Data & Policy (2023), 5: e42
doi:10.1017/dap.2023.47

CAMBRIDGE UNIVERSITY PRESS

TRANSLATIONAL ARTICLE

Accelerating and enhancing the generation of socioeconomic data to inform forced displacement policy and response

Patrick Michael Brock and Harriet Kasidi Mugera

World Bank UNHCR Joint Data Center on Forced Displacement, Copenhagen, Denmark
Corresponding author: Patrick Michael Brock; Email: brock@unhcr.org

Received: 27 March 2023; **Revised:** 26 October 2023; **Accepted:** 08 November 2023

Keywords: data science; displacement; innovation; living conditions; socioeconomic data; wellbeing

Abbreviations: JDC, The World Bank UNHCR Joint Data Center on Forced Displacement; UNHCR, The United Nations Refugee Agency

Abstract

There are now an estimated 114 million forcibly displaced people worldwide, some 88% of whom are in low- and middle-income countries. For governments and international organizations to design effective policies and responses, they require comparable and accessible socioeconomic data on those affected by forced displacement, including host communities. Such data is required to understand needs, as well as interactions between complex drivers of displacement and barriers to durable solutions. However, high-quality data of this kind takes time to collect and is costly. Can the ever-increasing volume of open data and evolving innovative techniques accelerate and enhance its generation? Are there applications of alternative data sources, advanced statistics, and machine-learning that could be adapted for forced displacement settings, considering their specific legal and ethical dimensions? As a catalytic bridge between the World Bank and UNHCR, the Joint Data Center on Forced Displacement convened a workshop to answer these questions. This paper summarizes the emergent messages from the workshop and recommendations for future areas of focus and ways forward for the community of practice on socioeconomic data on forced displacement. Three recommended areas of future focus are: enhancing and optimizing household survey sampling approaches; estimating forced displacement socioeconomic indicators from alternative data sources; and amplifying data accessibility and discoverability. Three key features of the recommended approach are: strong complementarity with the existing data-collection-to-use-pipeline; data responsibility built-in and tailored to forced displacement contexts; and iterative assessment of operational relevance to ensure continuous focus on improving outcomes for those affected by forced displacement.

Policy Significance Statement

To design effective responses and policies to improve the quality of life for people who have been forced to flee their homes, or who are otherwise affected by forced displacement, governments, and international organizations require relevant, high-quality, and timely data. The process of collecting this kind of data is costly and takes time. Are there ways that the ever-expanding universe of big data and the fast-developing analytical tools of data science can help? We convened a workshop to find out, and here present recommendations for the most effective ways to focus effort on applying innovative methods and using alternative data sources to accelerate and enhance the generation of socioeconomic data on forced displacement, towards improving outcomes for those affected by it.

1. Introduction

The United Nations Refugee Agency (UNHCR) estimates that there are 114 million forcibly displaced people worldwide, some 88% of whom are in low- and middle-income countries (UNHCR, 2023). As forced displacement has received increasing international attention during the last 6 years, so has the lack of appropriate data needed to design policies and responses to it (Baal, 2021). This was recognized, for example, by the 2016 World Humanitarian Summit, which led to the commitments of the Grand Bargain,[1] by the United Nations Secretary General's High-Level Panel on Internal Displacement as one of its five priority themes,[2] by the Global Action Plan to End Statelessness,[3] and by the Global Compact on Refugees (United Nations, 2018). It has also been recognized by development actors, such as the World Bank, for example in its flagship report on the topic (World Bank, 2017), and through its commitment to follow-up on priorities identified since. This data gap is the key challenge that this paper aims to contribute to addressing. Despite this data gap, there is a growing evidence base on how the socioeconomic data on forced displacement that is available can inform operations (e.g., targeting of social protection programs[4]), and policy[5] (e.g., national policies on labor market and public service access for migrants[6]), which emphasizes the importance of addressing it.

Despite tangible progress on data and evidence on the socioeconomics of forced displacement during the last 5 years,[7] the volume and complexity of the needs of those affected by forced displacement continue to increase (World Bank, 2023).[8] Given this need, and given the ever-increasing volume of openly available data, can the rapidly evolving advanced innovative techniques of data science and other expertise be applied to accelerate and enhance the generation of good quality data needed to inform effective policies and operations on, and responses to, forced displacement? As a catalytic bridge between the World Bank and UNHCR, the Joint Data Center on Forced Displacement (JDC) convened a workshop to answer these questions.

