Pretty vacant:
using job vacancies to understand the effects of labour market mismatch on UK output and productivity

Arthur Turrell,* Bradley Speigner, David Copple, James Thurgood
Bank of England

Jyldyz Djumalieva
Nesta

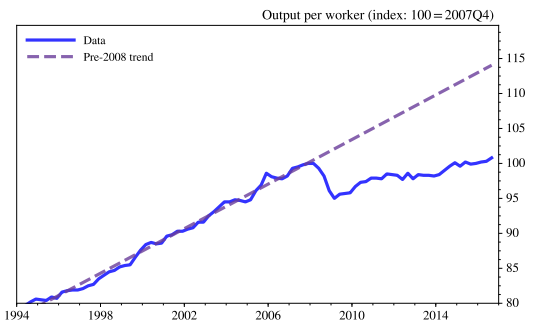
*arthur.turrell@bankofengland.co.uk

Introduction and question

The mismatch between unemployed jobseekers and job vacancies has been implicated as a driver of the UK's productivity 'puzzle'.

Question:

- To what extent would eliminating occupational or regional labour market mismatch have boosted productivity and output in the UK in the post-crisis period?

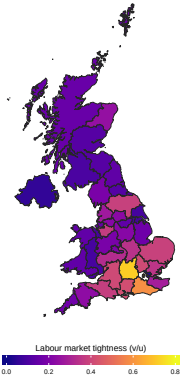


Method

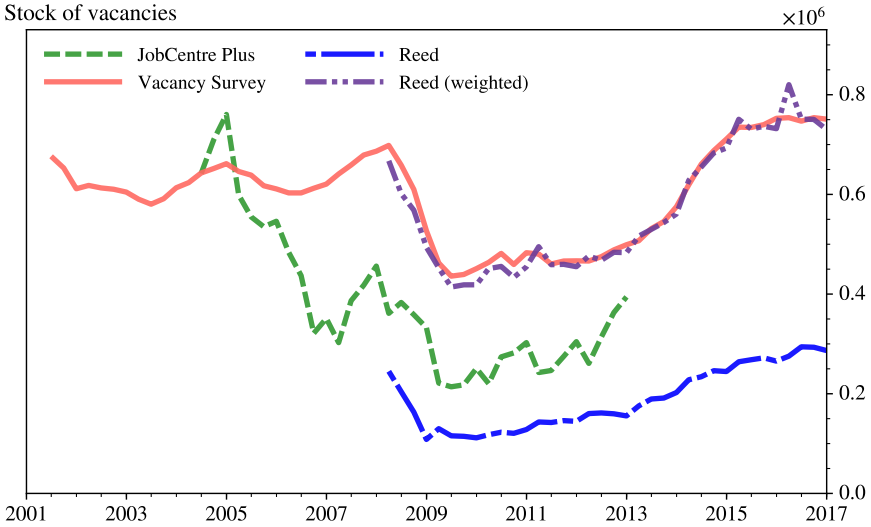
- Reweight data using overlap with official statistics by sector to overcome bias
- Map naturally occurring, messy vacancy data into Standard Occupational Classification (SOC) codes using text analysis
- New estimates of matching functions, $M = \phi u^{1-\alpha} v^{\alpha}$, by occupation and region
- Counter-factual simulations of output and productivity per worker using optimal allocation of unemployed

Example of SOC code assignment

job_title	Physicist
job_description	Make calculations about the universe, do research, perform experiments and understand the physical environment.
job_sector	Professional, scientific & technical activities
SOC_code	211

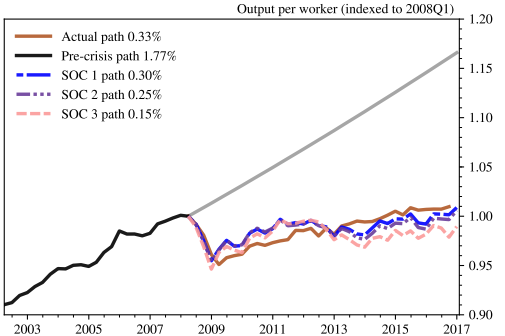


Data: 15 million job vacancies posted between 2008 and 2016



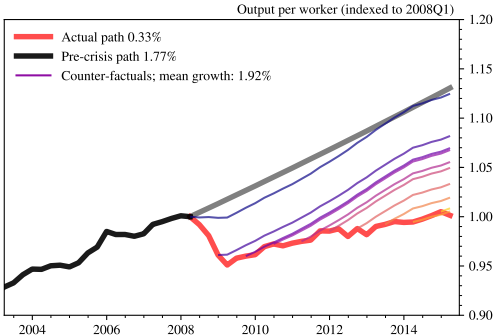
Results: occupation at 1, 2, and 3-digit SOC codes

- Significant dispersion in occupational productivity, tightness, and matching efficiency
- Unwinding occupational mismatch would have had a weak effect on productivity level (counter to previous work on JobCentre Plus data)
- Inverse relationship between worker productivity and matching efficiency
  - implies unemployment shock to higher skilled employment doubly damaging to short-term output due to high productivity, low job-finding rate combination



