Using machine learning to understand the earnings effects of import competition

Johanna Muffert Erwin Winkler

FAU Erlangen-Nuremberg

EEA, August 2023

What explains import-induced earnings losses?

Workers in import-competing industries: displacement & **earnings losses**

- e.g., Autor et al. (2014), Utar (2018), Dauth et al. (2021), Nilsson Hakkala & Huttunen (2018)

- Suggests existence of frictions to moving between jobs/industries
  1. imperfect transferrability of **specific human capital**?
  2. losses in **rents/wage premia** earned at initial employer?

- **Relative importance of channels unclear, but policy-relevant**
  - e.g., retraining of displaced workers more effective under 1.

This paper

Combine machine learning (causal forest) with IV from existing literature to study import-induced earnings losses

- Focus on import competition from China and Eastern Europe on German manufacturing industries
- Idea: Carefully analyze heterogeneity in effect to learn about underlying mechanisms (e.g., Smith 2022)
- **First paper to provide 'horse race' between competing channels**

Why machine learning?

Usual approach: evidence on <u>one</u> channel by sample splitting, e.g.:

- workers in high vs. low-rent firms (Dauth et al. 2021)
- workers with high vs. low specific human capital (Utar 2018)

- Issue (1): **Multiple hypothesis testing** if many channels are tested
- Issue (2): Estimates might be **'wrong-signed'** if sub-samples differ in many relevant characteristics
  - e.g., workers in high-rent firms might have less specific human capital
- Issue (3): Need to make ex-ante choices about **functional form**

$\Rightarrow$ **Use machine learning to circumvent these issues**

Preliminary findings

- **Specific human capital** and **losses in rents** play important role, to a similar extent
- **Conventional interaction effects yield misleading results**
  - Would favor rents losses over specific human capital as main channel of earnings losses
  - Some interaction effects are ’wrong-signed’ (e.g., age)

Data

- Sample of Integrated Labour Market Biographies (SIAB)
  - Full-time employed workers in manufacturing in the **base year**
  - Age 24-65 during observation period
  - Following workers over a 10-year period
- UN Comtrade Database
  - Bilateral trade data at 3-digit industry level

Empirical strategy (Autor et al. 2014)

Idea: Compare observationally identical workers who are differently exposed to imports due to different initial industry affiliation

$$CumulEarnings_{ikt} = \beta NetImp_{kt} + X'_{ikt}\gamma + \epsilon_{ikt}$$

- $CumulEarnings_{ikt}$: Cumulative earnings over 10 years relative to base year earnings of worker i, employed in industry k in base year t (1990 or 2000)
- $NetImp_{kt}$: 10-year-change in net import exposure on industry k
  - $\frac{\Delta Imports_{kt} - \Delta Exports_{kt}}{WageSum_{k(t-1)}}$
- X'_{ikt}: worker, plant, industry, and region controls (t)

Instrument (Autor et al. 2014, Dauth et al. 2014)

- Want $NetImp_{kt}$ to reflect increased import competition on domestic workers
- Instrument: industry-level increase in net import exposure in other countries (following Autor et al. 2013, 2014, Dauth et al. 2014, 2021)
  - Isolate increase in import competition driven by rise in productivity in China and EE
  - Instrument countries: Australia, Singapore, Japan, Norway, Sweden, Canada, UK

Baseline results

Table: Baseline estimates

Dep. Var.: 100 x Normalized cumulative earnings		
	(1)	(2)
	OLS	IV
NetImp	-0.169***	-0.194*
	(0.051)	(0.117)
Obs.	159,288	159,288
F-Stat. of excl. instrument		14.5

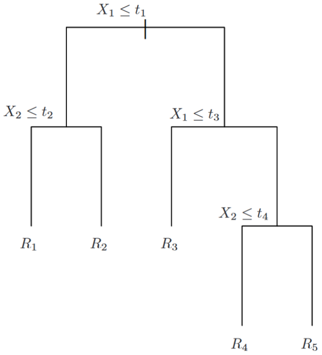
Magnitude:

- Worker at 90th vs 10th percentile of import competition: cumulative loss of 4,300EUR over 10 years for worker with mean base year earnings

