and no pre-processing of the data. Formally, the extent to which the two frequency counts are drawn from the same random variable is measured statistically by the p-value. In our case, the frequency counts correspond to the calls originating from district i and directed to each of the other districts j (represented as vector $\mathbf{o}_i$) for both the Turkish population, $\mathbf{o}_i^{\mathrm{TT}}$, and the refugee population, $\mathbf{o}_i^{\mathrm{RR}}$ (or $\mathbf{o}_i^{\mathrm{RT}}$). Thus,

$$p\text{-value}\left(\mathbf{o}_i^{\mathrm{TT}}, \mathbf{o}_i^{\mathrm{RT}}\right), \tag{1}$$

allows us to assess statistically if both communication patterns are indistinguishable or not. If the results of the test inform us that both samples come from a different distribution (the two call register samples differ significantly), we can conclude that there is segregation in the area in terms of communication. If the test does not allow us to reject the null hypothesis (H_o: both samples come from the same distribution) we cannot conclude segregation exists in that area. Note that, while we are measuring behavioural segregation, we are not trying to measure the level of interaction between the groups. Rather, we solely want to assess to what degree the two groups behave similarly.

In Eq. (1), we are measuring the patterns of outgoing calls from a particular district to all other districts. We have shown that the different districts have different refugee and local population (see Sect. 2). Thus, with a high probability, we are comparing two samples of different sizes. Considering there are more locals than refugees, we expect more calls originating from locals than from refugees. This does not affect our analysis, since the χ^2 test already accounts for these differences in absolute counts. However, when comparing between different destination populations, a difference in spatial distribution between populations can have a significant effect on the shape of outgoing call patterns, since a larger population of one community may mean that they tend to receive more calls than their counterparts, solely because they are more abundant. This problem is magnified if the ratios of local to refugee population are different from district to district. This issue presents itself only in the case of the comparison of TT to RR, since, when comparing TT to RT, the destination populations are identical. Thus, before conducting the χ^2 test, we need to normalise the RR call patterns by the refugee and local populations of the destination districts. In particular, we adjust the outgoing call counts of O^{RR} as $o_{ij}\frac{|T_j|}{|R_j|}$, where $|T_j|$ and $|R_j|$ are the size of local and refugees communities in district j.

The results of conducting the χ^2 test for each district of the Istanbul province shows us that refugee and Turkish calling patterns are always significantly different in every district, and for both comparison datasets (Refugee–Refugee and Refugee–Turkish).

4 Mitigating behavioural segregation through residential mixing policies

Politicians [11, 42], urban planers [13] and scholars [7, 43] have been debating the solutions to segregation and concentration of poverty in Europe and North America since the 70's. One of the primary mechanisms developed, along with some criticisms [43], is residential and social mixing [5, 44]. Policies developed under this approach aim at incentivising the mobility of the segregated communities to other neighbourhoods in order to increase spatial diversity. Rearranging the spatial distribution of each community would be in line with recent research suggesting that diversity within neighbourhoods can actually increase a positive contacts among citizens belonging to different groups [45]. Other than maximising geographic proximity, a parallel approach for increasing the mutual exposure of communities is to make individuals from different groups more similar, relying on the effects of homophily. Homophily is the well-known sociological principle which states that: the more similar individuals are, the more frequent their interactions are expected to be [23, 46].

Our work builds from these fundamental debates and hypotheses, and particularly relies on the principle of homophilic interactions. We assume calling behavioural can be understood as one behavioural feature [46] of individuals. Thus, reducing differences between communities in this regard (i.e. reducing behavioural segregation) may increase exposure, and subsequently, interaction between communities. With this aim in mind, in the following section we estimate the specific volumes of residents that would need to move from their current district, as well as the districts they would need to move to, in order to improve behavioural segregation as measured by variations in CNs.

4.1 Minimising segregation: residential mixing as an optimisation problem

As discussed, house or residential mixing aims at promoting the mobility of segregated communities into other less segregated neighbourhoods. Framing this idea within our definition of behavioural segregation, the problem can be rephrased as obtaining a mobility matrix $\mathbf{M}$, where each entry m_{ji} stands for the fraction of refugees living in district i that are required to be reallocated in district j, in order to maximise[a] the p-value of the χ^2

