Migrant by Proxy: How Big Data Defines and Redefines Human Mobility[1]

Tuba Bircan[1], Alina Sîrbu[2], Haodong Qi[3], Carlos Arcila Calderón[4], Albert Ali Salah[5]

[1]Vrije Universiteit Brussel
[2]University of Bologna
[3]Malmö University
[4]University of Salamanca
[5]Utrecht University

Corresponding Author: tuba.bircan@vub.be

I. Opening Essay: When Proxies become Migrant (Tuba Bircan)

Migrants are increasingly visible in data, yet rarely as themselves. They appear instead as traces: a geotagged social media post, a mobile phone subscription crossing network boundaries, a shift in night-time satellite luminosity, or a surge in online searches for migration-related terms. These traces do not record migration directly. They function as proxies: observable signals used to infer a social phenomenon that is not itself directly captured in the data. In doing so, they make migration measurable, scalable, and, increasingly, actionable.

Migration research has, of course, always relied on indirect measurement. Conventional indicators such as nationality, country of birth, or place of previous residence are themselves proxies, each capturing only selected dimensions of movement, status, or belonging (Willekens, 2019; Yildiz et al., 2025). The challenge of measuring migration has never been one of perfect observability. What distinguishes contemporary big-data environments is not the existence of proxies as such, but their scale, velocity, opacity, and their growing entanglement with privately governed digital infrastructures (Kitchin, 2022; Bircan, 2024). The rise of computational migration research therefore marks more than a methodological expansion. It raises a deeper question about how migration itself becomes knowable.

This paper introduces the concept of the migrant by proxy to capture that transformation. Our central argument is not that proxy-based measurement is inherently flawed, nor that indirect indicators should be abandoned. Rather, we argue that proxies do not merely approximate migration as a pre-existing social reality. They participate in defining what migration is understood to be. Migration, in this sense, is not simply observed through digital infrastructures; it is partially assembled through them.

This distinction matters because migration is not equivalent to mobility. Human movement is observable in many forms: commuting, tourism, seasonal circulation, forced displacement, evacuation, temporary relocation, and permanent settlement. Migration, by contrast, is a socially, legally, and institutionally constructed category,

[1] This is the uncorrected author proof. Please cite as: "Bircan, T., A. Sîrbu, H. Qi, C. Arcila Calderón, A.A. Salah, "Migrant by Proxy: How Big Data Defines and Redefines Human Mobility," *Journal of Immigration and Refugee Studies*, 1-14, 2026. https://doi.org/10.1080/15562948.2026.2689658

typically involving thresholds of duration, residence, legal status, or territorial boundary crossing. Yet many digital traces capture movement without capturing the classificatory dimensions that distinguish migration from other forms of mobility. A device changing location across regions does not reveal whether its user has migrated, travelled, fled, commuted, or simply lent a phone to someone else.

At the same time, digital traces do not merely collapse migration into crude approximations of movement. In some contexts, they do the opposite. High-frequency behavioural data can fragment migration into granular sequences of micro-mobility that no longer map cleanly onto demographic or legal definitions. The same infrastructure that simplifies migration into movement can also decompose it into activity patterns too fine-grained to align with conventional categories. Computational systems therefore do not simply distort migration in one direction. They reshape its boundaries in ways that depend on how traces are collected, modelled, and interpreted.

To clarify this transformation, it is useful to distinguish three dimensions through which proxies operate.

First, proxies function as technical indicators. They translate behavioural, infrastructural, or discursive traces into measurable signals. A “lost subscriber” in mobile phone data, a geolocated post suggesting cross-border movement, or a spike in migration-related search queries all exemplify technical attempts to infer migration-related phenomena from indirect observations.

Second, proxies operate as epistemic operators. They shape what becomes thinkable and measurable as migration by privileging certain forms of evidence over others. Some mobilities become highly visible because they leave dense digital traces, while others remain absent because they occur outside the infrastructures from which data are extracted. The undocumented migrant without stable connectivity, the elderly person with limited digital participation, or the refugee whose movements are constrained by infrastructural dependency do not merely appear less clearly in these systems; they may disappear from the analytic field altogether.

Third, proxies act as governance devices. Once stabilised as indicators, they can travel beyond research settings into dashboards, forecasting systems, humanitarian monitoring, and migration governance infrastructures. In these contexts, proxies do not simply describe migration; they may shape decisions about intervention, preparedness, risk assessment, or resource allocation (Salah, Korkmaz, & Bircan, 2022; Bircan, 2024; Fontana et al., 2025).

These three dimensions do not necessarily align. A proxy may perform adequately as a technical signal while remaining conceptually ambiguous as a representation of migration. It may be analytically useful while still introducing systematic distortions into governance contexts. This is consistent with a broader body of work in science and technology studies showing that classifications, indicators, and measurement systems are not passive reflections of reality but active instruments in its organisation (Espeland & Stevens, 2008; Agre, 1997; Sismondo, 2011). In migration research, proxy infrastructures do not simply observe categories; they participate in stabilising them.

A further complication emerges when proxies are used not only descriptively but predictively. The attraction of digital traces lies partly in their timeliness. Unlike censuses, surveys, or administrative registers, digital infrastructures can generate near real-time signals, making them appealing for early warning systems, displacement

monitoring, and migration forecasting. Yet computational social science has repeatedly shown that descriptive and predictive models operate under different epistemic assumptions and should not be conflated (Hofman et al., 2021). A signal that correlates with past mobility does not automatically constitute a valid predictor of future migration.