Causal Forest estimation

Generalized random forest

Figure: Regression Tree



Source: Hastie et al., 2021

- Splitting rule to maximize heterogeneity in the treatment effect
- Allows for non-linearities and interactions
- Forest consisting of 10,000 trees:
  - Each tree uses a bootstrapped sub-sample
  - Random subset of variables at each split
  - Honest approach for causal estimates

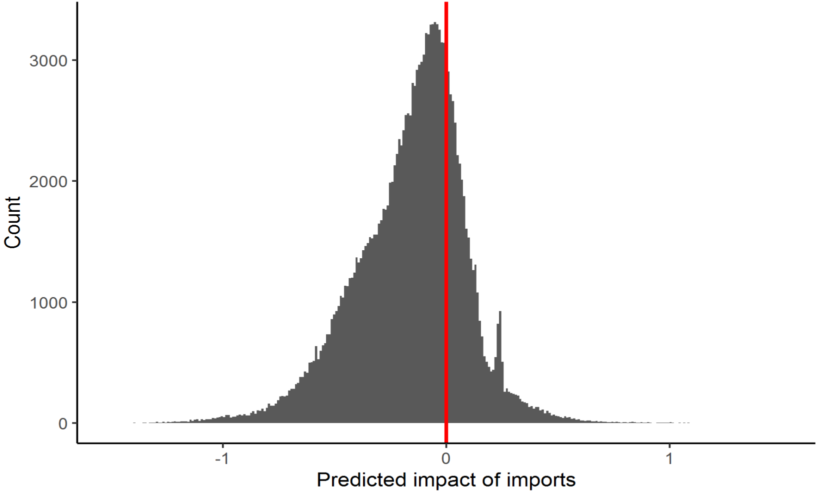
Details

Partitioning variables used in Forest

- **Measures of specific human capital**
  - 3-digit industry specificity (Utar 2018)
  - Manufacturing specificity (Utar 2018)
  - Industry tenure, firm tenure (Helm et al. 2022)
- **Measure of firm rents**
  - AKM firm wage premium (Dauth et al. 2021)
- **Demographics and others**

Detailed

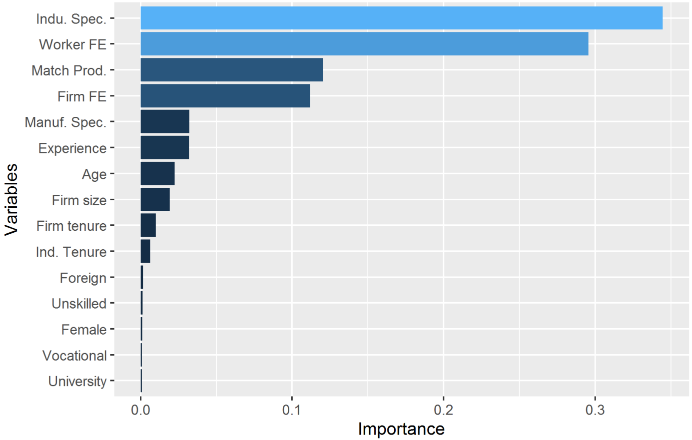
Individualized treatment effect



Mean = -0.15, close to linear IV (-0.19)

Variable importance measure points to industry-specific human capital (and workers' skill level)

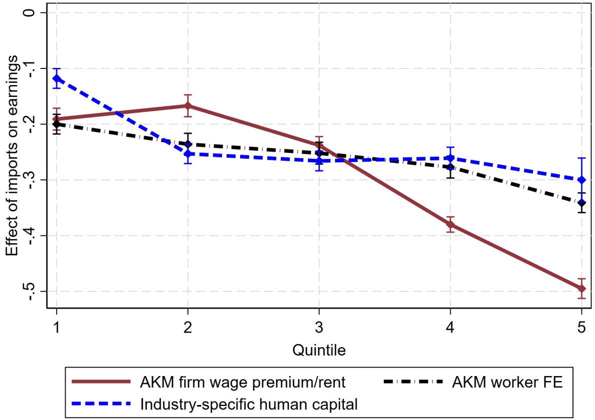
Figure: Variable importance of partitioning variables



Conditional Local Average Treatment Effects

$\Rightarrow$ Heterogeneity in effect over values of a partitioning variable (holding other variables constant!)