We brought together 70 specialists from 20 international organizations, development partner governments and universities, for the exploration of ideas, the identification of potential activities, and collective brainstorming on feasibility. We asked: *what are the best opportunities for applying innovative tools and data science to meet the challenges of forced displacement during the next 5 years?* We posed this question with a view to informing an agenda for the next 5 years to apply innovative tools and data science to strengthen the body of socioeconomic data and evidence that is in use to inform operations, programs, and policies toward improving outcomes for the forcibly displaced and their hosts. The workshop began with a phase of expansive collation during plenary sessions that were followed by a focused assessment during breakout groups. This paper summarizes the emergent messages from the workshop and recommendations for future areas of focus and ways forward for the community of practice on socioeconomic data for forced displacement.

The energy and the richness of discussions demonstrated that these conversations were needed. Data science is applied by humanitarian and development organizations more and more, including by UNHCR and the World Bank, and is having a significant positive impact, from the automation of reproducible analytical pipelines to prediction using machine learning.[9] The workshop identified concrete opportunities to build on and expand this use of data science through its application to accelerating and enhancing

[1] https://interagencystandingcommittee.org/system/files/grand_bargain_final_22_may_final-2_0.pdf
[2] https://www.un.org/internal-displacement-panel/sites/www.un.org.internal-displacement-panel/files/tor_of_the_panel.pdf
[3] https://www.unhcr.org/ibelong/global-action-plan-2014-2024/
[4] https://blogs.worldbank.org/voices/moving-toward-data-driven-policies-aid-people-forced-flee
[5] https://www.jointdatacenter.org/heres-how-other-refugee-emergencies-can-guide-government-response-to-the-ukraine-crisis/
[6] https://www.jointdatacenter.org/jdc-newsletter-january-2023/
[7] https://www.worldbank.org/en/programs/building-the-evidence-on-forced-displacement & https://www.jointdatacenter.org/
[8] https://pro.drc.ngo/resources/news/gdf-report-2023/
[9] https://www.unhcr.org/61bc6ae84.pdf

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

the generation of socioeconomic data on forced displacement, as well as approaches to cement the sustainability of these approaches within organizations.

2. Emergent Themes

Many discussions returned to the theme of applicability of innovative and data science approaches developed in other settings to forced displacement. Two considerations emerged: (i) a technical consideration, i.e., whether methods and innovative solutions developed in other settings can be adapted to provide useful insight in, for example, refugee camps; (ii) and as two-pronged capability consideration, i.e., whether organizations working on forced displacement are set-up to apply advanced analytical approaches drawing on alternative data sources, as well as whether they are set-up to effectively and responsibly use their outputs. Two distinct capability gaps were identified across multiple discussions: the expertise that analysts require to apply innovation and data science effectively, and the awareness and trust that non-analysts require to manage, understand, and make effective and appropriate use of the outputs. There was neat crossover between them in the form of a technical solution to increasing trust in model outputs as implemented for example by the World Food Programme Hunger Map Live.[10] The feature integrates estimated uncertainty as well as information on which key variables influence predictions. The capability discussion highlighted the value of iterative collaborative working between technical, data science, and operational teams, which not only builds trust but also ensures that outputs are operationally relevant. The collaboration between United Nations (UN) Global Pulse and UNHCR headquarters teams working with UNHCR Brazil on predicted border crossings provides a good and practical case study in this regard.[11]

Participants agreed that when it comes to socioeconomic data on forced displacement (as in many other spheres), the value of data is in its use, and that everyone involved in the data-collection-to-use pipeline (or data value chain) must do all they can to ensure the effective use of outputs with due ethical and legal considerations. This raised the question of the scale at which indicator estimates need to be representative (e.g., of a country, region, or other administrative unit, such as a refugee camp) to be actionable, on which perspectives differed in line with organizational priorities. For some organizations, for estimates to be actionable, they will usually need to be representative at a national or state level, given the decisions they will inform. For others, a priori insistence on national or state-level representative estimates can obscure the more important question of what is needed and actionable when it comes to informing a varied set of operational and policy decisions.

Even as workshop conversations catalyzed new collaborative links between participants, there was a recurrent question on who we mean by 'we'. A cross-organizational community of practice was implicit in many discussions and will be necessary for realizing much of the potential the workshop identified. Different organizations have different priorities, but there were many areas in which collaborative work could minimize the duplication of effort, build new partnerships, and accelerate progress. For example, there were calls for forums for information and knowledge exchange, as well as for new frameworks for sharing code, tools, data, and models. The pooling of expertise, experience, and example use cases would also help ensure that data responsibility is central to new initiatives, and that newly identified risks are quickly understood and mitigated, as frameworks for trusted information exchange allow risks experienced by one organization to be avoided by others. The Data Responsibility Working Group[12] Operational Guidance on Data Responsibility in Humanitarian Action[13] Annex of Examples of Data Responsibility in Practice is a good example.[14]

[10] https://hungermap.wfp.org/

[11] https://www.dropbox.com/s/m1s0fn8xjf3vkb8/DeNieves_et_al_Modeling_Population_Movements.pdf?dl=0