This distinction is especially consequential in migration research, where uncertainty is structurally high and mobility decisions are shaped by rapidly changing political, economic, and environmental conditions (Bijak, 2019). Search queries may reflect aspiration, anxiety, or curiosity rather than concrete migration plans. Social media activity may capture discourse rather than movement. Proxies developed to observe one phenomenon are often repurposed to anticipate another. In that transition, methodological convenience can become epistemic overreach.

A central implication of proxy-based migration measurement is the structural displacement of the migrant as a speaking subject. What proxy systems capture are externally observable traces: movement patterns, connectivity, discourse, behavioural signals, inferred affiliations. What they cannot access is the migrant's own account of movement, status, motive, identity, or lived experience. This is not merely an ethical concern about dehumanisation, though it may be that as well. It is an epistemological boundary. Certain dimensions of migration are structurally inaccessible to inferential data systems, regardless of scale or computational sophistication.

This is precisely why qualitative and participatory approaches remain indispensable, not as a normative corrective to computational methods, but as sources of knowledge about dimensions of migration that proxy infrastructures cannot render visible. The issue is not merely methodological pluralism as a matter of principle. It is epistemic adequacy.

The governance implications are equally significant. Many of the most valuable data sources for computational migration research are controlled by private actors. Access is often negotiated, restricted, or contingent on proprietary infrastructures, creating asymmetries in who can produce migration knowledge and under what conditions. At the same time, the populations whose traces generate these indicators typically have little awareness of how they are being represented and limited capacity to contest resulting classifications. In such settings, the central question is not simply whether data were collected with individual consent. It is whether systems of proxy-based representation are legitimate, contestable, and accountable, particularly where the consequences affect identifiable migrant populations.

Against this background, this reflection paper asks three interconnected questions. First, how do different proxy-based measures redefine the boundary between mobility and migration? Second, what epistemic assumptions are embedded in the proxies used to make migrants visible through digital infrastructures? Third, what standards of accountability are required when proxy-based indicators move from research into governance contexts?

The contributions that follow approach these questions from complementary perspectives. They examine how migrants are constructed through discursive proxies in online hostility, behavioural proxies in social media traces, infrastructural proxies in mobile telecommunications data, and demographic efforts to align digital traces with established population concepts. Together, they do not reject proxy-based migration research. Rather, they ask what becomes visible, what remains absent, and what

becomes actionable when migration is increasingly known through proxies rather than persons.

II. Invited Reflections:

Migrants in the Mirror of Discourse: Hate Speech, Sentiment, and Digital Proxies of Otherness (Carlos Arcila Calderón)

Big data migration research is often associated with measuring movement: estimating flows, detecting displacement, or inferring settlement patterns from digital traces. Yet migrants are also constructed in data systems in a fundamentally different way, not through mobility signals, but through discourse. In digital communication environments, migrants frequently appear not as demographic actors or mobile populations, but as symbolic figures produced through language: objects of hostility, securitisation, moral panic, or political contestation (Chouliaraki & Georgiou, 2022; Ekman, 2019). In this sense, online discourse itself functions as a proxy infrastructure, generating classifications of "the migrant" that may bear little relation to actual migration trajectories, yet have substantial consequences for public perception and governance.

Computational methods have made these discursive formations measurable at unprecedented scale. Techniques such as sentiment analysis, hate speech detection, topic modelling, and network analysis enable researchers to trace anti-immigrant narratives across large digital corpora, revealing temporal dynamics, amplification mechanisms, and transnational circulation patterns that would be difficult to capture through conventional qualitative approaches alone (Arcila Calderón et al., 2021, 2022; Fortuna & Nunes, 2018). Such methods have produced valuable empirical insights. Research has shown, for example, that anti-refugee sentiment in digital environments can translate into measurable offline harms. Müller and Schwarz (2021), analysing German municipalities, found that anti-refugee Facebook activity was associated with increased violent attacks against refugees, with the relationship weakening during periods of internet outage. Findings such as these illustrate that digital discourse is not merely expressive, but socially consequential.

Yet computational tractability should not be mistaken for conceptual clarity. The key question is not simply whether anti-immigrant discourse can be detected at scale, but what exactly is being measured when computational systems classify discourse about migrants.

Anti-immigrant rhetoric is rarely reducible to overtly hostile speech. It is often historically layered, politically coded, and rhetorically indirect. Exclusionary narratives may emerge through irony, euphemism, securitisation language, or invocations of demographic and economic threat that do not fit neatly within conventional computational categories of hate or negative sentiment, while nevertheless contributing to exclusion and stigmatisation (Ekman, 2019; Fortuna & Nunes, 2018; Blodgett et al., 2020). Sentiment analysis, in particular, assumes that affective polarity can meaningfully capture social harm embedded in discourse, an assumption that becomes increasingly fragile in politically contested communicative environments.

This is not merely a technical limitation. It is an epistemic one. Computational models require operational categories, but those categories are themselves interpretive constructs shaped by annotation practices, institutional norms, cultural assumptions,

and modelling choices (Blodgett et al., 2020). A classifier may therefore appear to identify anti-immigrant hostility while in practice reproducing the conceptual boundaries embedded in its training data. For example, a model trained primarily on explicitly racist slurs may successfully detect overt hate speech while failing to recognise exclusionary narratives framed through concerns about "security", "cultural incompatibility", or "demographic threat". The result is not simply a technical error, but a particular definition of hostility embedded within the classification system itself. In migration-related discourse, this matters because exclusionary narratives often operate through politically normalised framings that are not explicitly hateful in formal terms but nevertheless shape how migrants are socially understood.