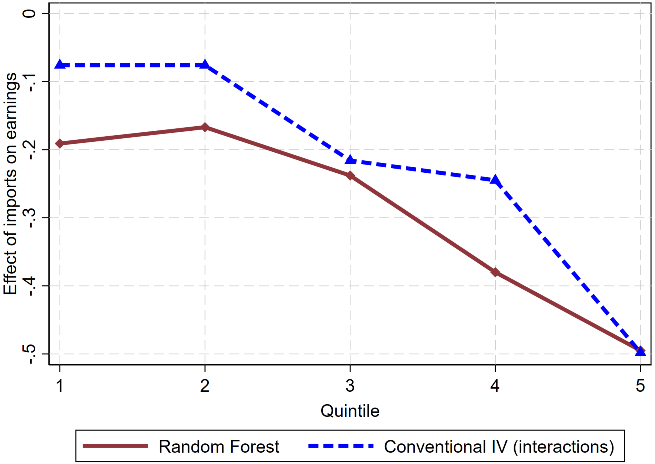
Strongest heterogeneity in AKM firm premium/rent, followed by industry-specific human capital



⇒ important role of losses in rents **and** industry-specific human capital

Forest vs. Interaction effects in conventional IV

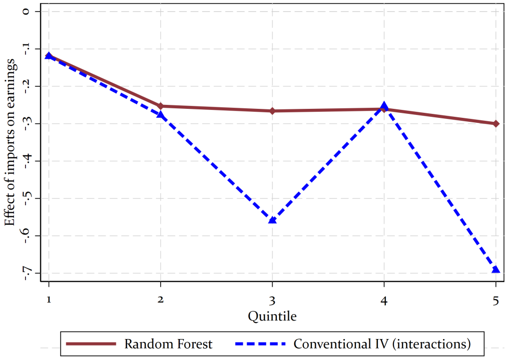
AKM firm wage premium/rent



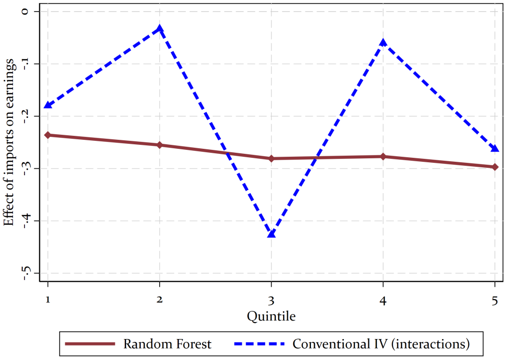
- Forest and conventional IV provide similar results

Industry-specific human capital

Industry-specificity



Industry tenure



- Forest: negative slope, in line with theory
- Conventional IV: non-linear/unclear pattern

Age

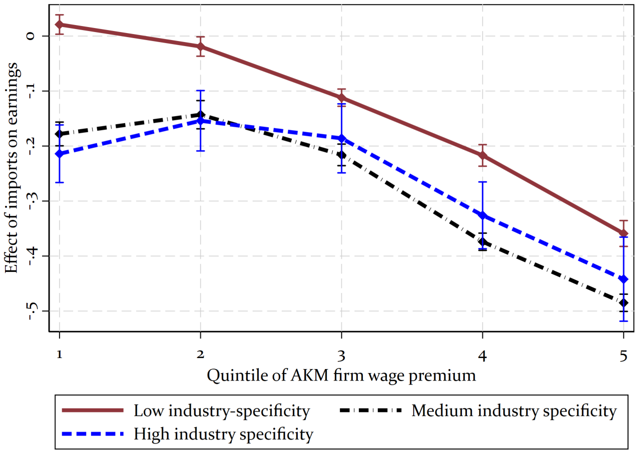
Conclusion

- Losses in rents and specificity of human capital seem to be important drivers of import-induced earnings losses
  - Retraining of workers could be effective
- Conventional interaction effects would yield misleading results

Next steps:

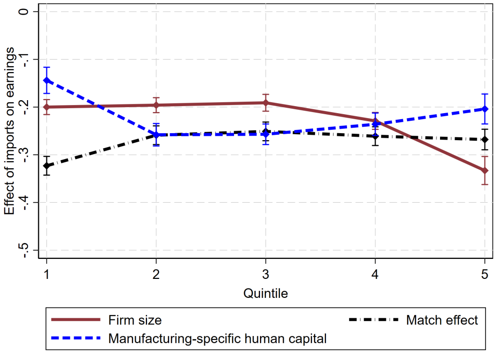
- Closer look at worker adjustment (mobility between firms, between/within industries sectors)
- Policy tree

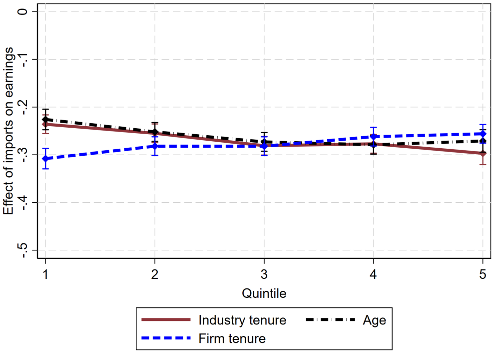
Firm wage premium & industry-specific human capital



- Positive effect for group with low industry specialisation and low firm wage premium

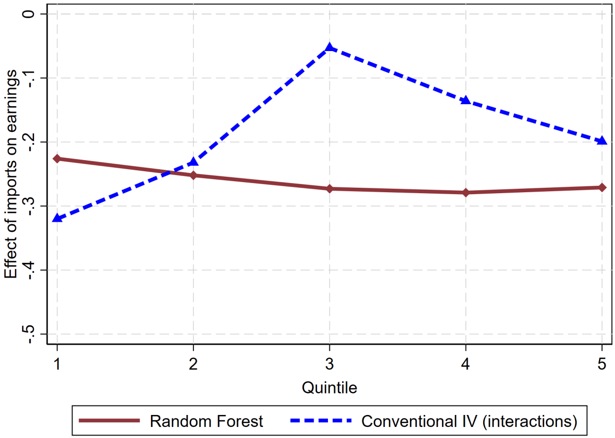
Less heterogeneity in all other variables





Back

Age



- Forest: negative slope, in line with expectation
- Conventional IV: non-linear/wrong-signed

Back

Descriptives

Table: Descriptives

	1990-2000		2000-2010	
	Mean	SD	Mean	SD
100 x earnings/base year earnings	872.735	416.767	906.218	370.318
Base year earnings	42,705.81	24,210.69	46,410.856	41,157.067
Dummy, 1 = female	0.231	0.421	0.216	0.411
Dummy, 1 = foreign national	0.123	0.328	0.094	0.292
Dummy, 1 = unskilled	0.214	0.410	0.138	0.346
Dummy, 1 = vocational training	0.714	0.452	0.761	0.426
Dummy, 1 = college degree	0.073	0.259	0.101	0.301
Δ net import exposure	0.673	0.468	0.309	0.462

Note: N=163,047

Net import exposure Back

$$NetImp_{kt} = \frac{\Delta Imports_{kt} - \Delta Exports_{kt}}{WageSum_{k(t-1)}}$$

- $\Delta Imports_{kt}$ = 10-year change in imports in industry k
- $\Delta Exports_{kt}$ = 10-year change in exports in industry k
- $WageSum_{k(t-1)}$ = Total domestic wage bill in industry k in t-1

Industry and manufacturing specificity

Back Part. var.

$$ManuSpec_{jt} = \frac{\textit{Number of workers in occupation j employed in manufacturing in the base year t}}{\textit{Total number of workers in occupation j in the base year t}}$$

$$InduSpec_{jt} = \frac{\textit{Number of workers in occupation j employed in Industry k in the base year t}}{\textit{Total number of workers in occupation j in the base year t}}$$

Firm wage premia (rents) Back Part. var.

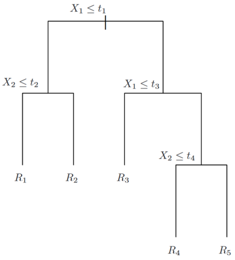
$$y_{it} = \alpha_i + \psi_{J(it)} + x'_{it}\gamma + r_{it}$$

- y_{it} = log daily wage of worker i in year t
- α_i = worker component of the wage of worker i
- x'_{it} = vector of year dummies and a cubic term in age fully interacted with education dummies of worker i
- $\psi_{J(it)}$ = proportional wage premium paid by firm J in year t to all workers
- r_{it} = error term...