[12] https://reliefweb.int/topics/data-responsibility-working-group-drwg

[13] https://interagencystandingcommittee.org/operational-response/iasc-operational-guidance-data-responsibility-humanitarian-action

[14] https://docs.google.com/document/d/1f5zOBLaL8mlitmOZBiLTnsVQCqyh5XnqJXt_WHooDXs/edit

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

There was a strong emerging message on the need for the application of innovative tools and data science to be complementary to existing approaches used in the current data collection-to-use pipeline for socioeconomic data on forced displacement. Many of the opportunities identified during discussions were dependent on the availability of high-quality survey data (for example, to train prediction models using alternative data sources as inputs), and others were innovative enhancements of traditional approaches (for example, geospatial data to inform household survey sampling approaches). The approach of the global community of practice on household survey statistics and the approach it has taken to the harmonization of concepts and definitions over time could provide some useful lessons and insights.

3. A New Agenda

Workshop discussions were divided into four topic areas, which were chosen based on need, and incorporated the possibility of scoping new tools as well as the adaptation of existing ones to forced displacement settings. The four topic areas were used as a framework for the expansive collation of ideas and experience, followed by focused assessment to examine and define the challenges and opportunities of possible new adaptations and approaches, and then to filter them according to feasibility and operational relevance. This line-of-sight to operational relevance was a key consideration of the workshop but the aim was not to exhaustively assess ideas with colleagues from country and regional operational teams. Instead, the workshop was a pre-stage to this, during which the participants scoped, identified, and explored promising tools and approaches, for further context-specific refinement with operational teams in contexts where they could be applied.

The four topic areas were:

- Socioeconomic indicator estimation using alternative data sources in forced displacement settings, which covered the challenges and opportunities of using satellite imagery and other alternative data sources for estimating the well-being and living conditions of refugees, host populations, and internally displaced people (IDPs).
- Natural language processing to improve forced displacement knowledge and data management, which looked at the most promising opportunities for increasing the discoverability and use of relevant, high-quality socioeconomic data and evidence on forced displacement.
- Integrating socioeconomic estimates into predictive models of displacement, which explored the potential for enriching the predictions of numbers of people displaced with estimates of their needs.
- Tools to facilitate the collection and improve the quality of survey data in forced displacement settings, which covered the best approaches to accelerate and enhance the rate at which relevant, high-quality socioeconomic survey data on forced displacement is generated.

The natural experiment of arranging discussions around these four topics had four clear outcomes. First, there are opportunities for the productive application of alternative data sources and innovative and data science approaches to enhance the sampling approaches of household surveys in forced displacement contexts. Second, there are opportunities for the application of natural language processing and machine learning to increase the discoverability to boost awareness and use of socioeconomic data on forced displacement. Third, there is potential for the estimation of socioeconomic indicators using alternative data sources and advanced analytics, and this potential is worth investigating and testing. Fourth, given the complexities of displacement prediction in general (Suleimenova et al., 2017; Huynh and Basu, 2020; Leasure et al., 2022; Pham and Luengo-Oroz, 2023),[15] including its intersections with environmental

[15] http://unglobalpulse.net/predictingdisplacement/; https://centre.humdata.org/catalogue-for-predictive-models-in-the-humanitarian-sector/; https://pro.drc.ngo/what-we-do/innovation-and-climate-action/predictive-analysis/foresight-displacement-forecasts/; https://elva.org/news/forecasting-displacement/; https://jetson.unhcr.org/; https://www.dropbox.com/s/m1s0fn8xjf3vkb8/DeNieves_et_al_Modeling_Population_Movements.pdf?dl=0; https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/15455.

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

modeling in humanitarian contexts[16] and conflict prediction,[17] any new drive for the integration of socioeconomic estimates into predictive models of displacement could prove premature, as well as giving rise to new data responsibility concerns.[18]

Given these main outcomes from discussions, as well as the themes that emerged across topic areas, particularly the emphasis on complementarity to traditional approaches, we have re-organized the potential contribution of innovation and data science to socioeconomic data on forced displacement identified by the workshop, emphasizing operational relevance and feasibility, by stage of the data-collection-to-use pipeline. This is shown below (Figure 1), with three pillars complementing, accelerating, and enhancing data collection, analysis, and dissemination. These would all be complemented by mechanisms for increasing information exchange and technical collaboration between organizations. It should be noted that these are specific to the intersection of data innovation with socioeconomic data on forced displacement, and do not apply more generally to the humanitarian and development sectors, for which prioritization would be different.