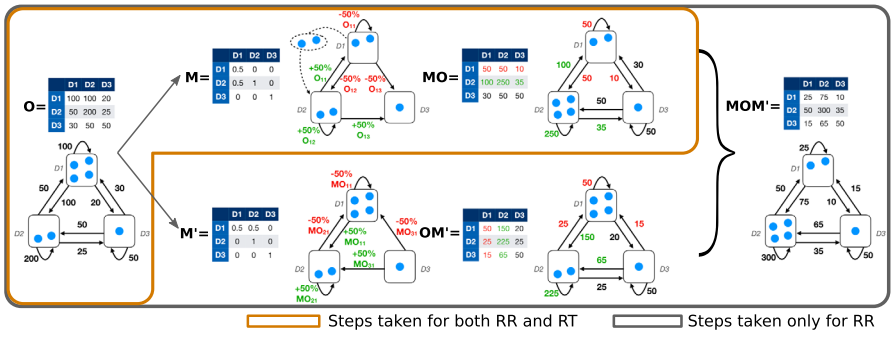

Figure 5 Description of the variables, structures and the process of the non-linear optimisation problem in Eq. (2). The process of minimising segregation considering the datasets TT and RT is exemplified within the orange square, where the effect of the application of the mobility matrix $\mathbf{M}$ is shown. The additional step of multiplying $\mathbf{O}$ by $\mathbf{M}'$, carried out when reducing segregation considering datasets TT and RR, is shown within the gray square. For both processes, we show an example together with the resulting CN and mobility matrices

homogeneity test. Our interpretation of the problem, although applied to call patterns and not spatial distributions, is very similar to the definition of the Dissimilarity Index [28, 29], which is usually interpreted as the percentage of the minority population that would need to relocate in order to perfectly spatially integrate the residential distributions in a region.

The estimation of the best $\mathbf{M}$ can be formally defined as an optimisation problem. We begin with the case of the "eliminating" differences between the RT and TT networks. The non-linear optimisation problem corresponds to

$$\text{maximize} \quad \sum_i p\text{-value}\left(\mathbf{o}_i^{\mathrm{TT}}, \hat{\mathbf{o}}_i^{\mathrm{RT}}\right) \tag{2}$$

$$\text{s.t.} \quad \sum_i m_{ji} = 1 \quad \forall i \tag{3}$$

$$\sum_j \hat{\mathbf{o}}_{ij} \leq f_i \quad \forall i \tag{4}$$

$$0 \leq m_{ji} \leq 1 \tag{5}$$

$$\text{where} \quad \hat{\mathbf{O}}^{\mathrm{RT}} = \mathbf{M}\mathbf{O}^{\mathrm{RT}}, \tag{6}$$

where $\mathbf{O}$ is the matrix of original communication records, $\hat{\mathbf{O}}$ is the resulting matrix of communication records after the mobility matrix has been applied, and each m_{ji} is an unknown to be obtained. The restriction in Eq. (3) guarantees that the total number of communications is maintained. That is, in the mobility matrix, the sum from each origin and to all the destinations must equal the total number of communications observed in the call record matrix $\mathbf{O}$. The restriction in Eq. (4) requires that no district has more than f_i refugees. This restriction is important, as the definition of enclaves has to do with a high fraction of immigrants living in an area with respect to the total immigrant population in the region. In our case f_i is obtained such that the fraction of refugees living in a district never exceeds 10% of the total population. This percentage was chosen as a rounded upper bound based on the empirical observation that, under current conditions, the highest percentage of refugee population in a single district is 8%. The restriction in Eq. (5) simply ensures that the different m_{ji} are bounded in the range $[0, 1]$.

Unlike in the comparison between the TT and RT networks, when comparing TT and RR networks, the destination groups of calls the are different. This requires some modifications to the optimization problem in Eq. (2) when applied to the RR case. First, as we explained in Sect. 3.3, it is necessary to normalise the call destination counts by the different volumes of the target populations of the two datasets when computing the p-value. Second, in order to account for the fact that the refugees being moved are the same ones receiving calls (refugees call refugees in the RR network), we need to apply an additional transformation to change the destination districts of the calls directed at the relocated refugees. This can be done by multiplying the result of $\mathbf{M}\mathbf{O}^{\mathrm{RR}}$ by $\mathbf{M}'$ (the transpose of $\mathbf{M}$). For the definition of the optimisation problem, this means replacing Eq. (6) with $\hat{\mathbf{O}}^{\mathrm{RR}} = \mathbf{M}\mathbf{O}^{\mathrm{RR}}\mathbf{M}'$. Figure 5 provides a simplified example of the optimisation problem we propose.

