Bocconi

# Applications of Data & Machine Learning in Economic Research: Part II – Refugee Allocation

*BAI 30545 – Foundations of Economic Sciences*

**Julian Streyczek** (Bocconi)

Previous session

# Survey results

Thanks to everyone who participated!

Results: About the **same pace,** slightly **more detail**

**Pace:** In the next lessons, I would prefer the pace to be:

*(Maximum 1 choices per person)*

| Answer | Choices | Average |
|---|---|---|
| Much faster | 1/23 | 4% |
| Slightly faster | 6/23 | 26% |
| About the same | 13/23 | 57% |
| Slightly slower | 3/23 | 13% |
| Much slower | 0/23 | 0% |

**Detail:** In the next lessons, I would prefer the amount of detail to be:

*(Maximum 1 choices per person)*

| Answer | Choices | Average |
|---|---|---|
| Much more | 5/23 | 22% |
| Slightly more | 11/23 | 48% |
| About the same | 5/23 | 22% |
| Slightly less | 2/23 | 9% |
| Much less | 0/23 | 0% |

# Introduction

# Research Paper: Bansak et al. (2018, Science)

## Improving refugee integration through data-driven algorithmic assignment

Kirk Bansak,[1,2]* Jeremy Ferwerda,[2,3]* Jens Hainmueller,[1,2,4]*† Andrea Dillon,[2]
Dominik Hangartner,[2,5,6] Duncan Lawrence,[2] Jeremy Weinstein[1,2]

Kirk Bansak et al.: *"Improving refugee integration through data-driven algorithmic assignment"*. Science **359**, 325-329 (2018)

# Why we care about allocation of refugees

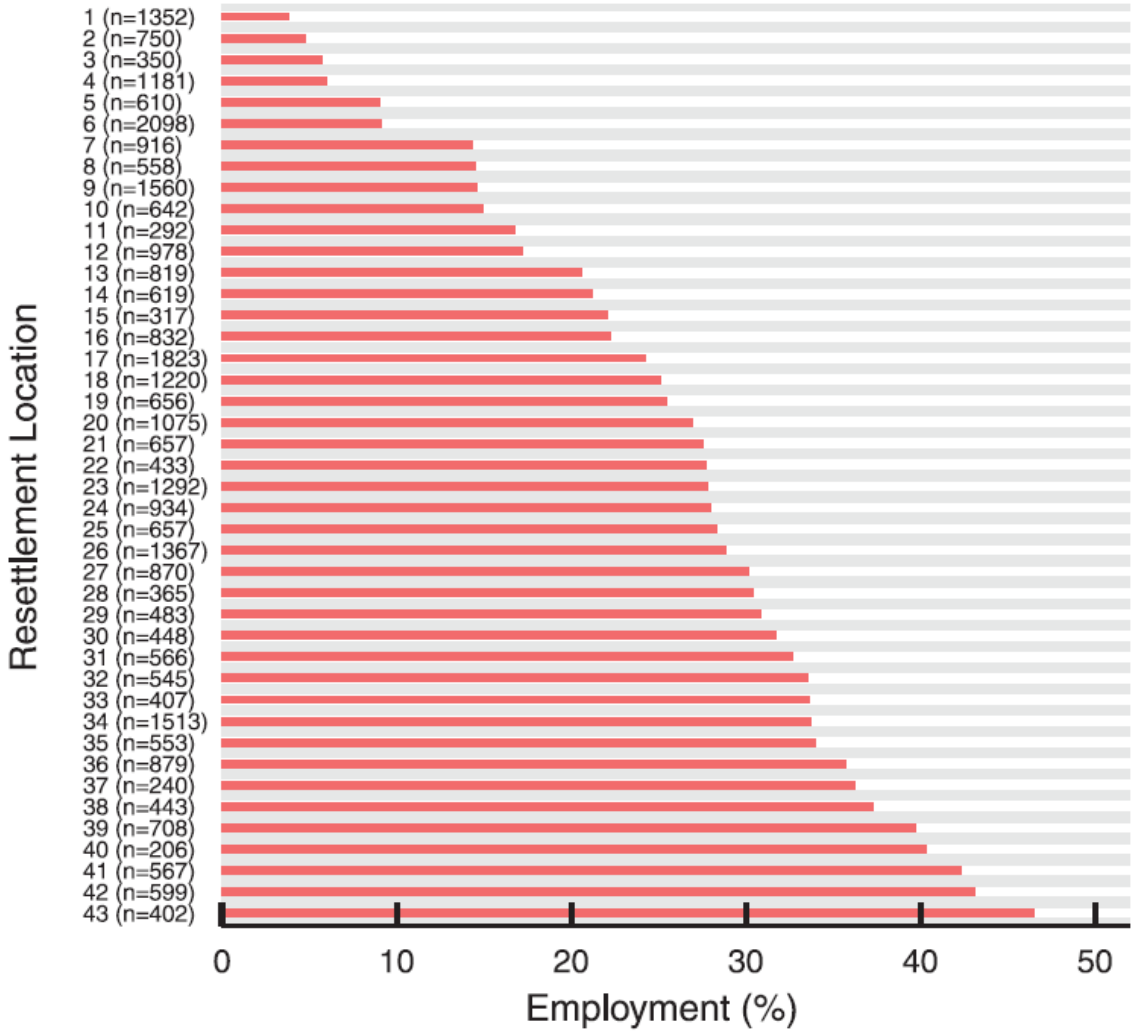
- Refugees are among the world's **most vulnerable populations:**
  - Endured **hardship**, **displacement**, **loss** of resources
  - Face **unfamiliar environment** and limited support networks
- **Employment** is key for successful **integration**:
  - **Contact** with local population
  - **Self-sufficiency** and contribution to society
- **Argument: Allocation** of refugees to resettlement locations is one of the **first** and **most consequential policy decisions**
  - Match quality matters: e.g., place French speakers into French-speaking locations
  - Long-term consequences

# Terminology: Refugee vs. Asylum Seeker

- **Definitions:**
  - **Refugee:** Granted official status before arrival in host country
  - **Asylum seeker**: Applies for protection after arrival in host country, remains in country until request is processed
- **Important implication:** Refugees usually permitted to work immediately, asylum seekers often not
- In this paper, two datasets:
  - **US:** Only **refugees**
  - **Switzerland:** Both **refugees** and **asylum seekers**, both allowed to work early after arrival
- Will use term „refugee“ for both groups throughout