3.1. Sampling

On enhancing sampling approaches of socioeconomic surveys on refugees, host communities and IDPs, there are immediate opportunities for the use of innovation and alternative data sources. Adaptable tools are already in use by WorldPop[19] in support of national surveys. For example, methods to generate gridded population estimates (Qader et al., 2020) and algorithmically designate enumeration areas (Qader et al., 2021).[20] The innovative approach to using multiple geospatial datasets to ensure the representativeness of survey samples is particularly promising for application to forced displacement settings, given the lack of or outdated sampling frames, as well as its potential to simultaneously cut costs and increase sample quality and efficiencies. The development and adaption of grid sampling applications to forced displacement settings by the World Bank has been piloted in the Democratic Republic of Congo (DRC). Automatic enumeration area designation to forced displacement settings has been kick-started by a UNHCR and WorldPop pilot supported by the UNHCR Data Innovation Impact Fund in Cameroon. The results of this pilot will be relevant to all those designing sampling approaches in resource-poor forced displacement contexts, so there would be a collective benefit in consolidating and disseminating the findings and implications.

The input layers for this process could also be adapted to specific purposes in forced displacement settings. For example, in situations where only parts of countries, states or administrative areas need to be sampled. Building footprint layers could be used for over-sampling of host communities around refugee camps by guiding field teams to households or small building-seeded listing areas. There would be collective benefit in testing the efficiency of this approach, particularly through a direct experimental comparison with traditional listing approaches to compare costs, accuracy, and representativeness. There is a growing number of building footprint layers created in different ways and with variable qualities, as well as open-source models for image classification to identify buildings and shelters. However, there is a lack of a systematic comparison between the existing layers and a technical assessment of the performance of the accuracy of the layers and performance of the models in forced displacement settings; another opportunity for a review that would have cross-organizational benefits. The findings of this investigation

[16] https://www.internal-displacement.org/sites/default/files/publications/documents/220906_IDMC_DroughtDisplacementModelling.pdf

[17] https://www.diva-portal.org/smash/get/diva2:1628162/FULLTEXT01.pdf; https://unis-sahel.org/wp-content/uploads/2022/11/SAHEL-PREDICTIVE-ANALYTICS-REPPORT-compresse.pdf; https://centre.humdata.org/assessing-the-technical-feasibility-of-conflict-prediction-for-anticipatory-action/

[18] https://data4migration.org/articles/new-publication-the-good-and-bad-of-anticipating-migration/?s=03

[19] https://wopr.worldpop.org/?/preEAs

[20] https://wopr.worldpop.org/?/preEAs; https://www.worldpop.org/2022/09/12/benin-adopts-worldpops-preea-tool-to-conduct-its-first-modern-digitised-census/; https://www.worldpop.org/2022/05/20/supporting-socio-economic-surveys-in-the-drc/;

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

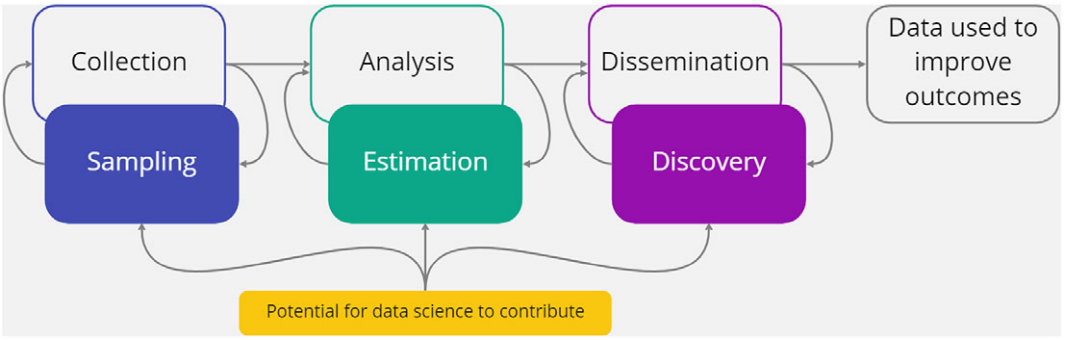

Figure 1. *The parts of the data-collection-to-use pipeline with the greatest potential for data science to contribute to accelerating and enhancing the generation and use of socioeconomic data on forced displacement.*

could be used to recommend the integration of open data sources into sampling applications[21] and tools for more effective use in forced displacement settings.

3.2. Estimation

On socioeconomic indicator estimation in forced displacement settings, there are a myriad of opportunities for the possible application of data science and use of alternative data sources; the challenge will be effectively prioritizing efforts in collective exploration, evidence gathering, and assessment. This potential is partly a result of the rapid and expansive progress in data handling and automation, for example, automated quality checks, the increasing availability of application programming interfaces (APIs), and frameworks for the integration of diverse data sources. The ability of models to update from live data sources, generate new estimates, and automatically disseminate these, shows how far such pipelines have come. Likewise, the potential is represented by synthetic data, which can be trained on national survey observations to produce non-sensitive microdata for a whole population, which allows for scenario exploration, including through simulation modeling, and hypothesis testing (Solatorio and Dupriez, 2023).