The high non-linearity of the problem, in both the RT and RR case, does not allow us to obtain satisfactory results optimizing directly the problem in Eq. (2). The fundamental complication is due to the very low p-values obtained with the initial call densities, $\mathbf{o}_i^{\mathrm{TT}}$

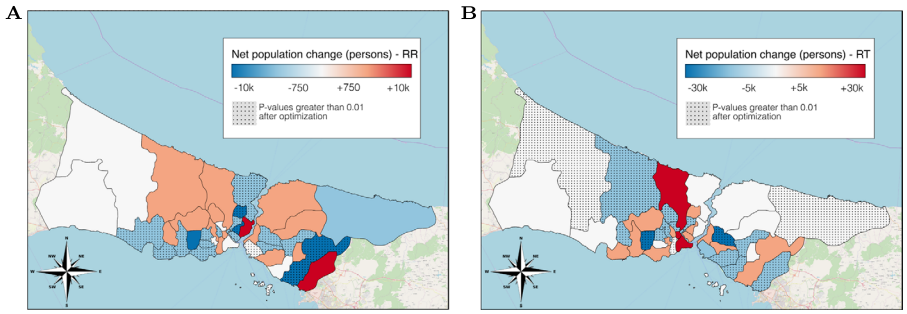

Figure 6 Maps of the results of the optimisation, for the Refugee–Refugee and Refugee–Turkish networks respectively. Districts are coloured in a gradient according to their relative change in population. Districts where Turkish and Refugee communication patterns were harmonised (*p*-value from χ^2 test indicating no significant difference) have a dotted pattern

and $\mathbf{o}_i^{\mathrm{RR}}$. From those values, we were unable to find good initialisations for unknowns m_{ji} that were close enough to a satisfactory mobility matrix solution. Instead, we developed a two-step procedure based on two similar optimisation problems. In the first step, we modified the objective function (with equivalent restrictions) to find the mobility matrix that minimises the mean squared difference between vectors $\mathbf{o}_i^{\mathrm{TT}}$ and $\hat{\mathbf{o}}_i^{\mathrm{RR}}$. In the second step, using as initialisation vector the mobility matrix outcome of the previous optimisation, we minimised the sum of the χ^2 value for the different vectors $\mathbf{o}_i^{\mathrm{TT}}$ and $\hat{\mathbf{o}}_i^{\mathrm{RR}}$. The solution to the optimisation problem was been obtained using the MatLab R2017a engine. We used the *fmincon* function configured to use the Interior-Point algorithm.

This two-step process, similar to the original objective function in Eq. (2), gives very satisfactory results, as Fig. 6 shows. Note again that, under the initial conditions, all of the districts indicated segregation in both the Refugee–Refugee and Refugee–Turkish case. Figure 6 Panel A shows the results mitigating segregation considering the Refugee–Refugee network. We observe that after the proposed mobility, we reduce segregation in 43% of the districts. When considering Refugee–Turkish communications, the results are also impressive (see Fig. 6B). After promoting mobility, segregation is reduced in 40% of the districts.

4.2 Optimising behavioural vs. spatial segregation: the potential trade-offs

In order to establish a baseline for the outcome of our method, we compared our results with a process directed to minimise the Dissimilarity Index (DI). That is, maximize

$$\frac{1}{2}\sum_{i}^{n}\left|\frac{c_i^T}{\sum_j^n c_j^T}-\frac{c_i^R}{\sum_j^n c_j^R}\right|,$$

where n corresponds to the number of districts, and c_j^T and c_j^R are the sum of all outgoing calls made from district j, serving as a proxy of population. As in Sect. 4.1, we preformed a separate optimisation for both the RR and RT networks. In each case, in order to have fair comparison with the results of our method, we impose a constraint to limit the total number of citizens to be relocated under the optimisation, which is set to the number relocated using our behavioural segregation optimisation described above. After optimisation, we compared the results in terms of the change in the DI, and in terms of the

number of districts in which refugee and local call patterns did not exhibit significant differences. Clearly, each optimisation will do its job better than the other (when minimising segregation, we expect a better outcome for segregation than when we minimise call pattern differences), but seeing how distinct the outcomes are can point us towards potential trade-offs.

We note that the original DI calculated using our call volume-based population estimation for the RT network was 24%, while for the RR network it was 32%. Both are quite similar to the value calculated using official population data (around 30%). The optimal mobility matrices found in the optimisation of behavioural segregation increased the DI, to 53% in the RT case and 57% in the RR case. When minimising the DI, we reach 2.5% and 3% for the RT and RR cases respectively. With respect to behavioural segregation: in both cases, RT and RR, when optimising the DI, all of the districts remain significantly segregated (p-value < 0.01). This is in contrast to the optimisation minimising behavioural segregation, which reduced segregation in 40% of the districts. These results suggest that spatial segregation as measured by the Dissimilarity Index and behavioural segregation as we measure it here present somewhat different objective functions with different optima. The optimisation of both measures may be taken as being desirable, and studying their mutual effects on one another could be useful. An interesting prospect for future work could go in the direction of designing a multi-objective objective function, in order to find points in the problem space where a positive outcome exists for improving both spatial and behavioural segregation.