# Employment results differ much by resettlement location (US)

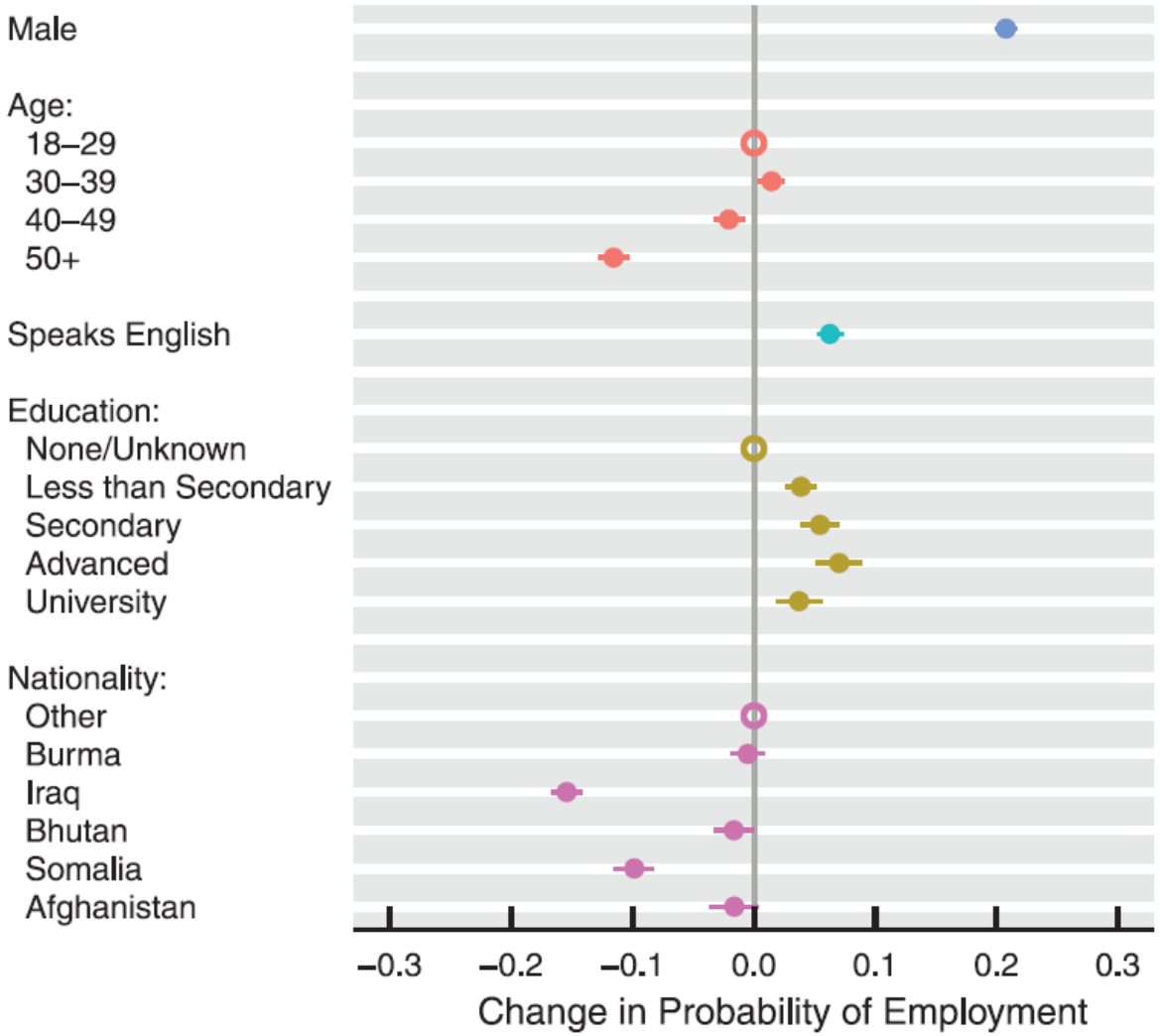
- **Figure:** Share of employed refugees 90 days after arrival, separately for 43 resettlement locations in the US
  (number of observations in brackets)
- **Result:** Some locations seem **more capable** at placing refugees into employment
- **Implication:** Might want to allocate more refugees to locations with high employment shares
  => Room for improvement!



Introduction

# Employment results differ much by refugee characteristics (US)

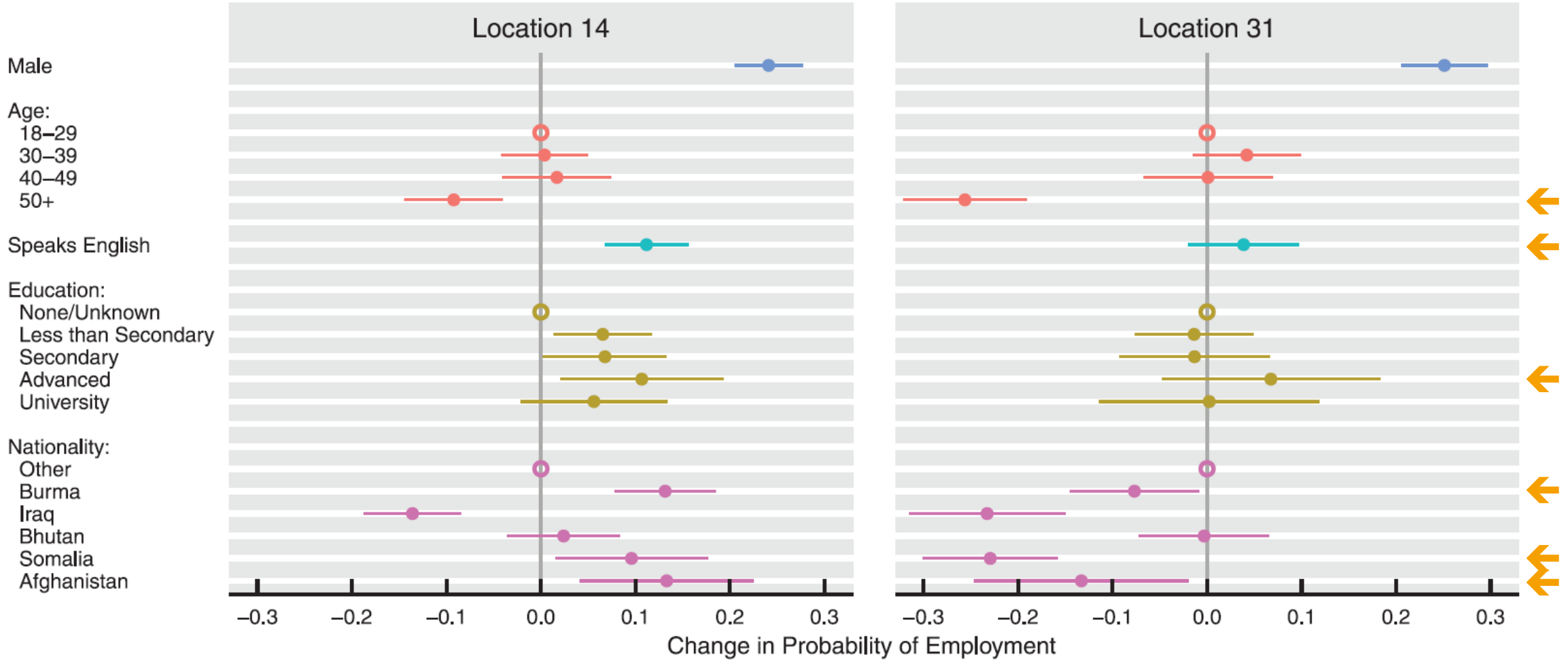
- **Figure:** Correlation of each refugee characteristic with probability of being employed 90 days after arrival
  (Ordinary Least Squares)
  - **>0:** Characteristic *increases* chance of employment*
  - **<0:** Characteristic *decreases* chance of employment*
  - **Horizontal lines:** Indicate *uncertainty* – Points are an *estimate* for the correlation, *lines* depict the range in which we expect the true value
- **Result:** Refugee characteristics **correlate highly** with probability of finding employment:
  - **Male:** +20 percentage points (pp.)
  - **Speaks English:** +8 pp.
  - **Advanced Degree:** +8 pp.




* **Note:** Relative to "omitted" category: Male vs. Female, Education level vs. "None/Unknown"

Introduction

# Importances of refugee characteristics change by location (US)



Introduction

# Importances of refugee characteristics change by location (US)

**Results:**

- **Age:** 50+ find employment more easily in Location 14
- **Speaking English:** Does not matter in Location 31
- **Education:** Does not matter in Location 31
- **Nationality:** Matters much in both locations

**Implications:** (assuming that predictions are causal…)

- If we place a **50+ years-old** into Location 31 instead of 14,
  we **increase probability of employment by ~15pp!**
- If we place an **English-speaking Somali** into Location 14 instead of 31,
  we **increase probability of employment by ~ (10+30) = 40pp!**

# This paper

- **Idea:** Data-driven approach to allocating refugees within countries to optimize employment
  - **How:** Allocate refugees so that total employment is maximized
- 3 Stages:
  1. Modeling
     - **Train model: Predict employment probability** for **refugee characteristics** and **locations**: One model for each destination, reveals which characteristics are important in which location
  2. Mapping
     - **Transform predictions** from **refugee-level** to „**case**"-**level** (family-level), as refugee families are allocated together
  3. Matching
     - **Find optimal allocation:** Use **numerical solving procedure** to maximize **total employment** over all refugees and locations, given constraints (location capacities)

# Background and Data

Background

# Allocation of refugees in the US and Switzerland

## USA:

- Department of State randomly assigns refugee cases to one of 9 agencies, who place them within their network
- **70%** of cases have fixed location: Need to be placed as close as possible to existing family/sponsor/etc.
- **30%** of cases are „free“ to be allocated anywhere: Usually place into location with largest available capacity

## Switzerland:

- Secretariat for Migration assigns refugees to one of 26 cantons (states)
- **5%** existing ties: Fixed location
- **95%** „free“: Assigned proportionally to local capacity, *blind* to refugee characteristics

=> For “free” cases, assigned location is random with respect to characteristics!