There remain methodological challenges to using alternative data sources and complex analytics to generate estimates of socioeconomic indicators. For example, the proportionate inclusion of expert human judgment, and on managing the tension between the comparability and cost-effectiveness of global models and their applicability to different contexts. The variability in data quality and data accessibility over time and in different places influences the accuracy, generalizability, and flexibility of global models. Given the potential of methods to estimate food prices from partial surveys,[22] and that food security was identified in discussions as a priority area, it would be of collective benefit to assess the feasibility of a partial survey modeling approach in terms of data availability and the validity of assumptions in forced displacement settings; likewise, the operational relevance of food security estimates made for forced displacement settings generated through cross-survey imputation.[23]

A complementary review that would also be of collective benefit could explore the applicability of an approach taking inspiration from the Hunger Map Live[24] and other food security estimation approaches

[21] http://www.worldbank.org/gems; https://mysurvey.solutions/en/; https://support.kobotoolbox.org/transcription-translation.html?utm_source=kobo&utm_medium=inapp

[22] https://documents1.worldbank.org/curated/en/185851639662039407/pdf/Estimating-Food-Price-Inflation-from-Partial-Surveys.pdf

[23] https://www.econstor.eu/bitstream/10419/216898/1/GLO-DP-0538.pdf; http://www.ecineq.org/milano/WP/ECINEQ2020-536.pdf.

[24] https://hungermap.wfp.org/

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

(Gholami et al., 2022) to forced displacement settings. This would involve an iterative examination of the utility of the outputs given their spatial resolution, asking what level of spatial disaggregation would be necessary to usefully inform operations related to forced displacement. Testing this operational relevance could also provide an opportunity to apply one of the innovative features of Hunger Map Live: the ability to show which input variables drive predictions (Martini et al., 2022). This could be used, for example, to highlight to UNHCR operations whether a sudden estimated increase in food insecurity in a particular place is driven by a displacement event (that UNHCR is likely to be aware of and already responding to) rather than an environmental change (that UNHCR may not be); the former was the case for a Hunger Map Live estimated increase in food insecurity in Malakal in South Sudan at the time of the workshop on 1 December 2022.

Although there has been mixed success in mapping poverty only using alternative data sources,[25] alternative data sources and advanced analytics can enhance household survey-derived estimates (Masaki et al., 2022), and may be especially helpful when survey data are old.[26] Given that household surveys in forced displacement contexts are newly aligning with international statistical standards and beginning to cover refugees, IDPs, and host communities in comparable ways, there is good potential for data scientists to add value by working alongside household survey teams collecting data on the forcibly displaced. For example, as UNHCR begins to collect comparable multi-topic household data on refugees and host communities in multiple countries,[27] there is an opportunity to collect relevant data from alternative sources in parallel. This could inform an assessment of the appropriateness of the data to generate operationally relevant insight (e.g., spatial resolution), and the ability of models trained on the household survey data to reliably estimate socioeconomic indicators from alternative data sources. These could be used to update indicator estimates between survey waves, generating indicator estimates at a fraction of the cost of a household survey and therefore feasible with a higher frequency.

On this topic in particular there is an emerging collective need for new approaches to sharing data, tools, insight, and experience, through ready access to data processed and documented in standardized ways that increase its usability and reduce the duplication of effort across organizations; for example, through the adoption and use of geospatial data schemas relevant to forced displacement. This could begin with forums for information exchange where they are non-duplicative, topic-focused, and can complement existing ones, for example, the UN Data Science Cell.

3.3. Data, evidence, and knowledge discovery

The use of natural language processing and associated modeling techniques is booming. This applies to the humanitarian and development space as well as to the private sector, even though the tools and applications used by the former are lagging behind those of the latter. Given the large amounts of unstructured text data relevant to forced displacement, there are valuable opportunities to use text analysis to increase the discoverability and thus, use of data and evidence to inform policies, responses, and programs. These include increasing the discoverability of microdata relevant to forced displacement through identification and tagging on dissemination platforms. This is underway for the World Bank Microdata Library,[28] and there would be collective benefit in applying the similar approaches on other platforms that host microdata with relevance for forced displacement, including government data portals (Ekhator-Mobayode and Hoogeveen, 2022).