4.3 Economic incentives towards integration

From one perspective, social integration can be framed in terms of cost-benefit analysis [47]. In this conceptualisation, language acquisition, distance from family, and exposure to unfamiliar cultures can be considered costs, though they are difficult to quantify in economical terms. Housing costs, in contrast, are relatively easy to quantify. Aside from the characteristics of individual houses, this cost reflects a variety of factors including access to services, employment, and city resources [48–50]. As previously mentioned, rent prices are negatively related to refugee population, as Fig. 1C show. This implies that some rent-reduction incentives might be effective in getting refugees to relocate out of enclaves. This could be an opportunity for public and private actors interested in increasing host-refugee integration in Turkey to adjust the cost-benefit analysis of refugee location choice by subsidising rent in targeted areas of the city, thereby encouraging refugees to live away from enclaves and making inter-group contact more frequent.

In support of the viability of using rental subsidies as a way to incentivise refugee location choice, we examined the overall change in rent payments that would occur under the new population distribution considering rental markets for the 2017 period [27]. The proposed optimisation problem in Eqs. (2)–(6) provides us with information about the volume of communications that need to be shifted from one district to another. The density of communications originating from an area is known to be related to the population density of the area [51–53] as Fig. 7, drawn from the real population and CDR data, confirms. We can thus use outgoing call volume as a proxy for the amount of citizens for whom we need to incentivise mobility. Performing the optimisation considering the RR communication, a total of 54,942 refugees are required to be relocated (12% of the refugee population). The resulting net increase in monthly rent cost is 11,709,295₺ (1,847,817€), which corresponds to 213₺ (34€) per person/month. Performed considering the RT communication

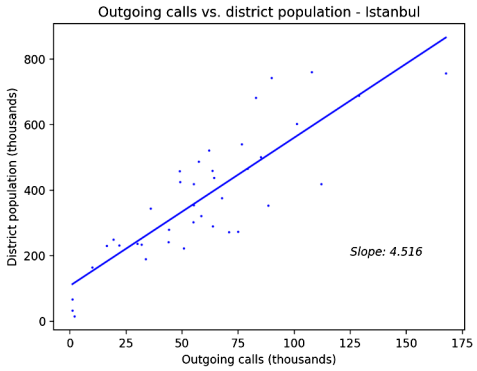

Figure 7 Observed relationship between the amount of communications originated in each district of Istanbul and the population living in the district. We confirm a linear relationship as expected. Using this relationship to make an estimate indicates that there are about 565,000 refugees living in Istanbul. This is above the 400,000 number cited by public authorities. On the other hand, using the relationship to estimate the local population of the city returns a figure of 10,000,000 people, below the official count of 13,500,000