# Importance of random allocation

- **Important:** For refugees who were assigned to locations randomly, **correlations** may (cautiously) be interpreted as **causal**
  - **Why:** Refugee characteristics are unrelated to employment probabilities of locations

- What happens without randomization? Example:
  - Assume:
    - No causal effect of speaking English on employment probability
    - Send English-speaking individuals to locations with overall high employment probability
  - We would measure a positive correlation between speaking English and employment, although there is no causal effect *("Selection Bias")*

Data

# Data on refugees, allocations, and employment

## USA:

- Complete data for 1 out of 9 resettlement agencies
- Only refugees, working-age (18-64 years)
- Training sample:
  - Q1/2011 – Q2/2016
  - 33,782 individuals from 22,144 cases (fixed + free cases)
- Testing sample:
  - Q3/2016
  - 919 refugees from free cases

## Switzerland:

- Complete data from Secretariat for Migration
- Refugees and asylum seekers (18-65 years)
- Training sample:
  - 1999 – 2012
  - 22,159 individuals (mostly free cases)
- Testing sample:
  - 2013
  - 888 refugees from free cases

Data

# Variables

| Variable | Levels | US | Switzerland |
|---|---|---|---|
| Male | Yes/No | ✓ | ✓ |
| Age | 18-29, 30-39, 40-49, 50+ | ✓ | ✓ |
| Education | University, High School, etc. *[5]* | ✓ | ✓ |
| Country of origin | *[many]* | ✓ (6) | ✓ (24) |
| Speaks English | Yes/No | ✓ | × |
| Speaks French | Yes/No | × | ✓ |
| Marital status | Single, Married, Widowed | × | ✓ |
| Christian | Yes/No | × | ✓ |
| Year/Month of arrival | *[many]* | ✓ | ✓ |
| Free case | Yes/No | ✓ | ✓ |
| Assigned location | *[many]* | ✓ (43) | ✓ (20) |
| Employment after 90 days | Yes/No | ✓ | × |
| Employment after 3 years | Yes/No | × | ✓ |

# Empirical Strategy: Modeling – Mapping - Matching

1) Modeling

# Predict employment for location + refugee features

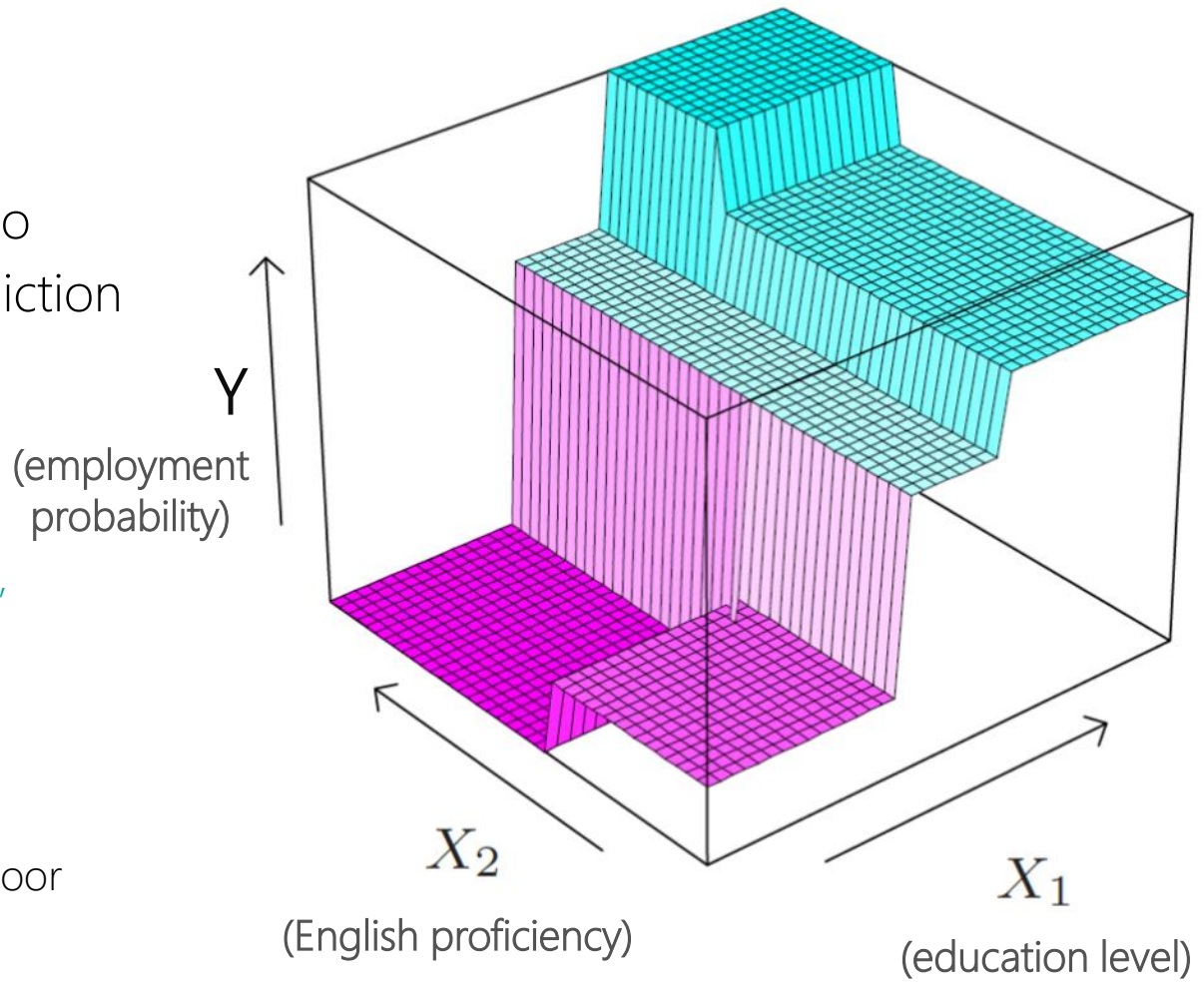
- Goal: Predict employment probability for refugee characteristics and locations
- Idea: Train one model for each location:
  Captures different relevance of refugee characteristics in different locations
  (e.g., speaking French only improves employment in French-speaking cantons)
- Strategy: For each location $\ell = 1,2, \dots, L$:
  - Select sample: Individuals in training sample that were *actually* assigned to location $\ell$
  - Train model using training data: Get predictions for which features predict employment
  - Predict employment for testing data: For *all* individuals in testing set, predict integration outcome in L
- Result: Integration prediction for each individual in testing data and location
  - Example - US: We predict for each of the 919 individuals the employment probability for each of the 43 locations (had they been allocated here)
    => 919 x 43 = 39,517 employment probabilities

1) Modeling

# How do we get predictions? (1/2)

- Estimator: „Decision tree“
- Idea: Create a **“step-function”**:
  Group combinations of **features** (X1, X2, …) into “regions”, and for each region make one prediction (best guess) for the **outcome** (Y)
- Example on the right:
  - X1 is **education level**:
    Pink = “Low”, Light cyan = “Medium”, Cyan = “High”
  - X2 is **English proficiency:**
    Low, Medium, High (not colored separately here)
  - Y is **predicted employment probability**

  ⇒ Low employment probability for (low education + poor English), high for (high education + good English)
- Out-of-the box algorithms to predict tree from data



# How do we get predictions? (2/2)

- **Actual estimator:** *Stochastic gradient-boosted decision trees* (R-package: *gbm*)
  (aka gradient-boosted random forest)
  - **Decision tree:** Grow a tree that predicts outcomes using features
  - **Gradient:** Try to find the "direction" to update the tree so that the prediction improves
  - **Stochastic:** Subsample from observations and build a new tree using the gradient above
  - **Boosted:** Update the initial tree: use an „average" of the initial and the new tree
  - … repeat a few times …
- **Additional challenge:** Need to select some *hyper-parameters*
  (number of trees (repetitions), tree depth, learning rate (updating)):
- **Solution:** *Cross-validation*
  Repeatedly train on subsets of data and evaluate performance on unseen ("out-of-sample") data.
  For final model, choose hyper-parameters that produce best out-of-sample performance