Forced displacement topic modeling, in addition to making possible microdata tagging, allows for trend analysis that shows how forced displacement has been covered over time by multilateral

[25] https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099419209132236954/idu0fcecbbc004dd3041cc088360d1c57c5ffe04?cid=DEC_PolicyResearchEN_D_INT

[26] https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099430309142231728/idu0660868530404c0414e0bf180797b525682a5

[27] https://data.unhcr.org/en/documents/details/103441

[28] https://microdata.worldbank.org/index.php/home

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

development bank public documents, and the coverage of this tool[29] could be usefully expanded to include academic research literature and new and newly accessible large language models also present many opportunities for building on this work. This tool[30] that integrates natural language processing and forced displacement topic modeling has facilitated the piloting of an automatic approach to microdata use detection, which aims to identify reports, papers, and research that have used data hosted on the UNHCR Microdata Library,[31] even when the data has not been cited.[32] There would be a collective benefit in optimizing and extending this tool to include within citation text analytics, which would distinguish different kinds of data use, for example, reports that cite data once in the introduction as context as compared with those that cite data throughout, using it to generate new insight.

The automatic identification of text relevant to forced displacement also creates opportunities for applications to operational and program data. To inform allocation and prioritization decisions, named entity recognition and topic modeling could provide insight into the geographical overlap of humanitarian and development programs to identify complementarity and duplication under different allocation scenarios. There are three opportunities for testing the feasibility of such a system focused on the socioeconomics of forced displacement. First, to test the effectiveness of the automatic identification of program text, including results indicators, relevant to forced displacement in systems with consistent coverage (e.g., the World Bank Projects & Operations Explorer). Second, to test the consistency of information content of summary extracts of the text of UNHCR operational data documentation by subnational location and sector. Third, to test the coverage of in-depth project documentation across-organizational development activity data sources, such as the International Aid Transparency Initiative.[33] Such a system would also provide opportunities to unify approaches to advanced text analysis between humanitarian and development settings, as it would bridge between knowledge management and event detection.[34]

There are potential pitfalls on the road to widespread robust use of natural language processing tools in humanitarian and development organizations. For example, there is a tendency to conflate natural language processing with qualitative data analysis. Although natural language processing can be used for qualitative data analysis, its applications are much broader. To allow the perception of natural language processing to become pigeon-holed would be to miss out on its wider impact. There is also a danger that a lack of appropriate skills combined with techno-solutionism gives rise to biased approaches to advanced text analysis, for example, through naïve modeling of social media text data. These risks can be countered by better communication and engagement around text analysis techniques and applications, which could raise awareness as well as amplify the sharing of lessons and experience.

4. Strategic Directions

The three areas of strategic focus for data science for socioeconomic data on forced displacement identified by the workshop (sampling, estimation, discovery), are well integrated with the existing data-collection-to-use pipeline (or data value chain). For example, there are overlaps in key areas identified as priorities for the household surveys during the next decade, on the integration of new data sources, the use of machine learning, enhancement of sampling efficiency, and increasing data access and discoverability.[35]

[29] www.nlp4dev.org
[30] www.nlp4dev.org
[31] https://microdata.unhcr.org/index.php/home
[32] https://ofatunde.github.io/microdata-citation-explorer/intro.html
[33] https://iatistandard.org/en/; www.d-portal.org; https://github.com/datasciencecampus/iati-partner-search
[34] https://www.internal-displacement.org/monitoring-tools/monitoring-platform
[35] https://blogs.worldbank.org/opendata/8-technical-priorities-position-household-surveys-next-decade?fbclid=IwAR3G5o99nf1pCV9loQg4EvBHaNmyLxsaz65Cgkouz2Bg1TEAyWfgD7qpfuw&deliveryName=DM161645

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

How can teams and organizations capitalize on the opportunities identified here? One way would be to collectively invest in shared data and model reviews. As different organizations explore and use alternative data sources, they are likely to be duplicating efforts. There is a need to document, catalog, describe and assess alternative data sources and models relevant to the socioeconomics of forced displacement as public goods, and develop the frameworks and schemas to make this possible, covering ground-truthing, validation and benchmarking. The review approach would be flexible enough to reflect that the utility of data varies by intended use, partly to provide insight to different users, but also to facilitate the scoping of new applications of data sources and approaches, for example, building damage assessment tools for monitoring the degradation of refugee camp shelters. It would also need to build on, link to and reinforce the value of existing repositories and approaches, for example, the OCHA Center for Humanitarian Catalogue of Predictive Models[36] and Humanitarian Data Exchange.[37]

