Refugee Resettlement via Machine Learning and Integer Optimization

Andrew C. Trapp
Worcester Polytechnic Institute
Worcester, MA
atrapp@wpi.edu

Alexander Teytelboym
University of Oxford
Oxford, UK
alexander.teytelboym@economics.ox.ac.uk

Narges Ahani
Worcester Polytechnic Institute
Worcester, MA
nahani@wpi.edu

Tommy Andersson
Lund University
Lund, Sweden
tommy.andersson@nek.lu.se

ABSTRACT
Around 100,000 refugees are resettled to dozens of countries from conflict zones every year. While there is growing evidence that the initial placement of refugee families profoundly affects their lifetime outcomes, there have been few attempts to optimize resettlement destinations within host countries. We describe how machine learning and integer optimization can be used to empower resettlement agencies to drastically improve refugee employment outcomes. We describe possible future work on multi-objective optimization, the dynamics of allocation, and the inclusion of refugee preferences.

Keywords:
refugees, humanitarian operations research, machine learning, integer optimization, multiple multidimensional knapsack problem, matching

1 INTRODUCTION
In 2017, there were 18.5 million refugees—the highest number ever recorded—under the mandate of the Office of the United Nations High Commissioner for Refugees (UNHCR 2018). Of those, the UNHCR considers 1.2 million to be in need of resettlement—permanent relocation from their asylum country to a third country (UNHCR 2017). In 2016, the number of resettlement submissions reached 165,000 (a twenty-year high) and 125,800 people departed for resettlement (UNHCR 2017). Clearly, resettlement places offered by host countries are incredibly scarce. Currently, most refugees departing for resettlement are Syrians who are seeking asylum in Jordan and Lebanon, but there are also thousands of resettled refugees from the Democratic Republic of the Congo, Iraq, Somalia, and Myanmar. Refugees in need of resettlement are particularly vulnerable: a quarter are survivors of torture and a third face persecution in their country of origin (UNHCR 2017, Annex 3).

Dozens of countries resettle refugees.1 The United States (US) is by far the world’s largest destination of resettled refugees with 78,340 admitted in 2016.2 The US State Department delegates the resettlement process

---

1For refugee allocation mechanisms across countries, see Moraga and Rapoport (2014) and Jones and Teytelboym (2017a).
2In terms of per capita refugee resettlement, US is behind Canada, Norway, and Australia.
to nine resettlement agencies, known as Voluntary Agencies (VolAgs). The VolAgs are responsible for developing their own networks of “affiliates”—communities that welcome refugees and help them integrate into a new life in the US. Affiliates (which we refer to as localities) offer resettlement capacity voluntarily.

There is ample empirical evidence that the initial placement of refugees determines their lifetime outcomes (Áslund and Rooth 2007, Áslund and Fredriksson 2009, Áslund, Öst, and Zenou 2010, Áslund, Edin, Fredriksson, and Grönqvist 2011, Damm 2014, Feywerda and Gest 2016). Therefore, ensuring the optimality of the initial match between the refugee family and the community is an important challenge from social, economic, and humanitarian perspectives. However, there is currently little evidence that resettlement capacity offered by localities is being used to maximize either the welfare of refugees or of the host population.

In this paper, we describe how machine learning and integer optimization can improve the decision-making of the resettlement agencies and outcomes of refugees. Like Bansak, Ferwerda, Hainmueller, Dillon, Hangartner, Lawrence, and Weinstein (2018), we aim to maximize the employment outcomes (at three months after arrival) of resettled refugees. However, we incorporate an important additional complexity in resettlement. Localities face a number of constraints on the services they can offer to refugees, such as whether they can provide single-family support, whether they have appropriate translators, and whether they can adequately house and help large families (with six or more members). Some of these constraints are hard zero-one constraints while others might be quantity constraints (for example, the number of spots in employment training).

As refugees typically arrive in large families with children, they often require myriad services. Therefore, localities are often constrained in the kinds of refugee families they can reasonably host. Hence, optimizing refugee outcomes, while meeting refugees’ needs and respecting various hard constraints of the localities, turns out to be a complex operations research task. Here, we describe the basic framework of our empirical and optimization models. The design of the software, the preliminary results of the optimization, and the experiences of the practitioners are laid out in detail in Trapp, Teytelboym, Ahani, and Andersson (2018). We conclude with an outlook for future work in refugee resettlement optimization within the growing field of humanitarian operations research (Holguín-Veras, Pérez, Jaller, Van Wassenhove, and Aros-Vera 2013, Pedraza-Martinez and Van Wassenhove 2016).