2) Mapping

# Transform individual-level to case-level predictions

- **Remember:** We have predictions for employment probability of each **individual** in each **location**
- **Problem:** Allocations to locations is not on **individual-level**, but on **case-level** (family-level)
  => **Have:** Employment probabilities for **individual-by-location**
  => **Need:** Employment probabilities for **case-by-location**
- **Solution:** Aggregate:
  **For each case, compute the probability that at least one individual in case is employed**
  (The process of assigning a new value to each existing value is called "mapping" in math/computer science)
- **Intuition**: Indicator for „self-sufficient refugee family"
- **Alternative mappings:**
  Minimum/Maximum/Average probability of employment among individuals in case

3) Matching

# Compute optimal allocation of refugees to locations

- **Target:** Total predicted employment (= sum of predicted employment probability over all cases)
- **Optimal matching:** Allocation of cases to locations such that:
  1. Each case is allocated to exactly one location
  2. Each location has no more cases than its capacity allows
  3. There is no other allocation that satisfies 1. and 2. that has higher total predicted employment
- This is a **constrained optimization problem:** Mathematicians have created out-of-the-box computer algorithms to solve it (here: *RELAX-IV* from R-package *optmatch*)
- **Result:** For each case in test set:
  1. Optimal location
  2. Employment probability in optimal location
- **Remember:** We can compare to the *actual location* and *actual employment*

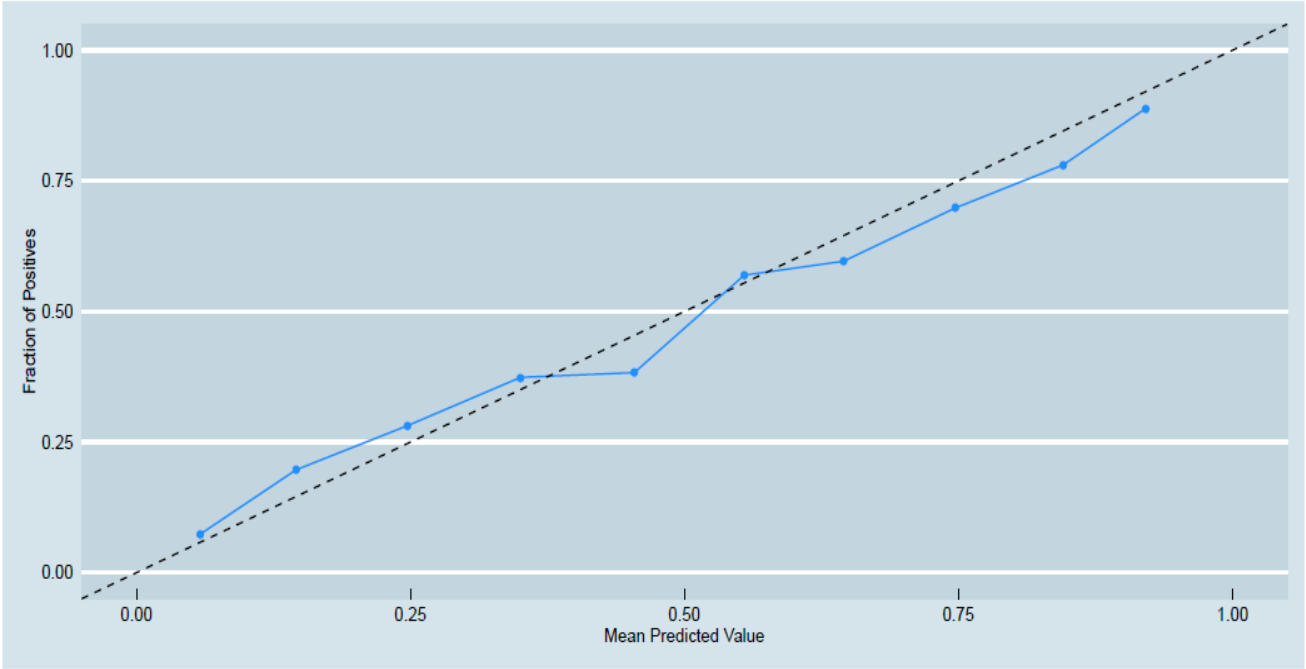
# Results

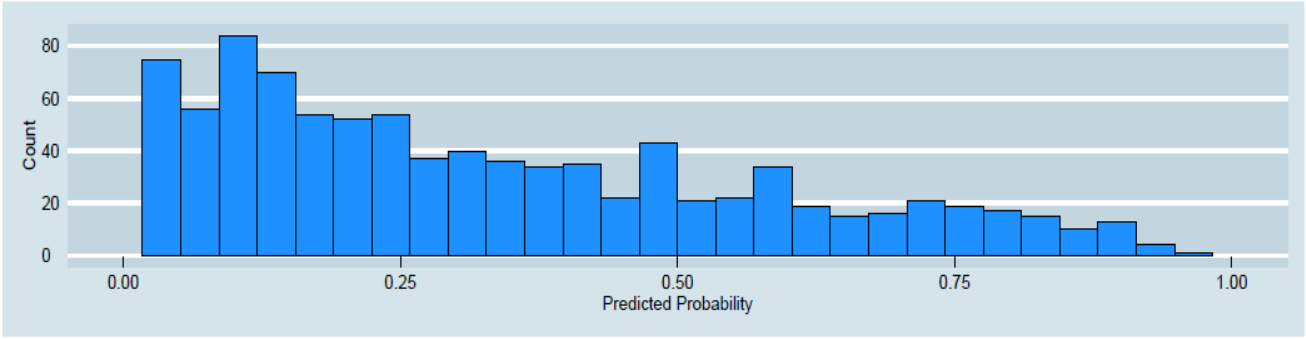
Results

# Model fit: US

- **"Backtest":** For individuals in test set, compare *actual* to *predicted* employment probabilities (in actual location)
- **Upper:**
  1. Sort test set individuals by *predicted* employment probability in their *actual* location
  2. Group by predicted probability: [0-10], (10,20], …, (90,100]
  3. For each group, plot actual employment share
  4. Dotted line: Best possible model

  => Model seems to fit decently
- **Lower:** Number of individuals in 30 groups of predicted employment probability