Given the potential vulnerability of those affected by forced displacement, and the specific legal and ethical dimensions of working with forced displacement microdata, the strong foundational data responsibility practice developed by UNHCR working with the World Bank should be built upon to understand and mitigate new risks. The use of alternative data sources could introduce new risks that are not already covered by existing experience, policies, and practice, especially if they are drawn from private sector systems that are not wholly transparent. Therefore, a continued focus on data responsibility is needed,[38] and the effective mitigation of newly emergent risks relies on the input of appropriate expertise.[39] As the temporal and spatial resolution of satellite imagery increases, data responsibility considerations change. New applications of existing data should be stress-tested for trade-offs between the benefit of additional socioeconomic insight and potential harm. Well-documented data and consistent systems for cataloging and searching datasets are key to keeping up with emergent data responsibility risks (Cervantes-Macías, 2022). Teams should likewise continuously assess the potential for algorithmic bias to cause harm and reduce the usefulness of analytical outputs. A principal driver of this agenda is the aim to make the needs of those affected by forced displacement more visible[40] so that operations, programming, research, and policy can better take them into account. However, there are well-documented risks of training algorithms on unrepresentative historical data,[41] and these should be carefully analyzed and assessed to counter the risk that the historical invisibility of the forcibly displaced in data and statistics undermines new attempts to better represent them.

There are efficiency gains to be made through enabling greater exchange of technical information and tools between actors, which would increase the overall rate of progress and bring benefits of greater transparency and consistency. Online collaboration infrastructure could collate code, data, and models, and provide a focus point for the development of harmonized definitions, standards, and protocols. Such collaboration tools, combined with the resource of data and model reviews, could enable more collaborative and agile procurement and sharing of private sector data between non-profit partners, which could save considerable investment in setting up new data-sharing agreements.

Given the complexity and dynamism of the challenges posed by forced displacement, and the rate of development of innovative tools and alternative data sources, teams and organizations will need to continue to work adaptively. Human resource strategies and organizational restructuring are likely to always lag the ever-changing wave of challenges and opportunities presented by forced displacement, so that the skills required are not commonly present. Nimble approaches to filling capability gaps, such as secondments and exchanges, are needed to deal with the resultant inevitable uneven distribution of skills and experience.

[36] https://centre.humdata.org/catalogue-for-predictive-models-in-the-humanitarian-sector/

[37] https://data.humdata.org/

[38] https://interagencystandingcommittee.org/operational-response/iasc-operational-guidance-data-responsibility-humanitarian-action

[39] https://reliefweb.int/topics/data-responsibility-working-group-drwg

[40] https://www.weforum.org/agenda/2021/11/refugees-internally-displaced-people-data/

[41] https://data4migration.org/articles/new-publication-the-good-and-bad-of-anticipating-migration/?s=03

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

On the question of whether effort and resources should be prioritized across the three identified areas of strategic focus, the workshop discussions highlighted both that there is a natural sequence from data collection through analysis to dissemination, and also that there are existing opportunities and needs across all of these areas. Therefore, these areas should be addressed in parallel, with the prioritization of effort and resources across them based on what can be done effectively now, with the understanding that as more relevant high-quality data becomes available (for example, through the UNHCR Forced Displacement Survey[42]), more analysis will be possible. How to define 'what is effective now' will require teams to make case-by-case judgment calls on balancing the direct operation-specific benefits of the generation of new data, and the indirect wider-impact benefits of generating new evidence to inform policy, without losing sight of the connectedness of the two undertakings.

5. Conclusion

The workshop was a deep dive into current and future opportunities for the application of innovative tools and data science for socioeconomic data on forced displacement. That this is an emergent area of opportunity should not detract from the broad positive impact that data science has already had, and continues to have, across humanitarian and development actors, especially when it comes to data management and reproducible analytical pipelines. A recent example from UNHCR is the use of a nowcasting analytical pipeline that draws on diverse data sources to impute numbers of people displaced, which informs country and regional operational planning. A recent example from the World Bank is a new synthetic data generation approach (Solatorio and Dupriez, 2023), one application of which is to responsibly test and optimize personal microdata risk assessment.[43]

Holding the hammer of new technology, there is always the danger of seeing every problem as a nail. However, the workshop discussions did not seem to flow from techno-enthusiasm or hype, but rather from a commitment to engage with the complexities of hard-to-use data sources and advanced analytical techniques driven by a belief in their potential to deliver value by informing responses to forced displacement. The current climate is favorable to investment in data-related activities, so we must collectively ensure that resources are directed to activities with impact. To do this, we need to build technical capability through hiring, training, and exchange, and to raise awareness through engagement and communication.

Acknowledgments. We would like to thank all the workshop participants for their contributions, and Olivier Dupriez, Tarek Abou Chabake, Aissatou Dicko, and Maja Lazić for their insightful comments.

Funding statement. The World Bank UNHCR Joint Data Center (JDC) is currently funded by the Government of Denmark represented by the Danish Ministry of Foreign Affairs, the European Union represented by the Directorate-General for International Partnership (INTPA), the U.S. Government represented by US Bureau of Population, Refugees, and Migration (PRM), the IKEA Foundation, and the Conrad N. Hilton Foundation. The JDC's mission is to improve the protection and well-being of those affected by forced displacement through evidence-informed humanitarian and development action and inclusive policies.