$\Rightarrow$ Estimated prior to the 10-year interval

Maximize heterogeneity in the effect at each split

Figure: Regression Tree



Source: Hastie et al., 2021

$\Rightarrow$ Use conditionally centered outcomes $(\tilde{Y}, \tilde{W}, \tilde{Z})$, leave-one-out estimates of (Y, W, Z) at x for orthogonalization

Sample split by:

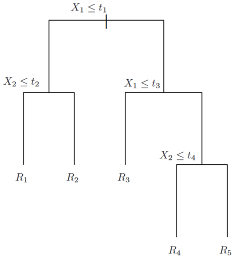
$$\hat{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} = \left(\sum_{\{i:X_i \in C_j\}} \rho_i \right)^2$$

$\Rightarrow$ Maximizing difference in treatment effects at each split

Back

Estimate a weighted local treatment effect in each leaf

Figure: Regression Tree



Source: Hastie et al., 2021

CLATE $\tau(x)$:

$$\tau(x) = \frac{Cov[Y_i, Z_i|X_i = x]}{Cov[Wi, Z_i|X_i = x]}$$

Estimation of individual weights at x:

$$\alpha_{bi}(x) = \frac{1(\{X_i \in L_b(x)\})}{|L_b(x)|}$$

$$\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x)$$

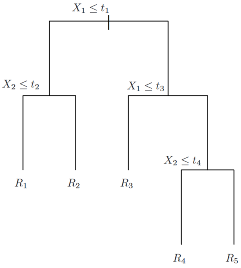
$\Rightarrow$ Comparable to weighted neighbourhood estimation

$\Rightarrow$ Assumption: homogeneous leaf-effects Back

Instrumental Forest

Back

Figure: Regression Tree



Source: Hastie et al., 2021

CLATE $\tau(x)$:

$$\tau(x) = \frac{Cov[Y_i, Z_i|X_i = x]}{Cov[Wi, Z_i|X_i = x]}$$

Estimation by moment functions:

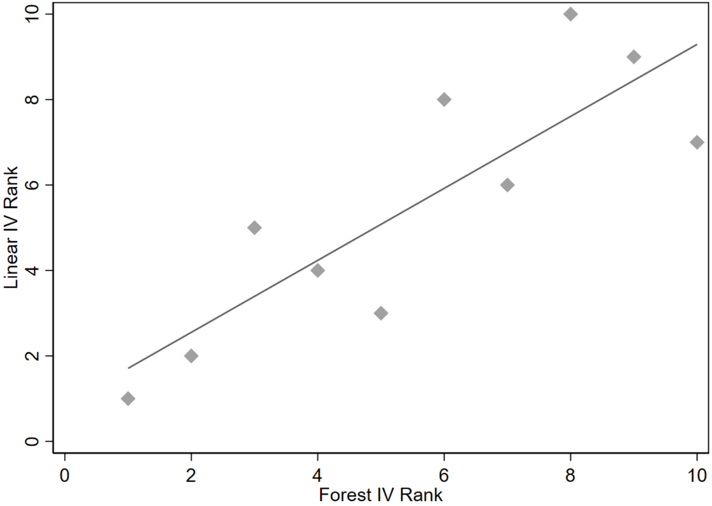
$$E[Z_i(Y_i - W_i\tau(x) - \mu(x))|X_i = x] = 0$$
$$E[Y_i - W_i\tau(x) - \mu(x)|X_i = x] = 0$$

Estimation of ρ_i:

$$\rho_i = (Z_i - \bar{Z}_P)((Yi - \bar{Y}_P) - (W_i - \bar{W}_P)\hat{\tau}_P)$$

$\Rightarrow$ Pseudo outcomes for each observation i
$\Rightarrow$ Find pseudo outcomes which maximize heterogeneity in the treatment