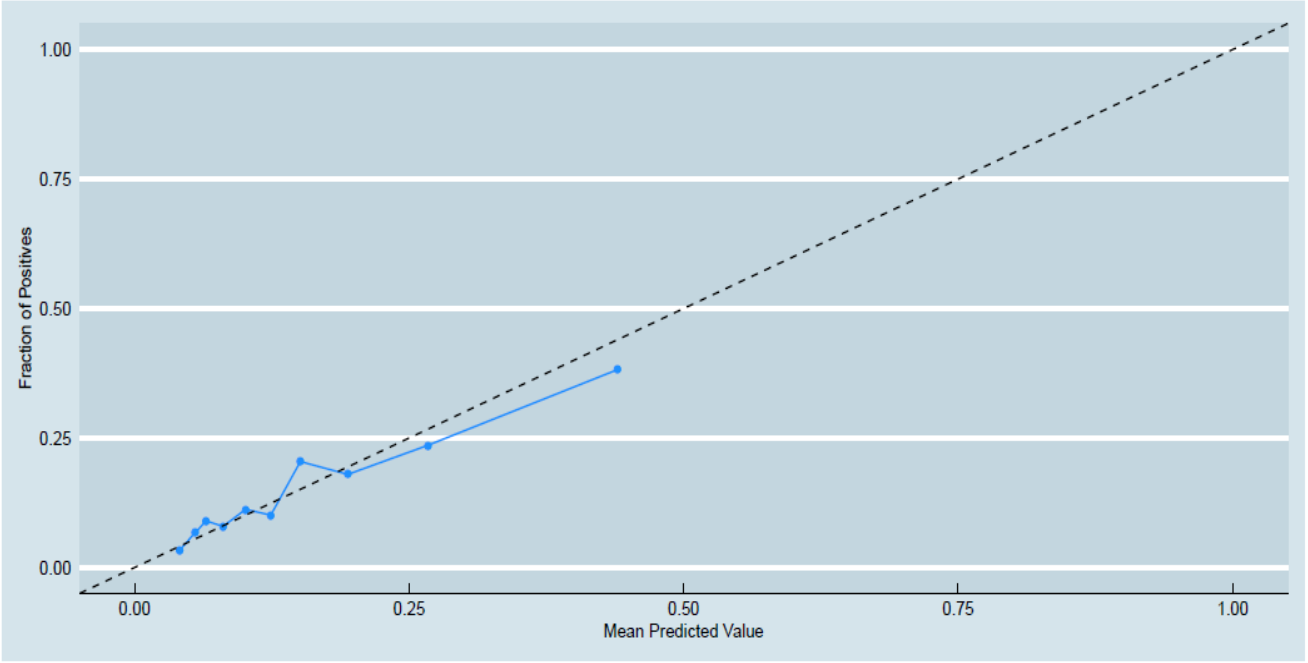



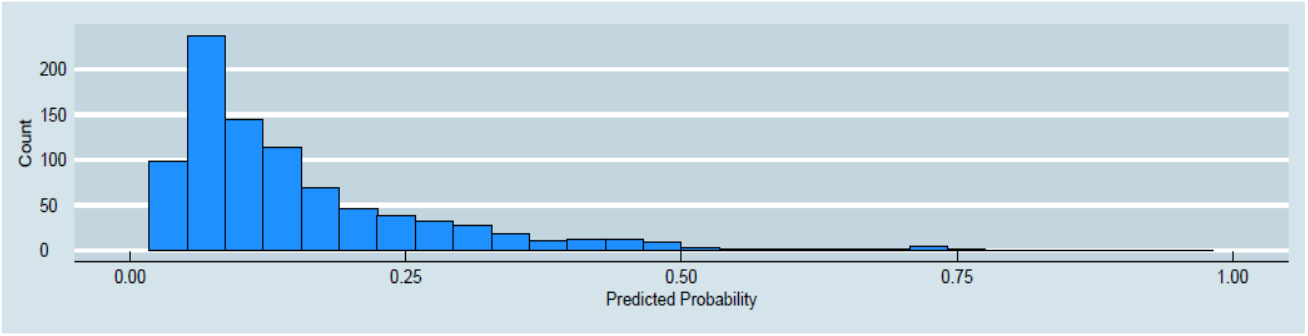
Results

# Model fit: Switzerland

- **"Backtest":** For individuals in test set, compare *actual* to *predicted* employment probabilities (in actual location)
- **Upper:**
  1. Sort test set individuals by *predicted* employment probability in their *actual* location
  2. Group by predicted probability: [0-10], (10,20], …, (90,100]
  3. For each group, plot actual employment share
  4. Dotted line: Best possible model

  => Model seems to fit decently
- **Lower:** Number of individuals in 30 groups of predicted employment probability



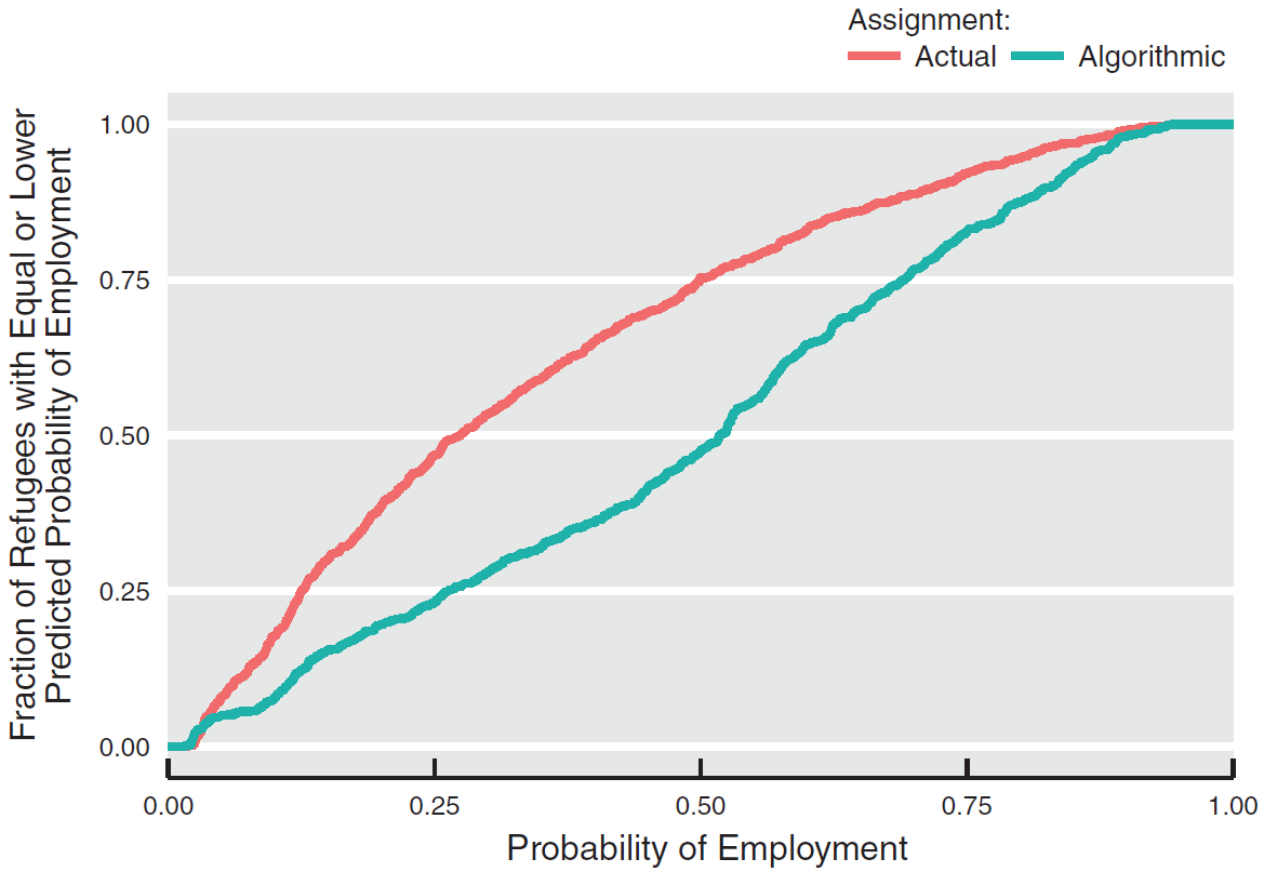



Results

# Employment gains from optimal allocation: US

- **Right:** *Empirical cumulative distribution function (CDF)* of employment, separately for actual and optimal allocation of refugees
  1. Sort individuals by predicted employment probability, for actual (red) and optimally assigned (cyan)
  2. For each probability value, plot share of individuals with *lower* employment probability

  => Lower = Better
- **Result:**
  - Actual: 34% employment
  - Optimized: 48% employment => +41%*
- **Interpretation: Optimal allocation improves employment by 41%**