Competing interest. The authors declare none.

Author contribution. Conceptualization: P.M.B.; Writing original draft: P.M.B. All authors approved the final submitted draft.

References

Baal NK (2021) *Forced Displacement Data.* Available at https://www.unhcr.org/people-forced-to-flee-book/wp-content/uploads/sites/137/2021/10/Natalia-Krynsky-Baal_Forced-Displacement-Data-Critical-gaps-and-key-opportunities-in-the-context-of-the-Global-Compact-on-Refugees.pdf (accessed 27 March 2023).

Cervantes-Macías ME (2022) Migration data collection and management in a changing Latin American landscape. *Data & Policy 4*, e40. https://doi.org/10.1017/dap.2022.34

[42] https://data2.unhcr.org/fr/documents/details/103441

[43] https://unece.org/sites/default/files/2023-08/SDC2023_S4_4_WB_Solatorio_D.pdf

https://doi.org/10.1017/dap.2023.47 Published online by Cambridge University Press

Ekhator-Mobayode UE and Hoogeveen J (2022) Microdata collection and openness in the Middle East and North Africa. *Data & Policy 4*, e31. https://doi.org/10.1017/dap.2022.24

Gholami S, **Knippenberg E**, **Campbell J**, **Andriantsimba D**, **Kamle A**, **Parthasarathy P**, **Sankar R**, **Birge C and Lavista Ferres J** (2022) Food security analysis and forecasting: A machine learning case study in southern Malawi. *Data & Policy 4*, e33. https://doi.org/10.1017/dap.2022.25

Huynh BQ and Basu S (2020) Forecasting internally displaced population migration patterns in Syria and Yemen. *Disaster Medicine and Public Health Preparedness 14*(3), 302–307. https://doi.org/10.1017/dmp.2019.73

Leasure DR, **Kashyap R**, **Rampazzo F**, **Dooley CA**, **Elbers B**, **Bondarenko M**, **Verhagen M**, **Frey A**, **Yan J**, **Akimova ET**, **Fatehkia M**, **Trigwell R**, **Tatem AJ**, **Weber I**, **Mills MC** (2022) Nowcasting daily population displacement in Ukraine through social media advertising data. https://doi.org/10.31235/osf.io/6j9wq

Martini G, **Bracci A**, **Riches L**, … **Omodei E** (2022) Machine learning can guide food security efforts when primary data are not available. *Nature Food 3*(9), 716–728. https://doi.org/10.1038/s43016-022-00587-8

Masaki T, **Newhouse D**, **Silwal AR**, **Bedada A and Engstrom R** (2022) Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS 38*(3), 1035–1051. https://doi.org/10.3233/SJI-210902

Pham KH and Luengo-Oroz M (2023) Predictive modelling of movements of refugees and internally displaced people: Towards a computational framework. *Journal of Ethnic and Migration Studies 49*(2), 408–444. https://doi.org/10.1080/1369183X.2022.2100546

Qader SH, **Lefebvre V**, **Tatem AJ**, … **Bird T** (2020) Using gridded population and quadtree sampling units to support survey sample design in low-income settings. *International Journal of Health Geographics 19*(1), 1–16. https://doi.org/10.1186/s12942-020-00205-5.

Qader S, **Lefebvre V**, **Tatem A**, … **Bird T** (2021) Semi-automatic mapping of pre-census enumeration areas and population sampling frames. *Humanities and Social Sciences Communications 8*(1), 1–14. https://doi.org/10.1057/s41599-020-00670-0.

Solatorio AV and Dupriez O (2023) REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. Available at https://arxiv.org/abs/2302.02041v1.

Suleimenova D, **Bell D and Groen D** (2017) A generalized simulation development approach for predicting refugee destinations. *Scientific Reports 7*(1), 1–13. https://doi.org/10.1038/s41598-017-13828-9.

UNHCR (2023) *Mid-Year Trends 2023.* Available at https://www.unhcr.org/mid-year-trends-report-2023 (accessed 20 October 2023).

United Nations (2018) *Global Compact on Refugees.* Available at https://www.unhcr.org/media/global-compact-refugees-booklet (accessed 27 March 2023).

World Bank (2017) *Forcibly Displaced.* Available at https://openknowledge.worldbank.org/entities/publication/a4bdb82b-01e7-5e8f-8b75-6dc1591d9da1 (accessed 27 March 2023).

World Bank (2023) *World Development Report 2023: Migrants, Refugees, and Societies.* Available at https://www.worldbank.org/en/publication/wdr2023 (accessed 20 October 2023).

Cite this article: Brock PM and Mugera HK (2023). Accelerating and enhancing the generation of socioeconomic data to inform forced displacement policy and response. *Data & Policy*, 5: e42. doi:10.1017/dap.2023.47