2 MATCHING MODEL

The model of matching with multidimensional constraints is based on Delacrétaz, Kominers, and Teytelboym (2016).

There is a finite set of refugee families $\mathcal{F}$. A family $F \in \mathcal{F}$ is itself a finite set of size $|F|$. That is, a family $F$ has $|F|$ members with a typical member denoted $f \in F$. There is a finite set of localities $L$. Families are inseparable so all family members must be placed in the same locality. Refugee families require multiple units of different services (for example, employment training slots, language support, single-parent support etc.) from a set $S$. We denote by $v$ the matrix of family service needs, with typical element $v^f_s \in \mathbb{Z}_{\geq 0}$ denoting the total number of units of service $s$ required by family $F$. A refugee family $F$ can only be assigned to locality $\ell \in L$ if $\ell$ can provide services to meet all the needs of $F$. We denote by $\kappa$ the matrix of locality service capacities, with typical element $\kappa^\ell_s \in \mathbb{Z}_{\geq 0}$ denoting the number of units of service $s$ that locality $\ell$ can provide.

We summarize the estimated quality of each refugee-locality match with a single number called the quality score. Let $q : \mathcal{F} \times L \to \mathbb{R}_{\geq 0}$ denote the quality score function. Setting $q = 1$ (or any other constant)

---

3 These are: Church World Service (CWS), Ethiopian Community Development Council (ECDC), Episcopal Migration Ministries (EMM), Hebrew Immigrant Aid Society (HIAS), International Rescue Committee (IRC), Lutheran Immigration and Refugee Services (LIRS), US Committee for Refugees and Immigrants (USCRI), US Conference of Catholic Bishops (USCCB), and World Relief Corporation (WR).

4 While most constraints are infrastructure constraints, there might also be social or political constraints.

5 The theoretical set-up is unchanged if we assume that $v^f_s \kappa^\ell_s \in \mathbb{R}_{\geq 0}$. As an example, we might wish to capture that a refugee may require $\frac{1}{3}$ of a dialysis machine because she needs to use it once a week.
would maximize the total number of families that are resettled, while setting \( q(F, \ell) = |F| \) would maximize the total number of refugees that are resettled. We will be interested in the case where \( q \) represents the employment outcome of family \( F \) in locality \( \ell \) and can be estimated from data using observable locality and family characteristics.

Let us now state the problem of maximizing the overall observed quality of the match via integer optimization. We introduce a binary variable \( x(F, \ell) \) which is equal to 1 if family \( F \) has been matched to locality \( \ell \) and 0 otherwise. In order to maximize the outcome within the feasibility constraints, the social planner solves the following outcome optimization problem (OOP):

\[
\text{max } \sum_{F \in \mathcal{F}} \sum_{\ell \in \mathcal{L}} q(F, \ell) x(F, \ell) \quad \text{subject to: } \sum_{F \in \mathcal{F}} \nu_s^F x(F, \ell) \leq \kappa_s^\ell \quad \forall \ell, s \\
\sum_{\ell \in \mathcal{L}} x(F, \ell) \leq 1 \quad \forall F \\
x(F, \ell) \in \{0, 1\} \quad \forall F, \ell
\]

The OOP is an example of a 0–1 multiple multidimensional knapsack problem (Song, Zhang, and Fang 2008)—a combination of the 0–1 multiple knapsack problem (Martello and Toth 1980) and the 0–1 multidimensional knapsack problem (Fréville 2004). Clearly, the multiple multidimensional knapsack problem is NP-hard. Even so, instances of OOP involving tens of localities and hundreds of families run in just seconds on a modern laptop, even with open-source software.

Figure 1 illustrates an instance of the outcome optimization problem with five families, four localities, and two services (left and right). Here, family \( F_2 \) requires two units of the left service and one unit of the right service whereas locality \( \ell_1 \) provides four units of the left service and two units of the right service. Figure 2 presents a feasible outcome—which satisfies all the constraints—for this instance (a feasible outcome always exists).

3 PREDICTIVE MODEL

Our predictive model is based on recent work by Bansak, Ferwerda, Hainmueller, Dillon, Hangartner, Lawrence, and Weinstein (2018). Our empirical data come from a resettlement agency over the course of several years. The basic model is a (regularized) logistic regression. The dependent variable is a binary outcome \( q(f, \ell) \) of whether an adult refugee \( f \) has managed to secure employment within three months of arrival to the US in locality \( \ell \). The independent variables include: gender, English-speaking ability, age, presence of a severe medical condition, whether the refugee required urgent resettlement, family size,
education, as well as nationality dummies and time trends. We run a different logistic regression for each locality, dropping the localities with few observations. Feywerda and Gest (2016) argue that, since the assignment of refugees to locality is random conditional on observables, the model estimates should deliver the causal effect of initial placement in a locality on employment. Once the model has been estimated, we can use it to predict \( \hat{q}(f, \ell) \) out-of-sample. This gives us an estimate of the probability of employment for every adult refugee in every locality. We then set \( q(F, \ell) = \sum_{f \in F} \hat{q}(f, \ell) \) so the quality score for a family-locality pair is the sum of the expected employment probabilities of the adult members of the family. Our predictive model can explain heterogeneity in observed employment outcomes for different refugees within localities.