From the perspective of this paper, the deeper issue is that discursive proxies do not simply measure attitudes toward migrants. They participate in constructing the migrant as a particular kind of social object. The migrant who emerges in these datasets is often not a person with a migration trajectory, legal status, or self-defined identity, but a figure assembled through the language of others. A refugee becomes a security concern. A foreign-born person becomes an economic burden. A migrant becomes a symbolic proxy for broader anxieties about nationhood, social change, or political control. Computational systems do not create these narratives, but by translating them into measurable indicators, they may stabilise and legitimise them as analytically meaningful categories.

This distinguishes discursive proxies from the movement-oriented proxies examined elsewhere in this paper. Mobile phone records geolocated behavioural traces, or search-based indicators attempt, however imperfectly, to infer migration-related activity from observable actions. Discursive proxies may classify migrants without observing migration at all. A group can become hyper-visible in hostile digital discourse without appearing in mobility datasets, while others who have migrated extensively may remain discursively absent. Visibility here is not about movement, but symbolic representation.

The governance implications are significant. Computational classifications of migration-related discourse are not merely descriptive tools. They increasingly shape content moderation systems, public narratives, policy debates, and broader imaginaries of migration governance (Gillespie, 2018). In such contexts, classification errors are not simply methodological inaccuracies. They can reinforce distorted representations of already vulnerable populations, amplify particular constructions of social threat, or narrow the range of discourse considered legitimate. The operational logic of such systems is often only partially transparent, making critical scrutiny all the more important.

At the same time, rejecting computational approaches would be neither analytically defensible nor practically realistic. Contemporary anti-immigrant discourse unfolds at scales that demand computational methods. The challenge is therefore not whether such tools should be used, but how their outputs should be interpreted. Scalable computational analysis can reveal important patterns, but it cannot substitute for contextual interpretation, qualitative inquiry, or critical reflection on how categories are constructed and operationalised.

Within the migrant by proxy framework developed in this paper, discursive proxies operate across all three dimensions identified in the opening essay. As technical indicators, they reduce complex rhetorical environments into measurable signals. As

epistemic operators, they shape what kinds of migrant subjects become visible through computational analysis. As governance devices, they influence moderation infrastructures, public narratives, and policy imaginaries that may affect migrants directly. The migrant constructed in these systems is not observed as a moving subject, but assembled through digital discourse, spoken about rather than speaking.

Migrants in Digital Traces: Identifying Mobility through Networks, Language, and Location (Alina Sîrbu)

Social media data are among the most visible big data sources in computational migration research. Unlike the discursive proxies discussed in the previous section, these data are often used to infer migration-related behaviour more directly, including migration stocks, flows, and aspects of integration. Yet here too, migrants do not appear directly. They are inferred through proxies: changes in geolocation, patterns of language use across contexts, network ties, or inferred affiliations. The apparent behavioural immediacy of these signals should not obscure the fact that they remain inferential constructs.

We can distinguish between two classes of methods. The first concentrates on geolocation changes for an anonymous user, considering the geolocation tags of posts for that user as a proxy for their location (Zagheni et al., 2014; Fiorio et al., 2017). When the location changes from one country to another, we assume the user crossed the border. The location can be estimated by actual geo-location of posts, but also through other proxies, such as employer location, affiliation, declared profile location, etc. By considering a large number of users, we can estimate migration flows. A second approach attempts to assign residence and nationality to anonymous users (Kim et al., 2020, Mazzoli et al., 2020). Residence is assigned in a similar manner to the first class of methods, by considering the location of posts. Nationality instead is estimated through a combination of the languages used by the user, their location, and the location of their friends, with the latter being the most informative factor. Besides being able to estimate both flows and stocks, this approach also allows to study integration and other similar processes, by analysing the content produced by the migrant users, and their interactions with other (local) content (Kim et al., 2022, 2023). Importantly, these methods aim to infer behavioural mobility more directly than proxies based on discourse or intention, although the distinction between observable movement and migration as a socially or legally defined category remains analytically important.

These approaches have been applied across multiple digital platforms, with implementation depending on data access and platform architecture. For Facebook, the Marketing API has been used to estimate stocks without downloading micro-level data (Zagheni et al., 2017; Palotti et al., 2020). Twitter (now X), instead, has been traditionally one of the most popular data sources (Pfeffer et al., 2023), due to free APIs accessible to academic researchers, allowing to obtain large datasets of user-level posts and profiles. These APIs are currently no longer available, so Twitter/X social media research has decreased (Murtfeldt et al., 2024). Linked-in data has been used to estimate brain drain (State et al., 2014). Scientific publication data, although not social media, has been used to estimate migration of scientists (Sanliturk et al., 2023; Pollacci et al., 2025). In this case, affiliation is used as the location proxy, while the first place of study can be used as nationality proxy.

Both types of methods make assumptions that can be wrong, leading to pointwise estimations that are noisy, for reasons that depend on many factors. For instance, the accuracy of the moment of border crossing depends on the posting frequency of the user. For a user who posts daily, we can accurately identify the migration event in time, while for users who post three times a year, we cannot. In scientific publication data, affiliation changes can appear delayed even by whole years, since the publication process is lengthy. Similarly, multi-language users do not necessarily correspond to migrants, so nationality estimation needs to take multiple factors into account. Even if, taken one by one, proxies are noisy, the hope is that, since we are in the realm of big data, we have enough measurements so that the overall signal is still strong. By aggregating measurements across many users, we can still observe patterns similar to what we see in official register data (e.g. Zagheni et al., 2014; Kim et al., 2020), but with the advantage of real time and high spatial coverage. Furthermore, corrections can be applied through modelling techniques to further improve performance (Rampazzo et al., 2021; Alexander et al., 2022).