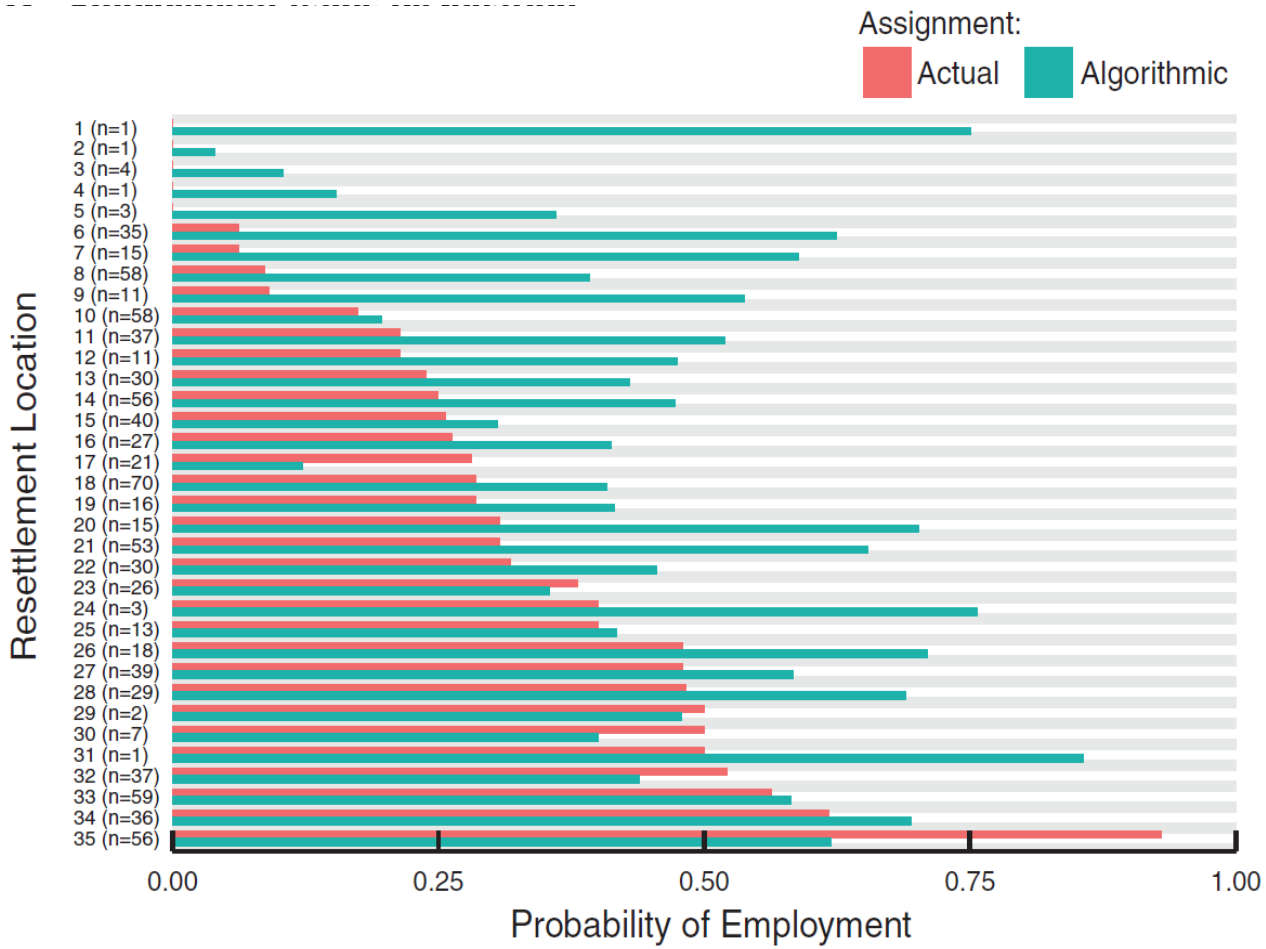

* Note: Difference between *percent* (relative to some baseline) and *percentage points* (absolute)

Results

# Employment gains from optimal allocation: US

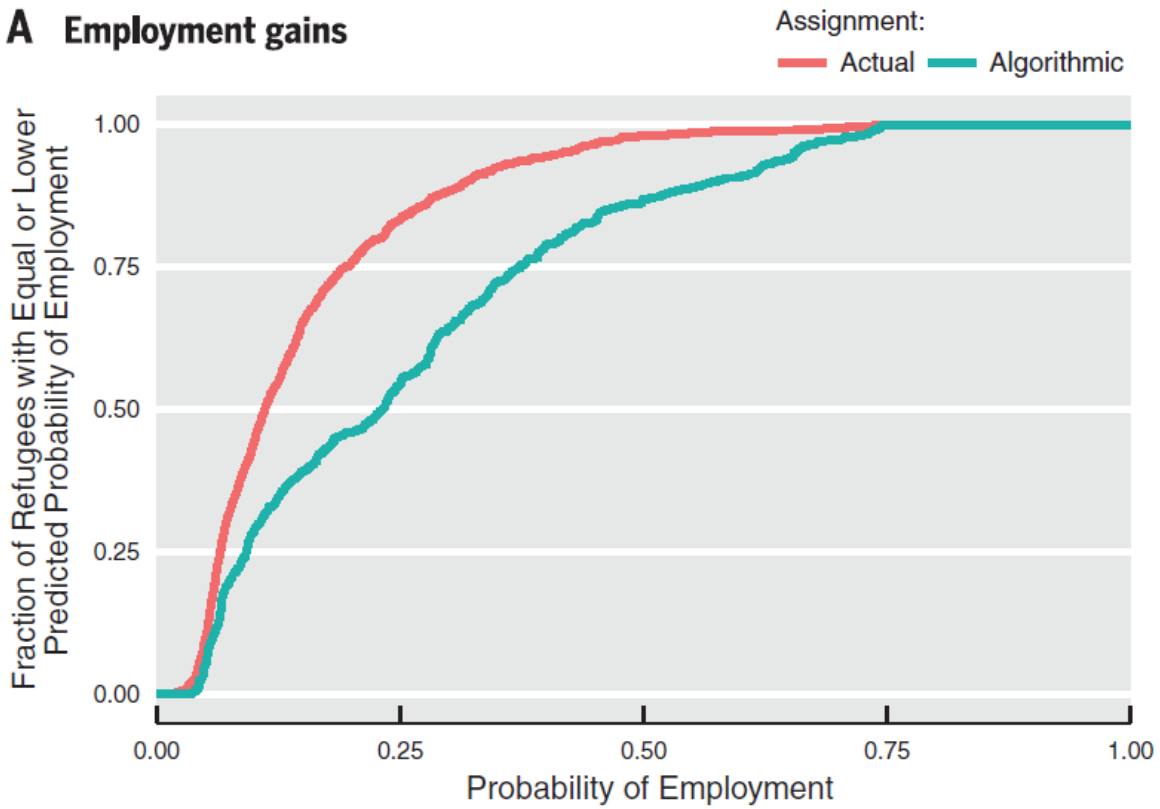
- **Right:** For each location, plot:
  1. Employment shares under actual allocation
  2. Predicted employment shares under optimal allocation

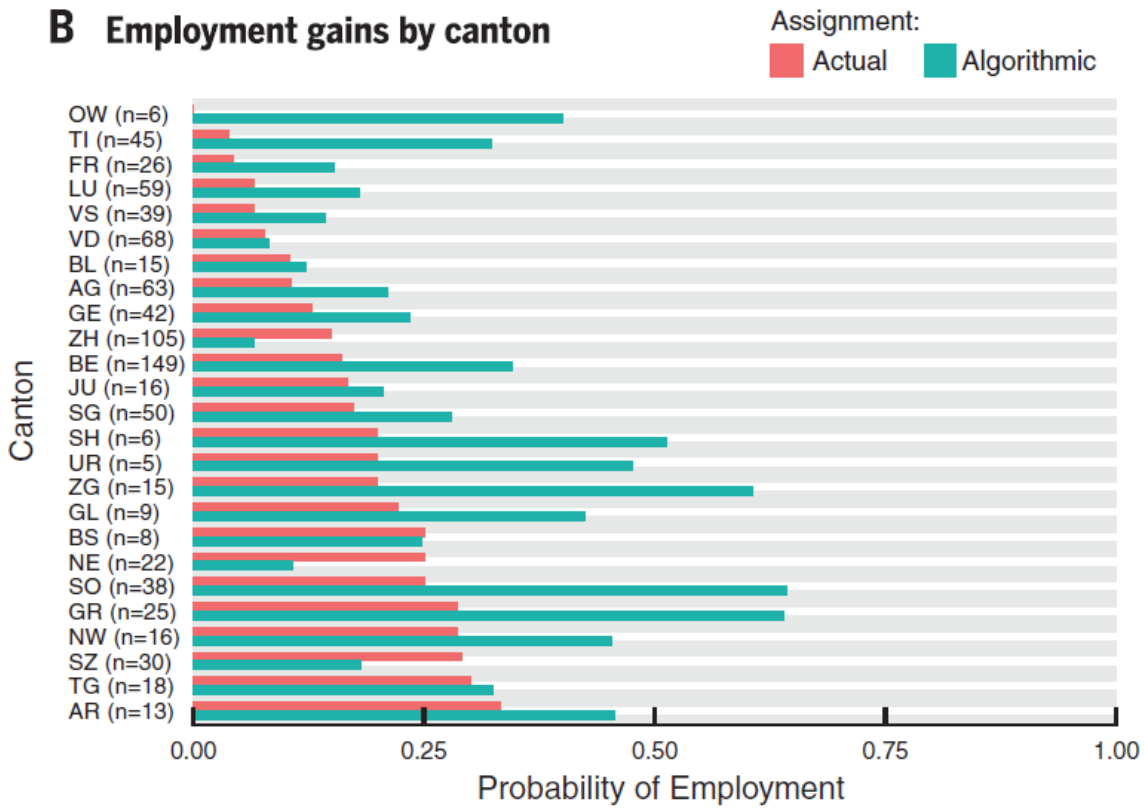
  => Higher = Better
- **Result:** Increases in employment in virtually every location
- **Implication:** (Almost) no location "loses"