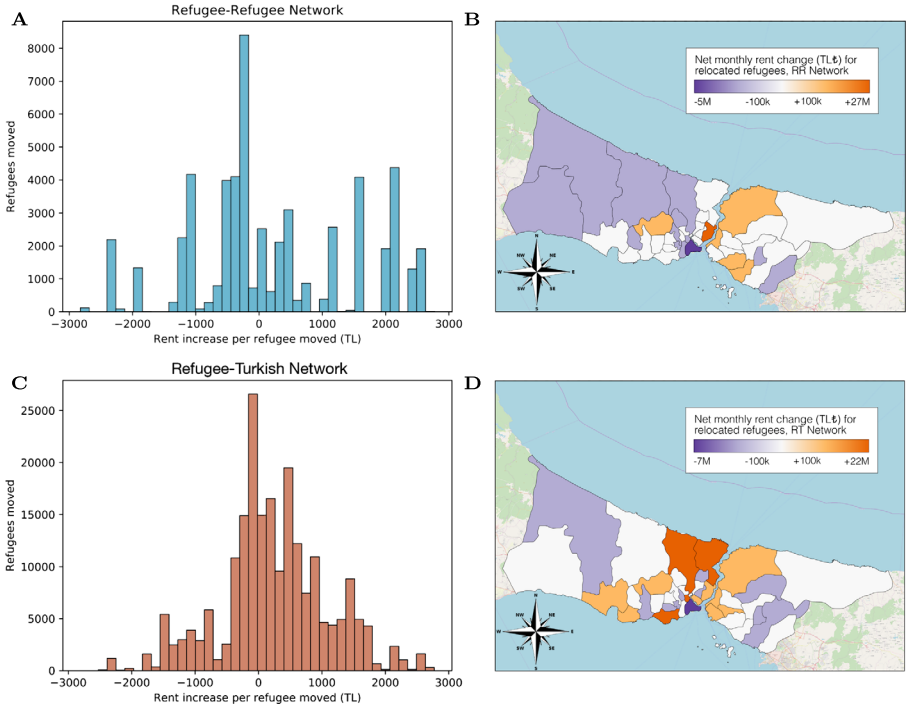

Figure 8 The histograms at left illustrate the distribution of monthly rent changes after optimisation, for the Refugee–Refugee and Refugee–Turkish networks respectively. The height of the vertical bars indicates the number of relocated refugees whose rent payments increased or decreased by the value indicated on the horizontal axis. Subsequently, the maps at right indicate, for each respective network, the total monthly rent change per district. That is, the product, for each origin district j and destination district i, of the difference in rent cost between j and i and the number of refugees moved m to the destination district. Thus, formally, the reported rent change per district i is obtained as $v_i = \sum_{\forall j} m_{ij}(c_i - c_j)$

network, the optimisation resulted in a relocation of 212,100 refugees (approx. 40% of the population). This corresponds to a net rent increase of 52,430,540₺ (8,273,946€), or 247₺ (39€) per person per month.

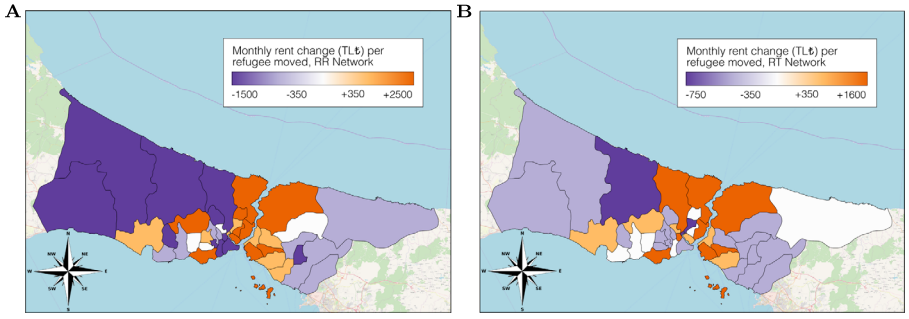

Figure 9 Maps, for each corresponding network, of the average monthly rent change per refugee arriving in each district. These maps complement the distributions of Fig. 8

As it can be seen in Fig. 8A and C, the changes in rent payment approximate a normal distribution with a large variance, meaning that, under the adjusted population distribution, some refugees would considerably increase their savings on rent, and others would pay a higher price. The overall tendency, though, is a positive increase in the rent costs. The distribution of these changes in rent cost over the districts of Istanbul at the level of the individual is provided in Fig. 9. These figures provide an individual (refugee) point of view in terms of the increase or reduction in cost of living. Panels B and D of Fig. 8, on the other hand, provide a governmental or organisational perspective. The maps indicate the total investment that would be required in each district in order to fully offset the increased rent payments of refugees. As we can see, the subsidies would be larger at the districts near the Bosphorus Strait. Surprisingly, these largest subsides are not regularly distributed among adjacent districts, and they correspond to the densest areas of the province.

5 Discussion

This work essentially makes two contributions. On the one hand, we perform a large-scale data analysis of behavioural segregation in Istanbul on the basis of call patterns. On the other, we provide a framework for reducing the level of segregation based on the normative assumption that lowering behavioural segregation can increase social integration.

The method we propose allows for the quantification and potential mitigation of refugee segregation within a geographical area. The method goes beyond the spatial dimension typically considered, and accounts for behavioural aspects of the different communities. From the combined analysis of communication data, the first step is to establish if and to what extent the two group of interest behave differently. Our analysis confirms that differences in communication patterns were always significantly different (p-values lower than 0.01) comparing the two groups. The two plausible reasons accounting for these differences are the existence of strong cultural differences and residential enclaves, and the combination of both factors are reflected in the segregation of refugees in specific areas. According to the classical assumptions of public policy, indistinguishable communication patterns between Turkish population and Syrian refugees would reflect a situation of integration; that is, if their patterns of communication reached a kind of natural equilibrium, considering the geo- and socio-economical situation of the city. From this approach we hypothesise that, by merging the differing communication patterns of the Turkish and

Syrian refugee populations into a single one, we can increase the potential for more intercommunication and integration between them, following the principle of homophily in social interactions.