However, this promise of big data is not always fulfilled. Platform data are never complete: they carry intrinsic selection bias and APIs provide access to only a subsample of the full dataset. This bias can be somewhat addressed with models taking into account platform adoption data. The migrants least likely to appear in these datasets, those who are undocumented, elderly, rural, digitally marginalised, or constrained in their platform use, are not randomly missing. They are disproportionately among the populations most exposed to the policy decisions that proxy-based indicators may ultimately inform. These filters compound one another. For geolocation-dependent migration research, only a small fraction of posts carries usable location data. Migrants represent roughly 3.7% of the world population (United Nations, 2024), and some categories, irregular migrants, refugees, may avoid social media altogether. The cumulative effect can be dramatic: one Twitter study (Kim et al., 2022) began with 59.4 million user profiles and obtained only 4,940 migrants after applying all filters. Augmented and more stable data access is therefore a structural requirement. Initiatives such as Meta's Data for Good (Weber et al., 2021) and Google's COVID-19 Community Mobility Reports (Aktay et al., 2020) show that company-led indicator production is feasible and has already supported substantial migration and mobility research (Minora et al., 2023; Bailey et al., 2018; Sulyok & Walker, 2020). Whether such voluntary efforts can substitute for more systematic governance of data access, through frameworks such as the EU Digital Services Act, remains an open question.

For proxy-based indicators to inform policy responsibly, three conditions must hold. First, infrastructural continuity: APIs, access schemas, and extraction rules change frequently, and an indicator that shifts because a platform has changed its architecture is not measuring migration consistently. Second, methodological transparency: indicators should be accompanied by documentation of code, modelling assumptions, and known limitations, this applies especially to company-generated indicators whose procedures are rarely disclosed. Third, validation: indicators should be benchmarked against independent sources before operational use, and where they serve early warning functions, ex post evaluation should be standard practice to avoid the silent institutionalisation of unreliable signals.

Validation against independent data sources is also necessary to assess whether indicators correspond to meaningful migration-related trends rather than platform

artefacts or sampling distortions. Where such indicators are used for early warning or anticipatory purposes, ex post evaluation should become standard practice, both to assess predictive validity and to avoid the silent institutionalisation of unreliable signals.

An important aspect of social media research is ethics and privacy. In terms of privacy, studies should adhere to GDPR rules, working on anonymised, minimised data, and publishing aggregated indicators. However, for migration research, group privacy should also be considered, so simple aggregation may not be enough, but some further data minimisation could be required for the indicators themselves. For instance, migrant groups could be further aggregated to offer protection, or publication could be delayed protecting migrant groups. Finally, dual-use risks must be treated as a core governance issue rather than an afterthought. Real-time migration indicators can support humanitarian preparedness, but under different institutional conditions they can also become instruments of surveillance, exclusion, or anticipatory control. Researchers and private companies producing such indicators should therefore evaluate not only methodological validity, but governance risk. The question is not simply whether an indicator works, but what kinds of decisions it may legitimise, and for whom.

Mobile Phones and the Migrant by Proxy: Mobility through Call Detail Records (Albert Ali Salah)

Mobile phone data, particularly Call Detail Records (CDRs), have become one of the most prominent sources for studying human mobility, especially where conventional official data are unavailable or delayed. Their use in displacement and humanitarian research has expanded substantially over the past decade, with systematic reviews documenting applications across crisis response, seasonal mobility, and forced displacement in low- and middle-income contexts (Aydoğdu et al., 2025a). Their analytical appeal lies in temporal granularity, scale, and near real-time responsiveness. Yet, as with other proxy-based systems discussed in this paper, what becomes visible is not migration directly, but behavioural traces inferred through telecommunications infrastructures. In CDR, each interaction between a mobile device and telecommunications infrastructure generates a timestamped record linked to a cell tower, producing large-scale datasets that reveal patterns of movement at high spatial and temporal resolution.

In migration research, such data have been used to detect displacement, estimate seasonal mobility, and reconstruct trajectories that are difficult to capture through surveys or administrative systems (Salah et al., 2019). Especially in Africa and South Asia, such data are used extensively (Blumenstock, 2012; Blanchard & Rubrichi, 2025). Governments and international organisations purchase such data (in a processed way) from telecommunications operators or intermediaries to fuel predictive or descriptive models. In some contexts, official statistics may be delayed, incomplete, or politically contested, making alternative mobility indicators analytically valuable. In our own research, we used CDR/XDR analysis to check what actually happened when Turkey opened the borders to Europe in March 2020 (Arcila Calderón et al., 2025), and to understand how Syrian refugees were affected by the Turkey-Syria earthquakes in 2023 (Aydoğdu et al., 2025b). Other work used mobile data to understand the aftermath of disasters, such as the earthquake in Haiti (Lu et al., 2012), and a cyclone in Mozambique

(Cumbane & Gidófalvi, 2021), and after earthquakes. Aydoğdu et al. (2025a) summarize the uses of mobile data in such contexts.

From the perspective developed in this paper, such data constitute infrastructural proxies, where mobility is inferred from interactions between devices and network architecture rather than from attributes of individuals. Migrants do not appear directly; instead, they are inferred from patterns such as a SIM card disappearing from one region and reappearing in another, or clusters of devices crossing administrative or national boundaries. There is also a difference between what kinds of analyses can be performed within the telecom operator, and what is possible on data that are shared by the operator. While the operator has access to a large set of demographic and usage-based indicators internally, due to privacy and data sensitivity, data sharing dissociates these variables through anonymisation and aggregation. In rare cases, a single demographic tag is used to separate native users from some group of interest based on certain intuitive indicators. In the Data for Refugees Challenge, for example, user's registration ID, and whether they were able to purchase a subsidised "refugee tariff" were used to tag Syrian refugees in the data (Salah et al., 2019). While this was a noisy indicator, it still allowed to observe the larger patterns between the refugee and native groups.