Results

# Employment gains from optimal allocation: Switzerland




- Optimal allocation increases employment from 15% to 26%: **+73%**

Results

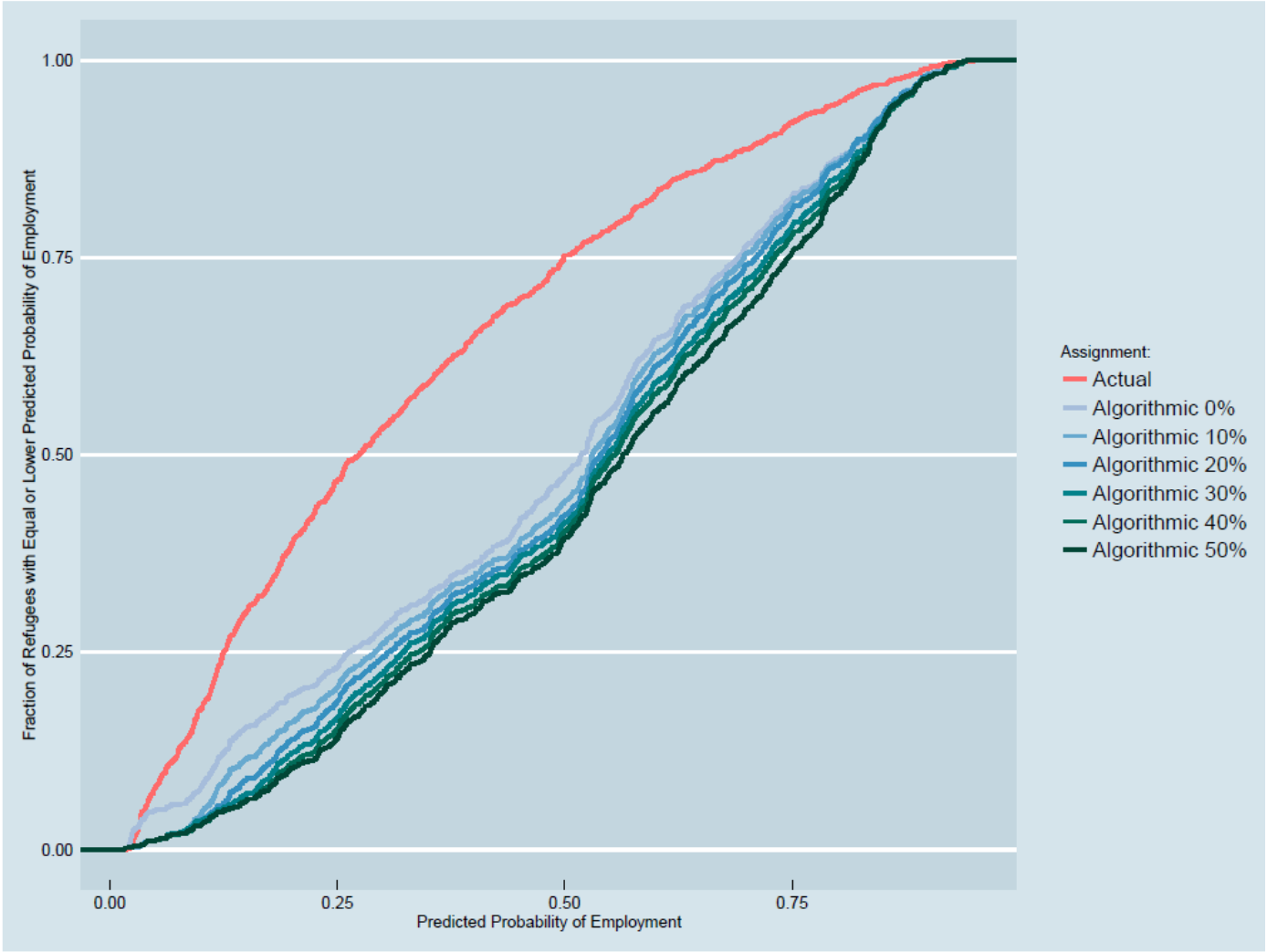
# Sensitivity checks

- Different time periods for testing:
  - US: Instead of Q3/2016, use Q2/2016, Q1/2016, Q4/2015
  - CH: Instead of 2013, use 2012, 2011, 2010 **Details**
- Different time periods for training:
  - US: Instead of 5.5 years, use 4.5, 3.5, 2.5
- Alternative mapping schemes (from individual to case):
  - US: Instead of probability of at least one employed, use minimum / maximum / median probability of employment
- Shorter- and longer-term outcomes
  - CH: Instead of employment after 3 years, check also after 2 and 4 **Details**

Results

# Relaxing location-specific capacity constraints (US)

- **Simulation experiment:** Optimize allocation as before, but allow 10%, 20%, … more capacity per location (more refugees can be placed in each)
- **Result:** Even larger increases in employment
- **Intuition:** Want to increase capacity in regions that are better at employing refugees
- **Implication:** Can predict which locations to expand
  => Policy implication!



# Discussion

Discussion

# Open questions

- **Drivers:** Which types of matches improve employment outcomes the most?
  - Nationality, language, etc.?
- **Distribution of gains:**
  - Do some types of refugees benefit more from optimal allocation than others?
  - Do any types lose?

Discussion

# Advantages

- **Cheap:** Data driven
- **Easily implementable:** Integrate with existing allocation process
- **Low effort:** No additional costs for refugees (allocation determined before arrival, no extra tasks)
- **Adaptable:** Easy to change integration success metrics (e.g., hours worked) and optimality criteria (e.g., average employment)
- **Complementary to existing knowledge:** May use as recommendation for allocation officers (combine expert knowledge + data-driven recommendation)
- **Policy implication:** Identify in which locations to increase capacity

Discussion

# Limitations

- **No testing** of findings, just simulations
- **Causality I: How random is allocation of refugees, actually?**
  - Is it credible that officers do not match on expected employment? Little empirical evidence.
- **Causality II: Training and testing** on **different populations** (all cases vs. free cases)
  - **Potential problem:** Training data includes individuals placed into **existing network** => easier to find work?
- **Causality III:** Predictions ignore **"general equilibrium effects":** Hold everything else fixed ("ceteris paribus")
  - **Potential problem, A:** The more French speakers we place into French-speaking cantons, the higher the competition for jobs among French speakers, leading to lower (employment) payoff from speaking French
  - **Potential problem, B:** As pre-existing nationality networks grow, two opposing effects (not yet modeled):
    - Easier to find job
    - Lower incentives to integrate into local society

# Appendix

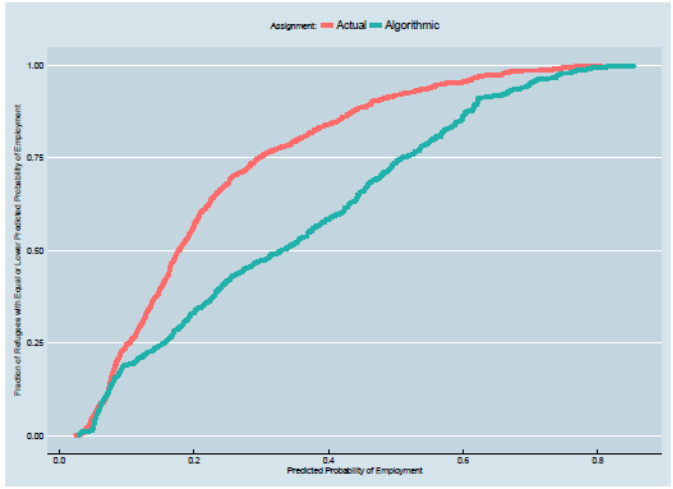
# Modeling: Notes

- Note: Train using *all* cases, test only for *free* cases
  - Train using all cases:
    - Pro: More data => more precise estimates
    - Con: different types => potential bias
    - Here: Include all, but add „free case“ indicator to capture differences
  - Test for free cases: Only group for which allocation can be optimized (for others, location is pre-determined)