4 EXTENSIONS

There are at least three ways in which our approach can be extended. First, while localities agree on annual quotas, refugees arrive stochastically over the course of a year. Therefore, it is important to schedule the arrival of refugees. Andersson, Ehlers, and Martinello (2018) tackle this problem in the Swedish context. Second, we could include multiple objectives such as longer-term employment, income, health and education outcomes. Unfortunately, at the time of writing, no data on these objectives for resettled refugees arriving in the US appears to be systematically available. Third, we could include preferences of refugees and priorities of localities explicitly (Delacrétaz, Kominers, and Teytelboym 2016, Jones and Teytelboym 2017b, Aziz, Chen, Gaspers, and Sun 2017). Preferences could be collected during the refugee pre-arrival orientation using a questionnaire that elicits how refugees might trade off features of areas (for example, crime, climate, population density, amenities, quality of schools). However, including preferences while optimizing for a particular observable outcome can be a challenging problem (Andersson and Ehlers 2016, Biró and Gudmundsson 2018).

5 CONCLUSION

Refugee resettlement is a complex humanitarian problem which requires insights from a number of disciplines, including operations research, economics, statistics, political science, and sociology. Many new systems and approaches are urgently needed to improve the livelihoods of the resettled refugees and the communities into which they integrate. In this paper, we show how combining tools from machine learning and integer optimization can potentially improve refugees’ outcomes within a particular hosting country. In forthcoming work, we report on our initial attempt to implement the model described in this paper in practice (Trapp, Teytelboym, Ahani, and Andersson 2018).

ACKNOWLEDGMENTS

We are incredibly grateful to a number of dedicated practitioners for helping us implement this research. Emily Aiken provided invaluable software development assistance. Scott Duke Kominers provided valuable comments on the first draft. Andersson gratefully acknowledges the financial support of the Jan Wallander and Tom Hedelius Foundation (Research Grant P2016-0126:1) and the Ragnar Söderberg Foundation (E8/13). Teytelboym gratefully acknowledges the Skoll Centre for Social Entrepreneurship Research Accelerator Grant (“Implementing the Local Refugee Match”). This work was supported by the Economic and Social Research Council grant number ES/R007470/1. Trapp and Ahani are grateful for the support provided by the Foisie Business School at Worcester Polytechnic Institute.

REFERENCES


**AUTHOR BIOGRAPHIES**

**ANDREW C. TRAPP** is Associate Professor of Operations and Industrial Engineering at Worcester Polytechnic Institute (WPI), and holds courtesy professorships in Mathematical Sciences and Data Science.
His research focuses on using analytical techniques, in particular mathematical optimization, to identify optimal decisions to problems arising from a diverse cross-section of sectors such as humanitarian operations, sustainability, healthcare, and data mining. Andy develops new theory, models, and computational solution approaches to tackle such problems. http://users.wpi.edu/~atrapp/

ALEXANDER TEYTELBOYM is Associate Professor of Economics at the University of Oxford. He is also a Tutorial Fellow at St. Catherine’s College and Deputy Director of the Economics of Sustainability Programme at the Institute for New Economic Thinking at the Oxford Martin School, University of Oxford. http://www.t8el.com

NARGES AHANI is a doctoral student at Worcester Polytechnic Institute in Data Science program. She received a Bachelor’s degree in Electrical Engineering at Shahroud University of Technology (2001—2005). She has ten years of experience in software development, application architecture, and hardware programming and design. She is completing an MS in Operational Analytics and Management at Foisie School of Business at WPI (2017—2018).

TOMMY ANDERSSON is Professor in Microeconomics at the Department of Economics, Lund University, and an External Fellow at the Center for Mechanism and Institution Design at University of York. He is a Board Member of the Society for Economic Design (2014—present), an elected Board Member of the Society for Social Choice and Welfare (2016—2022), and a member of the Scandiatransplant Steering Committee for the Scandiatransplant Paired Kidney Exchange Program. https://sites.google.com/site/tommyanderssonlunduniversity/home