While analytically powerful, mobile proxies face a fundamental limitation: they detect mobility, but do not directly identify migration as a demographic or legal category. The core issue lies in the ambiguity of the unit of observation. In mobile phone data, the observable entity is not the individual, but the device or subscription. A SIM card is not a person, even though it can be legally tied to a person. Yet the relationship between devices and individuals may be unstable - phones may be registered to other individuals and family members, shared, replaced, or used intermittently, and registration practices vary across contexts. In some areas there are practices of phone sharing, whereas in others, females rarely own the phone line they use, and the SIM card is registered on a husband, father, or brother. Children cannot own a phone line legally and are missing as a demographic. As a result, positional changes in network data cannot be straightforwardly translated into changes in residence, legal status, or membership in categories such as migrant or refugee. A woman whose mobility is tracked through her husband’s registered SIM card does not appear in displacement estimates as an individual. She remains invisible even when she has moved, her trajectory effectively absorbed into someone else’s data trace. These are not incidental data quality issues. They determine whose mobility becomes measurable and whose remains structurally absent from the analytic field.

Interpreting such data as migration therefore relies on a chain of assumptions: that one device corresponds to one individual, that device movement reflects human movement, and that sustained presence indicates relocation. These assumptions may hold in some contexts and fail in others, meaning that CDR-based indicators are constructed rather than directly observed measures of migration. This is a well-known trade-off in computational social science: individual traces are inherently noisy and carry little meaning in isolation, yet aggregate patterns derived from large volumes of such traces can still reliably reflect real population dynamics. The strength of CDR-based measurement lies precisely at this aggregate level, not in the individual record, but in the convergence of millions of traces. This aggregate utility, however, should not be mistaken for representational completeness: a proxy can perform well at population level while systematically excluding specific categories of people.

This does not mean that mobile data are analysed only at aggregate level. Many analyses begin by inferring home and work locations from individual communication patterns. In a recent paper, we proposed that this view is limiting in a sense, and a broader approach that computes an "activity space" for the individual may be preferable for some analyses (Aydoğdu et al., 2025b). In the aftermath of an earthquake, for example, we could look at shifts in the usage of activity space to realise how a person's patterns changed. For migration research specifically, this is consequential: displacement and forced mobility tend to compress, fragment, or restructure a person's activity space in ways that a home-location proxy alone cannot capture. A refugee whose activity space collapses from a city to a camp, or expands across borders, is misrepresented by any indicator that reduces their presence to a single anchor point. But ultimately, the individual patterns are aggregated to derive concrete indicators.

More granular Extended Data Records (XDRs) address some limitations of CDRs by capturing continuous data interactions, Internet sessions, application usage, rather than discrete call events, improving temporal resolution where infrastructure allows. However, XDRs deepen the abstraction: social connections visible in call records are replaced by denser behavioural traces whose interpretive meaning is less transparent. Both CDR and XDR data capture behavioural traces of movement and connectivity without contextual information about motives or legal status, a structural difference from the discursive and intentional signals captured by social media or search data. This makes triangulation with demographic and qualitative sources not optional but necessary for any policy-relevant use.

The use of mobile phone data raises broader questions of governance legitimacy, access, and accountability. These data are typically controlled by private companies and shared under restricted agreements, creating asymmetries in who can observe mobility and under what conditions. Legal frameworks governing access vary substantially across countries. In some countries, governments can access such data easier, and in some, even at a personal level. The purchase of telco mobility data by European governments following the 2022 Ukrainian displacement crisis illustrates both the speed at which such arrangements are made and the governance gaps that accompany them: data were acquired and deployed in policy contexts before standards for transparency, validation, or migrant notification had been established. At the same time, individuals whose data generate these proxies have limited awareness of how their traces are used, and limited capacity to contest resulting classifications.

From the perspective of the “migrant by proxy”, mobile phone data exemplify both the analytical promise and the structural limits of proxy-based migration measurement. They offer an unusually powerful lens on population mobility at scale, yet what they capture remains a selectively measurable behavioural footprint rather than migration as lived social experience. Their apparent immediacy should not obscure the infrastructural assumptions, exclusions, and governance arrangements on which they depend. Ensuring that such proxies enrich rather than distort migration research requires clearer documentation of how indicators are constructed, systematic triangulation with other data sources, and careful scrutiny of the decisions these indicators may ultimately inform.

From Proxies to Populations: Aligning Data Science with Demographic Research (Haodong Qi)

Digital traces, including mobile phone records, geolocated social media activity, remittance transactions, and online search trends, have expanded the empirical repertoire of migration research by providing timely and scalable indicators of migration-related behaviour. Yet these data differ fundamentally from official population statistics. They emerge from selective digital infrastructures rather than defined sampling frames or administratively enumerated populations, with user bases that vary substantially across platforms, countries, and time periods (Hargittai, 2015, 2020). The central demographic question is therefore not simply what these data reveal, but whom they represent.

Demographic methods offer a structured framework for assessing the representational limits and conditional interpretability of organic digital trace data. At their core, demographic approaches emphasize population definitions, coverage, and selectivity, which are essential when working with non-random digital populations. Techniques such as benchmarking, post-stratification, and reweighting allow researchers to compare the observable characteristics of users generating digital traces with known population distributions derived from censuses, registers, or large-scale surveys. For example, when age, sex, or regional distributions of platform users can be inferred or approximated, these can be aligned with official demographic structures to evaluate who is systematically over- or under-represented among observed digital migrants.