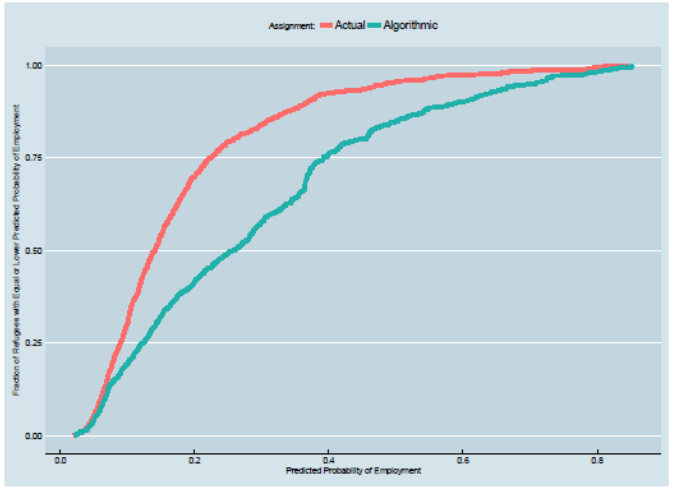
Appendix

# Refugees that arrived in different years (CH)

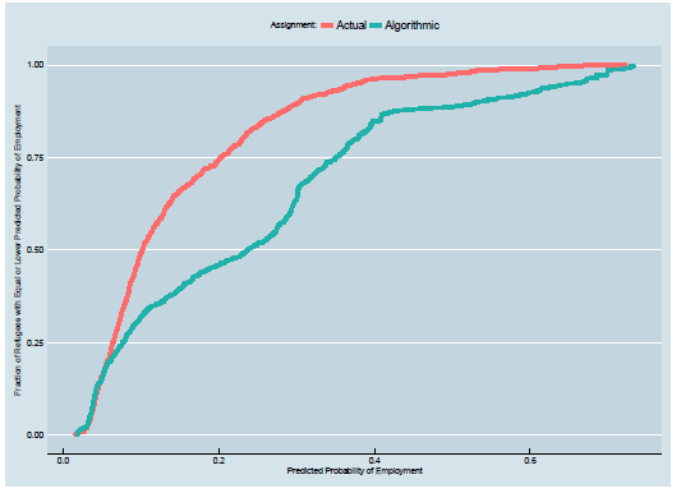
(a) Employment gains for 2010 arrivals



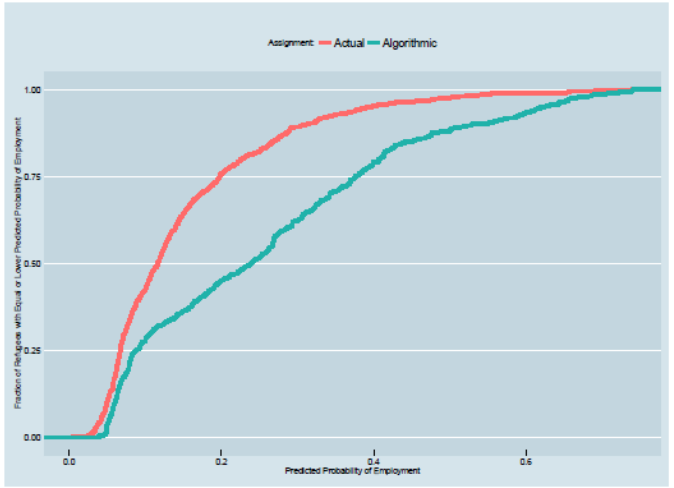

(b) Employment gains for 2011 arrivals




(c) Employment gains for 2012 arrivals




(d) Employment gains for 2013 arrivals

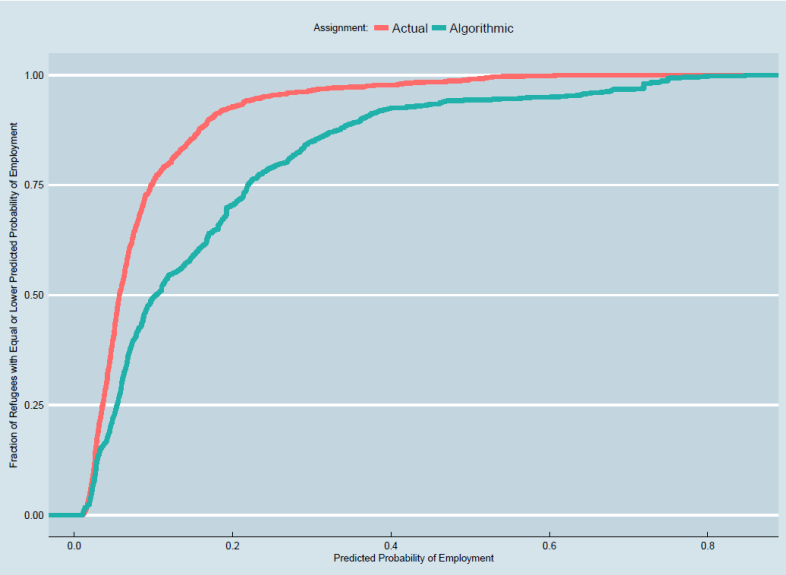



Back

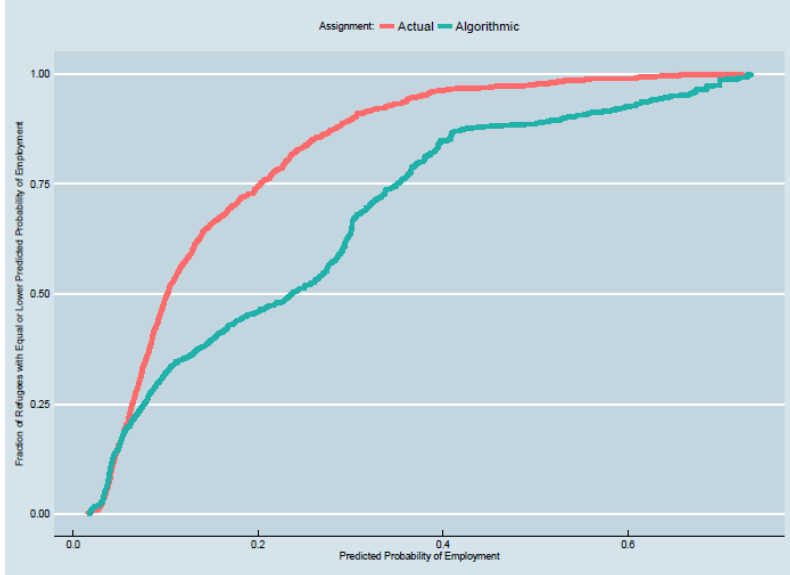
Appendix

# Employment outcomes after 2, 3, 4 years (CH)

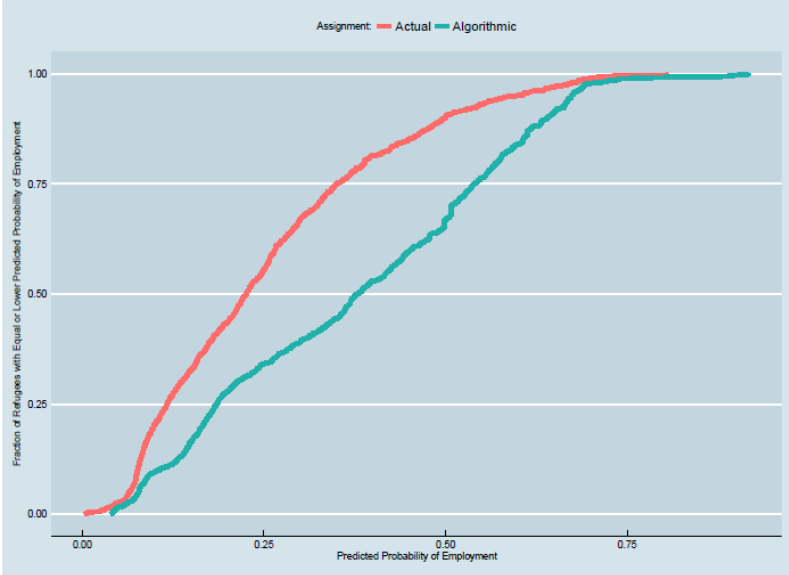
(a) Gains for second-year employment




(b) Gains for third-year employment




(c) Gains for fourth-year employment




Back