Nevertheless, while the reported results presented here have a number of potentialities, there are also a number of limitations that should be further assessed in future studies, most of them with an interdisciplinary approach in mind. First, our model works within an idealised situation that does not address some well-known and important factors for integration, such as Syrian-Turkish cultural differences, which should be taken into account for a well-designed public policy aimed at improving social cohesion. These elements are important not only when a model such as the one presented here might be used as an input for public planning, but also for further research in the academic field. An example of this are the long-term effects that achieving similar communication patterns might have on the cultural aspects at the individual or inter-group level. On the other hand, local dynamics of urban politics should be taken into account. In our case, this would include for instance the complex relationship between the Turkish state, local government, real estate businesses, and residents in the context of the trend of "urban transformation" [54–56]. Second, concerning the methods used here, further research could also include other variables that we have not addressed here. While our analysis is aggregated and anonymised, a similar procedure could be carried out with data tracking individuals over a period of time, to draw related but distinct conclusions. Additionally, "quality" of communications could be taken into consideration. Here, SMS and phone calls are given the same value. Perhaps even call duration could provide some measure of communication quality. All in all, these elements would allow to better address the interaction between individual, group and contextual factors that determine spatial patterns of segregation. Finally, the comparison we performed between our behavioural optimisation problem and an optimisation problem minimising the Dissimilarity Index opens up an interesting possibility of attempting a multi-objective optimisation, in the attempt to find an outcome beneficial to both spatial and behavioural segregation. This could be especially relevant given the fact that both optimisations were ineffective at reducing the segregation measure lowered by the other.

While we admit that these unaddressed aspects regarding other socio-cultural factors or different levels of detail of the Communication Network [57] should be considered in subsequent works, the estimations given here can be of practical use in several ways. First, the developed procedure provides estimations as to the level of integration that can be achieved by using social and residential mixing strategies. Secondly, we provide a systematic method that can give consistent quantitative evidence about the volumes and destinations required if a group—in our case Syrian refugees—were to be relocated in a particular urban or regional area. This recommendation can be seen as a good starting point for governments and NGOs to analyse the situation, target their campaigns, and optimise their economic investment in the area. Lastly, the optimisation framework proposed here can be easily complemented with other interesting parameters. In this work, we applied only one restriction to mobility: the one limiting the proportion of refugees per district. That said, mobility can be easily restricted in other ways. For instance, assuming the availability of the data, a restriction could be applied using employment data or labor demand in each district in order to achieve more socially accurate results.

An illustrating example of this exercise is if we consider how refugees' choice of residential location influences their integration with the local community [58, 59]. Policy-driven incentives such as rent subsidies could facilitate those who, for example, might choose to move away from an ethnic enclave if rent prices outside were lower [56]. In our case, we have estimated that average rent paid among the relocating refugee population would not rise by more than by 39€ per family per month. This is a barrier that could be too high for refugees who already have difficulties. However, it is also a barrier that governmental and NGO policy could reduce. Governments and NGOs have a range of options available to them to incentivise locational choice which are out of the scope of the scientific work presented here. However, several well-known approaches to the problem exist. In this sense, our method could be used as an input to design programs involving a differential rent subsidy, or "voucher" [60], based on the relative rent price in target districts.

As a final note, it is worth saying that it is not our objective to advocate particular policies, but instead to provide methods to quantify and give indications of what could be expected from house mixing policies. An optimal integration of the refugee and host population should probably be considered an organic process, as the meaning of integration here is connections between people, and connections are made voluntarily and maintained only by individual choice. In the event that governments and non-governmental entities decide to take a hands-off approach to integration policy, the proposed framework can be useful for analysing how the situation evolves and providing early warnings of recessive or problematic conditions.

Acknowledgements
We acknowledge Türk Telekom for providing the communication data necessary for this work in the frame of the Data 4 Refugees Challenge, as well as Boğaziçi University and Tübitak for participating in the organisation of the challenge.

Funding
D.R., I.S., A.S.-R. and J.B.-H. acknowledge the support of the Spanish MICINN project PGC2018-096999-A-I00.

Abbreviations
CN, Communication Networks; CDR, Call Detail Records; ICT, Information and Communications Technology; NGO, Non-Governmental Organization; SMS, Short Message Service.