Beyond compositional comparisons, demographic methods also focus on rates and transitions, which are central to migration analysis. Digital traces often capture events or behaviours, such as changes in inferred location, remittance activity, or cross-border communication, that can be interpreted as migration-related transitions. Demographers can assess the plausibility of these indicators by comparing derived migration rates, durations of stay, or origin–destination patterns with those observed in administrative or survey-based data. Life table approaches and consistency checks across age and time can help identify implausible patterns that may signal measurement error, platform artifacts, or algorithmic bias rather than genuine population processes.

Recent studies have sought to mitigate sampling biases by reconciling digital trace data with surveys, censuses, and population registers (Rampazzo et al., 2021; Yildiz et al., 2025; Qi et al., 2025). These efforts demonstrate that, under certain conditions, digital indicators can approximate migration trends and population distributions well. However, they also reveal persistent and uneven challenges. The availability and continuity of digital trace data vary widely across countries, reflecting differences in infrastructure, regulation, platform penetration, and data access arrangements. For example, data on users of most western digital services, including Google, Facebook, Twitter, are overwhelmingly missing in Mainland China due to legal restrictions. Moreover, platform usage is neither stable nor uniform over time and space, for instance, when Twitter transitioned to X, the user population has dropped sharply in many parts of the world. Such instability driven by changes in company's ownership or management complicate longitudinal and comparative analyses of migration processes.

Additional challenges stem from the opacity of data-generating process. Algorithms used to infer users' locations, movements, or residence histories are typically proprietary, undocumented, and subject to change, making demographic validation difficult. Moreover, the denominator problem remains acute: while digital traces provide rich

information on users, the total user population, and its relationship to the general population, is often poorly understood. Usage rates of major platforms such as Google or Facebook vary by age, socioeconomic status, country of origin, and/or nationality, yet these variations are rarely observable directly, limiting the effectiveness of standard demographic adjustment techniques, e.g., standardising population by age, gender, or other characteristics. Those most systematically absent from platform data, irregular migrants, those in detention, those without smartphones or stable connectivity, are precisely the individuals whose situations most urgently require accurate measurement, yet who leave the fewest traces for demographic methods to work with.

Addressing these challenges requires both methodological discipline and institutional accountability. From a research perspective, greater emphasis on methodological triangulation (Bircan & Qi, 2025), using multiple digital sources alongside traditional demographic data, can help identify robust signals and reduce reliance on any single platform. Sensitivity analyses, uncertainty quantification, and explicit documentation of assumptions about coverage and selectivity should become standard practice. At the institutional level, different stakeholders have distinct yet complementary roles to play. Governments and national statistical offices can facilitate validation by expanding access to secure, privacy-preserving administrative and register data that enable demographic benchmarking. They can also develop official guidelines for the responsible use of organic data in population statistics, clarifying where such data can supplement, but not replace, traditional sources. Companies that own digital trace data can contribute by increasing transparency around data generation processes, platform coverage, and algorithmic changes, and by supporting standardized metadata and aggregate statistics that enable demographic assessment without compromising user privacy.

III. Discussion Essay: Consent, Feasibility, and Accountability (Tuba Bircan)

Together, the reflections collected here show that migrants never appear directly in big data: they are always mediated through stand-ins. These proxies differ in kind, discursive framings, social network traces, geolocated signals, search queries, demographic alignments, but they share two features. They are partial, capturing only fragments of the trajectories they claim to represent, and they are consequential, because once adopted they travel into media narratives, policy dashboards, and governance systems that shape migrants' lives. The concept of the "migrant by proxy" names this double movement: measurement that simultaneously reveals and constructs its object.

The contributions to this paper illuminate how this construction operates across three dimensions. As technical indicators, proxies translate behavioural or infrastructural traces into tractable signals, at the cost of ambiguity, context, and temporal complexity. As epistemic operators, they determine which forms of mobility become thinkable and which populations remain invisible: the undocumented, the elderly, the rural, those who do not post, those whose SIM cards are registered to someone else. As governance devices, they feed early warning systems, displacement estimates, and policy models, stabilising particular interpretations of migration before their assumptions have been examined. A central finding across the sections is that these three dimensions do not align: a proxy that performs well technically may distort migration as an object of knowledge, while being operationalised as a decision-support tool regardless.

A second shared finding concerns the structural absence at the centre of proxy-based measurement: the migrant increasingly appears as a data point rather than as a person with a voice, a trajectory, and a stake in how they are classified. Put differently, the migrant increasingly appears as a data point rather than as a social actor. In proxy-based systems, what becomes visible is not the person in their social, legal, and biographical complexity, but a collection of measurable attributes, locations, interactions, transactions, or inferred characteristics. The migrant becomes legible as a pattern in data rather than as an individual capable of explaining their own trajectory. The expression “migrant as a data point” therefore does not imply simple dehumanisation; it describes a shift in how migration becomes known through computational systems, where traces stand in for persons and measurement substitutes for direct experience. Discursive proxies construct the migrant as a target of others' hostility; mobility proxies construct them as a pattern of device movements; demographic proxies construct them as a residual category after weighting adjustments. What none of these can access, and what big data approaches cannot substitute for, is the migrant's own account of their movement, status, and identity. This is not merely an ethical point but an epistemological one: self-narratives, qualitative interviews, and community-based research generate knowledge about migration that is structurally unavailable to any proxy. A fully accountable approach to migration data must be explicit about this boundary rather than eliding it.

This boundary is sharpest at the point of consent and governance accountability. Individual consent frameworks are inadequate for proxy-based data, since migrants are typically unaware that their traces are being used, and because the relevant harms are often collective rather than individual, affecting populations defined by nationality, ethnicity, or movement pattern rather than identifiable persons. Resolving the individual consent problem does not resolve the group privacy problem: aggregated and anonymised data can still produce actionable classifications of migrant populations that expose them to harm. This risk is compounded when proxy-based indicators are embedded in AI-driven governance systems, in border management, asylum determination, and predictive policing, where migrants are rendered into data points with little transparency, oversight, or right of contestation (Bircan & Korkmaz, 2021; Molnar, 2019; Beduschi, 2021; Molnar, 2021). The question that must become standard is not only how proxies are built, but which decisions about migrants they can legitimately inform, and under what conditions of accountability.