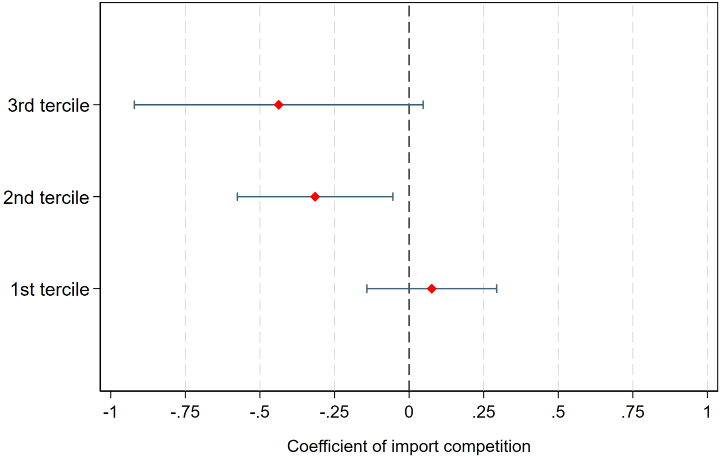
Forest accuracy by rank



$\Rightarrow$ Rank Correlation= 0.842

Significant differences between groups

Figure: Treatment effect by tercile of predicted impact on earnings



Back

Chosen settings of Forest

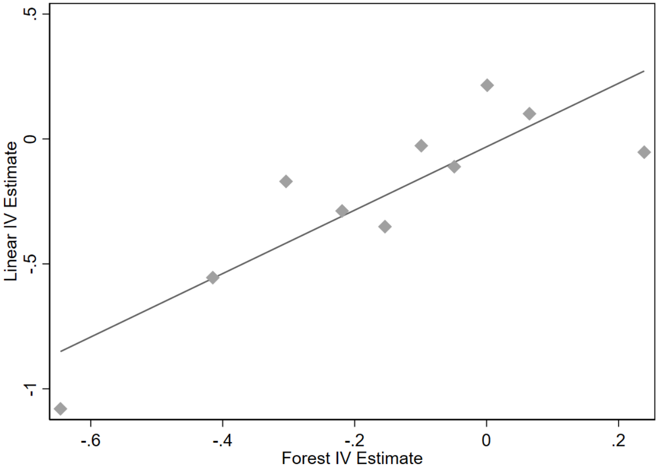
Parameter settings in GRF-algorithm:

- *num.trees*: 4.500
- *sample.fraction*: 0.5
- *mtry*: 16
- *min.node.size*: 40
- α: 0.05

$\Rightarrow$ Different settings still to test

Forest accurately predicts 'true' heterogeneity

Figure: Forest IV vs Linear IV (10 bins based on Forest)



$\Rightarrow$ Correlation= 0.864

Measure of variable importance

- Measure for the importance of a single variable for detecting heterogeneities in the treatment effect
- Higher splits (indicating a more important feature) get larger weights
- Weighted sum of the frequency feature i was split on at each depth in the forest

Related literature

Import competition and workers' labor market outcomes

- Autor et al. (2013, 2014), Utar (2018), Nilsson Hakkala and Huttunen (2018), Dauth et al. (2014, 2021), Huber and Winkler (2019), Traiberman (2019), Helm (2020), Keller and Utar (2021)

⇒ First paper to differentiate between competing explanations

Using machine learning to study treatment effect heterogeneity

- Athey et al. (2019), Lechner (2019), Gulyas and Pytka (2020), Cockx et al. (2022), Kleifgen and Lang (2022)

⇒ First paper to apply method to trade

Long-lasting earnings effects of displacement

- Jacobson et al. (1993), Couch and Placzek (2010), Davis and von Wachter (2011), Lachowska et al. (2020), Fackler et al. (2021), Helm et al. (2022), Schmieder et al. (2023)

⇒ Industry-level instead of firm-level shock

Variable Importance

$\Rightarrow$ Which variables are often used in the first splits of the trees?

Further partitioning variables used in Forest

- **Demographics and others**
  - Age, education, gender, nationality
- **Others**
  - Firm size (proxy for firm productivity, Melitz 2003)
  - AKM worker FE (unobserved skills)
  - Worker-firm match effect (Gulyas/Pytka 2022 Helm et al. 2022)
  - Experience

Back