Availability of data and materials
The source code and the aggregated data to reproduce the experiments within the article can be downloaded from http://cosin3.rdi.uoc.edu/data.html.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the paper. All authors read and approved the final manuscript.

Endnote
[a] Recall that we reject the H_o when the p-value is less than our significance level, so maximising the p-value has the effect of forbidding the rejection of the H_o. That is, we cannot say there is segregation in the area.

Publisher's Note

Received: 5 April 2019 Accepted: 25 February 2020 Published online: 06 March 2020

References
1. Phillips D (2013) Minority ethnic segregation, integration and citizenship: a European perspective. In: Linking integration and residential segregation. Routledge, London, pp 49–66
2. Sampson RJ (2012) Great American city: Chicago and the enduring neighborhood effect. University of Chicago Press, Chicago
3. Slater T (2011) From 'criminality' to marginality: rioting against a broken state. Hum Geogr 4(3):106–115

4. Massey DS (2012) Reflections on the dimensions of segregation. Soc Forces 91(1):39–43
5. Atkinson R, Kintrea K (2001) Disentangling area effects: evidence from deprived and non-deprived neighbourhoods. Urban Stud 38(12):2277–2298
6. Jargowsky PA (1996) Take the money and run: economic segregation in us metropolitan areas. Am Sociol Rev 61:984–998
7. Duncan OD, Duncan B (1955) Residential distribution and occupational stratification. Am J Sociol 60(5):493–503
8. Charles M, Grusky DB (2005) Occupational ghettos: the worldwide segregation of women and men, vol 71. LIT Verlag, Münster
9. Blair SL, Lichter DT (1991) Measuring the division of household labor: gender segregation of housework among American couples. J Fam Issues 12(1):91–113
10. Maccoby EE, Jacklin CN (1987) Gender segregation in childhood. In: Advances in child development and behavior, vol 20. Elsevier, Amsterdam, pp 239–287
11. Smith A (2001) Religious segregation and the emergence of integrated schools in northern Ireland. Oxf Rev Educ 27(4):559–575
12. Hu J, Zhang Q-M, Zhou T (2019) Segregation in religion networks. EPJ Data Sci 8(1):6
13. Musterd S (2005) Social and ethnic segregation in Europe: levels, causes, and effects. J Urban Aff 27(3):331–348
14. Malmberg B, Andersson E, Östh J (2013) Segregation and urban unrest in Sweden. Urban Geogr 34(7):1031–1046
15. Phillips D, Harrison M (2010) Constructing an integrated society: historical lessons for tackling black and minority ethnic housing segregation in Britain. Hous Stud 25(2):221–235
16. Cheong PH, Edwards R, Goulbourne H, Solomos J (2007) Immigration, social cohesion and social capital: a critical review. Crit Soc Policy 27(1):24–49
17. Valente P (2010) Census taking in Europe: how are populations counted in 2010? Popul Soc 467:1
18. Shearmur R (2015) Dazzled by data: big data, the census and urban geography. Taylor & Francis, London
19. Hilbert M (2016) Big data for development: a review of promises and challenges. Dev Policy Rev 34(1):135–174
20. Lokanathan S, Kreindler GE, de Silva NN, Miyauchi Y, Dhananjaya D, Samarajiva R (2016) The potential of mobile network big data as a tool in Colombo's transportation and urban planning. Inf Technol Int Dev 12(2):63
21. Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. Science 353(6301):790–794
22. Pinkovskiy M, Sala-i-Martin X (2016) Lights, camera… income! Illuminating the national accounts-household surveys debate. Q J Econ 131(2):579–631
23. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27(1):415–444
24. Wong DW, Shaw S-L (2011) Measuring segregation: an activity space approach. J Geogr Syst 13(2):127–145
25. Blumenstock J, Fratamico L (2013) Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. In: Proceedings of the 4th annual symposium on computing for development. ACM, New York, p 11
26. Kirişci K, Brandt J, Erdoğan MM (2018) Syrian refugees in Turkey: beyond the numbers. Brookings. https://www.brookings.edu/blog/order-from-chaos/2018/06/19/syrian-refugees-in-turkey-beyond-the-numbers/
27. ISTANBUL Rental Apartments Price Statistics. Endeksa. https://www.endeksa.com/analiz/istanbul/endeks/kiralik/daire
28. Massey DS, Denton NA (1988) The dimensions of residential segregation. Soc Forces 67(2):281–315
29. Duncan OD, Duncan B (1955) A methodological analysis of segregation indexes. Am Sociol Rev 20(2):210–217
30. Louf R, Barthelemy M (2016) Patterns of residential segregation. PLoS ONE 11(6):0157476
31. Frey WH, Myers D (2005) Racial segregation in us metropolitan areas and cities, 1990–2000: patterns, trends, and explanations. Population studies center research report (05-573)
32. Farber S, O'Kelly M, Miller HJ, Neutens T (2015) Measuring segregation using patterns of daily travel behavior: a social interaction based model of exposure. J Transp Geogr 49:26–38
33. Wang Q, Phillips NE, Small ML, Sampson RJ (2018) Urban mobility and neighborhood isolation in America's 50 largest cities. Proc Natl Acad Sci USA 115:7735–7740
34. Algan Y, Bisin A, Manning A, Verdier T (2013) Cultural integration of immigrants in Europe. Oxford University Press, London
35. Juzwiak T, McGregor E, Siegel M (2014) Migrant and refugee integration in global cities: the role of cities and businesses. The Hague Process on Refugees and Migration and UNU-MERIT & Its School of Governance
36. Canton N (2012) Cell phone culture: how cultural differences affect mobile use. CNN Online 28
37. Newman M (2010) Networks: an introduction. Oxford University Press, London
38. Salah AA, Pentland A, Lepri B, Letouzé E, Vinck P, de Montjoye Y-A, Dong X, Dağdelen Ö (2018) Data for refugees: the D4R challenge on mobility of Syrian refugees in Turkey. arXiv preprint. arXiv:1807.00523
39. Türk Telekom (2018) Data for refugees Turkey: D4R. d4r.turktelekom.com.tr/
40. de Montjoye Y-A, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD (2014) D4D-Senegal: the second mobile phone data for development challenge. arXiv preprint. arXiv:1407.4885
41. Wagner CH (1982) Simpson's paradox in real life. Am Stat 36(1):46–48
42. Andersson R, Bråmå Å (2004) Selective migration in Swedish distressed neighbourhoods: can area-based urban policies counteract segregation processes? Hous Stud 19(4):517–539
43. Lees L (2008) Gentrification and social mixing: towards an inclusive urban renaissance? Urban Stud 45(12):2449–2470
44. Musterd S, Andersson R (2005) Housing mix, social mix, and social opportunities. Urban Aff Rev 40(6):761–790
45. Kaufmann E, Harris G (2015) "White flight" or positive contact? Local diversity and attitudes to immigration in Britain. Comp Polit Stud 48(12):1563–1590
46. Axelrod R (1997) The dissemination of culture: a model with local convergence and global polarization. J Confl Resolut 41(2):203–226
47. Danzer AM (2011) Economic benefits of facilitating the integration of immigrants. CESifo DICE Report 9(4):14–19
48. Selim H (2009) Determinants of house prices in Turkey: hedonic regression versus artificial neural network. Expert Syst Appl 36(2):2843–2852
49. Selim S (2011) Determinants of house prices in Turkey: a hedonic regression model. Doğuş Üniv Derg 9(1):65–76