Our author team itself exemplifies the definitional tensions this paper describes. Each of us has experienced migration in different forms and at different moments of our lives, yet bibliometric databases classify us by our current European institutional affiliations. We appear in these systems not as migrants but as European researchers. In this sense, we too are migrants by proxy, visible in one register, invisible in another, classified by institutional convenience rather than biographical reality. This is not merely a reflexive observation: it demonstrates concretely that the choice of proxy is always also a choice about who counts, and for what purposes. Different proxies produce different migrant populations, with different policy implications. Recognising this is the first condition for responsible use.

The lesson that emerges is not that big data proxies should be abandoned, the contributions here demonstrate their genuine analytical value across multiple domains, but that their use requires a discipline that the field has not yet institutionalised. Three

concrete demands follow from this paper's analysis. First, proxy documentation: every indicator used in research or policy should specify its data source, the population it captures and the population it misses, its known failure modes, and the governance decisions it is and is not appropriate to inform. Second, proxy triangulation: no single proxy should anchor a policy-relevant finding; convergence across independent proxies, and between proxy-based and qualitative sources, should be a standard of evidence. Third, proxy accountability: when the conditions of data collection change, platforms close APIs, companies alter algorithms, crises disrupt normal mobility patterns, the indicators built on them should be re-validated rather than silently extended. These are not calls for more careful research in the abstract. They are institutional requirements, for funders, journals, statistical offices, and the private companies whose data increasingly define who migrants are.

More broadly, our discussion calls for a recalibration of expectations surrounding big data in migration research. Digital traces can reveal patterns that are difficult or impossible to observe through conventional methods, particularly in contexts where timely information is unavailable. Yet no increase in volume, velocity, or computational sophistication can eliminate the inferential character of proxy-based measurement. Big data can enrich migration research, but it cannot fully resolve the conceptual, representational, and ethical challenges involved in knowing migration. Recognising both the possibilities and the limits of proxies is therefore a precondition for their responsible use in science and governance.

Competing Interests: The authors declare no competing interests.

References

Agre, P. (1997). *Computation and human experience*. Cambridge University Press.

Alexander, M., Polimis, K., & Zagheni, E. (2022). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, 41(1), 1–28.

Arcila Calderón, C., Aydoğdu, B., Bircan, T., Gunduz, B., Ones, O., Salah, A.A., & Sîrbu, A. (2025). Combining Twitter and Mobile Phone Data to Observe Border-Rush: The Turkish-European Border Opening. *Journal of Computational Social Science*, 8(26).

Arcila Calderón, C., Blanco-Herrero, D., Frías-Vázquez, M., & Seoane-Pérez, F. (2021). Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius. *Sustainability*, *13*(5), 2728. https://doi.org/10.3390/su13052728

Arcila Calderón, C., Sánchez-Holgado, P., Quintana-Moreno, C., Amores, J., & Blanco-Herrero, D. (2022). Hate speech and social acceptance of migrants in Europe: Analysis of tweets with geolocation. *Comunicar*, 71, 21-35. https://doi.org/10.3916/C71-2022-02

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinger, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., Vispute, S., Vu, M., Wellenius, G., Williams, B., & Wilson, R.J. (2020). *Google COVID-19 Community Mobility Reports: Anonymization process description (version 1.1)*. arXiv preprint arXiv:2004.04145.

Aydoğdu, B., Bilgili, Ö., Güneş, S., & Salah, A.A. (2025a). Mobile phone data for anticipating displacements: Practices, opportunities, and challenges, *Data and Policy,* Vol.7, e5. https://doi.org/10.1017/dap.2024.94

Aydoğdu, B., Danış, D., Bilgili, Ö., Yıldızcan, C., Yağcıklı, S. N., Güneş, S., & Salah, A.A. (2025b). A novel activity space approach to discover displacement patterns via mobile phone data: An analysis of the 2023 Türkiye-Syria Earthquakes. *EPJ Data Science,* 14(61).

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *32*(3), 259-280.

Beduschi, A. (2021). International migration management in the age of artificial intelligence. *Migration Studies*, 9(3), 576–596.

Bircan, T. (2024). Augmentation or replication? Assessing big data's role in migration studies. *Data & Policy*, 6, e51.

Bircan, T., & Qi, H. (2025). New methodological approaches for migration and mobility studies: from traditional to big data. *Frontiers in Human Dynamics*, 7, 1710558.

Bircan, T., & Korkmaz, E. E. (2021). Big data for whose sake? Governing migration through artificial intelligence. *Humanities and Social Sciences Communications*, 8(1), 241.

Blanchard, P., & Rubrichi, S. (2025). A highly granular temporary migration dataset derived from mobile phone data in Senegal. *Scientific Data*, 12(1), 1051.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. Information Technology for Development, 18(2), 107-125.

Chouliaraki, L., & Georgiou, M. (2022). *The Digital Border: Migration, Technology, Power* (Vol. 44). NYU Press.

Cumbane, S.P., Gidófalvi, G. (2021). Spatial distribution of displaced population estimated using mobile phone data to support disaster response activities. *ISPRS Int J Geo-Inf* 10(6):421.

Ekman, M. (2019). Anti-immigration and racist discourse in social media. *European Journal of Communication*, 34(6), 606-618.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives européennes de sociologie,* 49(3), 401-436.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). *Using Twitter data to estimate the relationship between short-term mobility and long-term migration*. Proceedings of the 2017 ACM Web Science Conference, pp. 103–110.

Fontana, M., Belmonte, M., Bosco, C., Jusselme, D., Peters, A. M., Minora, U., ... & Verhulst, S. (2025). Anticipating human mobility: Methods, data, and policy in forecasting and foresight. *Data & Policy*, 7, e70.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR),* 51(4), 1-30.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Hargittai, E. (2015). Is bigger always better?? Potential biases of big data derived from social network sites. The ANNALS of the American Academy of Political and Social Science, 659, 63–76. https://doi.org/10.1177/0002716215570866

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38(1), 10-24.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Kim, J., Pratesi, F., Rossetti, G., Sîrbu, A., & Giannotti, F. (2023). *Where do migrants and natives belong in a community: a Twitter case study and privacy risk analysis*. Social Network Analysis and Mining, 13(1), 15.

Kim, J., Sîrbu, A., Giannotti, F., & Gabrielli, L. (2020). *Digital footprints of international migration on Twitter*. In International Symposium on Intelligent Data Analysis, pp. 274–286. Springer.

Kitchin, R. (2022). *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures, 2nd edition*. Los Angeles: Sage

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci* USA 109(29):11576–11581. https://doi.org/10.1073/pnas.1203882109

Mazzoli, M., Diechtiareff, B., Tugores, A., Wives, W., Adler, N., Colet, P., & Ramasco, J.J. (2020). Migrant mobility flows characterized with digital data. *PLOS ONE*, 15(3), e0230264.

Minora, U., Belmonte, M., Bosco, C., Johnston, D., Giraudy, E., Iacus, S.M., & Sermi, F. (2023). *The war in Ukraine and the potential of Facebook's Social Connectedness Index to anticipate human displacement*. Migration Research Series, N° 73. International Organization for Migration (IOM), Geneva.

Molnar, P. (2019). Technology on the margins: AI and global migration management from a human rights perspective. *Cambridge International Law Journal*, 8(2), 305–330.

Molnar, P. (2021). Robots and refugees: the human rights impacts of artificial intelligence and automated decision-making in migration. In M. McAuliffe (Ed.), *Research Handbook on International Migration and Digital* Technology (pp. 134–151). Edward Elgar Publishing.

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4), 2131-2167.

Murtfeldt, R., Paik, S., Alterman, N., Kahveci, I., & West, J. D. (2024). RIP Twitter API: A eulogy to its vast research contributions. *arXiv preprint arXiv:2404.07340*.

Palotti, J., Adler, N., Morales-Arilla, J., Purdie, M., Zagheni, E., Weber, I., & Fatehkia, M. (2020). *Quantifying international human mobility patterns using Facebook Network data*. PLOS ONE, 14(10), e0224134.

Pfeffer, J., Mooseder, A., Lasser, J., Hammer, L., Stritzel, O., & Garcia, D. (2023). This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. In *Proceedings of the international AAAI conference on web and social media* (Vol. 17, pp. 720-729).

Pollacci, L., Milli, L., Bircan, T., & Rossetti, G. (2025). Academic mobility from a big data perspective. *International Journal of Data Science and Analytics*, *20*(1), 107-120.

Qi, H., Reed, H.E. & Bevelander, P. Can internet search data predict human migration intentions?. *Comparative Migration Studies* 13, 28 (2025).

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., & Zagheni, E. (2021). A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom. *Demography*, 58(6), 2193–2218.

Salah, A.A., A. Pentland, B. Lepri, E. Letouze, Y.-A. de Montjoye, X. Dong, P. Vinck. (2019). *Guide to Mobile Data Analytics in Refugee Scenarios*, Springer International Publishing.

Salah, A. A., Korkmaz, E. E., & Bircan, T. (Eds.). (2022). *Data science for migration and mobility*. Oxford University Press.

Sanliturk, E., Zagheni, E., Dańko, M.J., Theile, T., & Akbaritabar, A. (2023). *Global patterns of migration of scholars with economic development*. Proceedings of the National Academy of Sciences, 120(4), e2217937120.

Sismondo, S. (2011). *An introduction to science and technology studies*. John Wiley & Sons.

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). *Migration of professionals to the U.S.: Evidence from LinkedIn data*. In L. M. Aiello & D. McFarland (Eds.), Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol. 8851, pp. 531–543. Springer.

Sulyok, M., & Walker, M. (2020). Community movement and COVID-19: A global study using Google's Community Mobility Reports. *Epidemiology & Infection*, 148, e284.

United Nations (2024). *International Migrant Stock 2024: Key facts and figures*. UN DESA/POP/2024/DC/NO. 13. United Nations, New York. https://www.un.org/development/desa/pd/content/international-migrant-stock

Weber, I., Imran, M., Ofli, F., Mrad, F., Colville, J., Fathallah, M., Chaker, A., & Seed Ahmed, W. (2021). Non-traditional data sources: Providing insights into sustainable development. *Communications of the ACM*, 64(4), 88–95.

Willekens, F. (2019). Evidence-based monitoring of international migration flows in Europe. *Journal of Official Statistics*, 35(1), 231-277.

Yildiz, D., Wiśniowski, A., Abel, G. J., Weber, I., Zagheni, E., Gendronneau, C., & Hoorens, S. (2025). Integrating traditional and social media data to predict bilateral migrant stocks in the European Union. *International Migration Review*, 59(1), 90-118.

Zagheni, E., Garimella, V.R.K., Weber, I., & State, B. (2014). *Inferring international and internal migration patterns from Twitter data*. Proceedings of the 23rd International Conference on World Wide Web (WWW '14), pp. 439–444. ACM.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4), 721–734.