50. Limsombunchai V (2004) House price prediction: hedonic price model vs. artificial neural network. In: New Zealand agricultural and resource economics society conference, pp 25–26
51. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888–15893
52. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. EPJ Data Sci 4(1):4
53. Khodabandelou G, Gauthier V, El-Yacoubi M, Fiore M (2016) Population estimation from mobile network traffic metadata. In: 2016 IEEE 17th international symposium on a world of wireless, mobile and multimedia networks (WoWMoM). IEEE Press, New York, pp 1–9
54. Torus B, Yönet NA Urban transformations in Istanbul
55. Gündoğan AZ (2019) Divergent responses to urban transformation projects in Turkey: common sense and state affinity in community mobilization. Urban Geogr 40(7):893–917
56. Kahraman C, Alkan G (2018) Istanbul's third airport in terms of transportation geography: geopolitics, regional and economic effects. PEOPLE Int J Soc Sci 3(3)
57. Carpi LC, Schieber TA, Pardalos PM, Marfany G, Masoller C, Díaz-Guilera A, Ravetti MG (2019) Assessing diversity in multiplex networks. Sci Rep 9(1):4511
58. Danzer AM, Yaman F (2013) Do ethnic enclaves impede immigrants' integration? Evidence from a quasi-experimental social-interaction approach. Rev Int Econ 21(2):311–325
59. Bauer T, Epstein GS, Gang IN (2005) Enclaves, language, and the location choice of migrants. J Popul Econ 18(4):649–662
60. Varady DP, Walker CC (2003) Housing vouchers and residential mobility. J Plan Lit 18(1):17–30