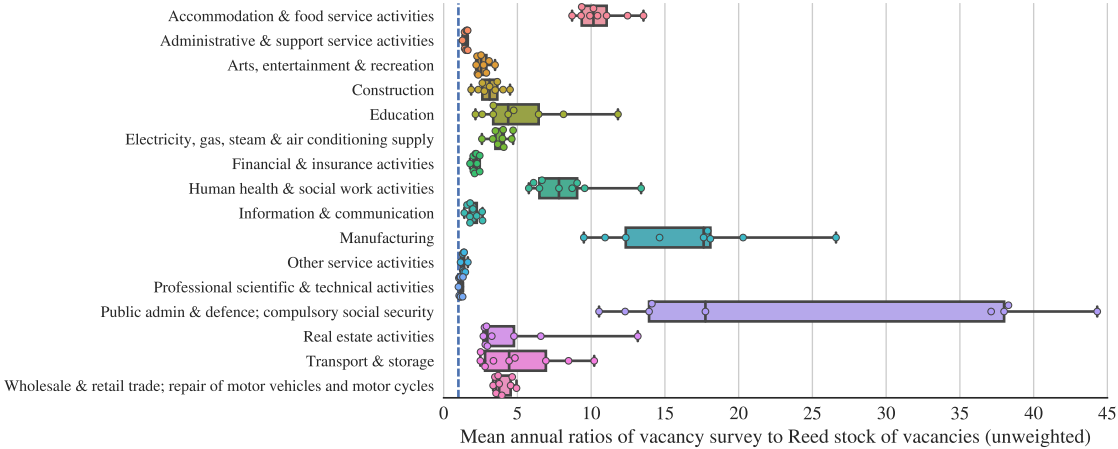
Results: region at the UK NUTS-1 level

- not clear that regional mismatch increased due to crisis; so start simulations at every 3rd quarter from start of data
- growth rates in the legend are means of all quarterly year-on-year growth rates
- unwinding regional mismatch would have substantially boosted output and productivity relative to their realised paths, bringing them in line with their pre-crisis trend rate of growth



Extra slides for discussion

Reweighting the data



Correlation between vacancy time series

	JobCentre Plus	Vacancy Survey	Reed	Reed (weighted)
JobCentre Plus	1	0.71	0.68	0.69
Vacancy Survey	-	1	0.93	0.98
Reed	-	-	1	0.90
Reed (weighted)	-	-	-	1

Algorithm to assign SOC codes

- Job title & description into algorithm which outputs 3-digit SOC code
- Term frequency-inverse document frequency represents all possible SOC codes with a matrix; t as term from within SOC descriptions and d as SOC code
- $\text{tf-idf}(\text{t}, \text{d}) = \text{tf}(\text{t}) \times \text{idf}(\text{t}, \text{d}) = \text{tf}(\text{t}) \times \left[\ln \left(\frac{1+\text{D}}{1+\text{df}(\text{t},\text{d})} \right) + 1 \right]$
- job vacancy description plus title expressed as $\hat{\vec{\text{v}}}'$
- 3-digit SOC code represented by $\hat{\vec{\text{v}}}_\text{d}$
- to calculate which SOC codes are closest to $\hat{\vec{\text{v}}}'$, we solve

$$\arg\max_\text{d} \left\{ \hat{\vec{\text{v}}}' \cdot \hat{\vec{\text{v}}}_\text{d} \right\}$$

- of top five matches, the known title with the closest fuzzy match is chosen

Econometric results on matching function estimation

The baseline empirical matching regression is

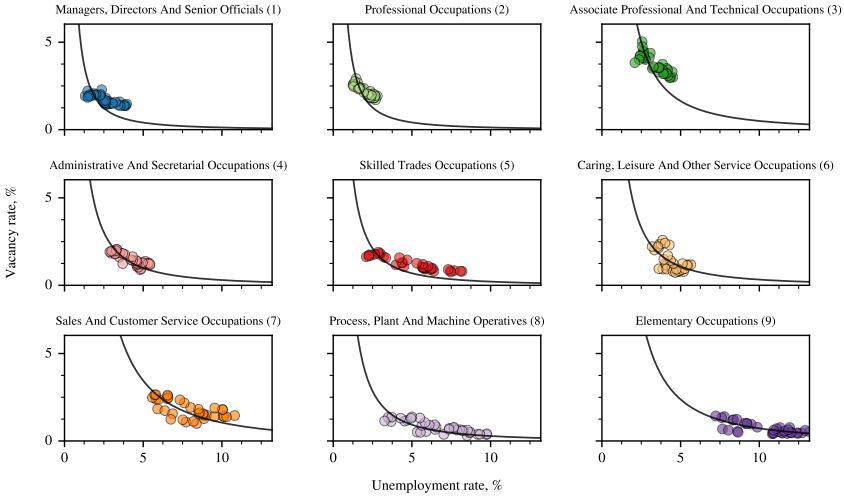
$$\ln\left(\frac{\mathrm{h_{i,t}}}{\mathrm{U_{i,t-1}}}\right) = \ln \phi_{\mathrm{i}} + \alpha \ln\left(\frac{\mathrm{V_{i,t-1}}}{\mathrm{U_{i,t-1}}}\right) + \epsilon_{\mathrm{i,t}} \tag{1}$$

where ϕ_{i} capture cross-section fixed effects.

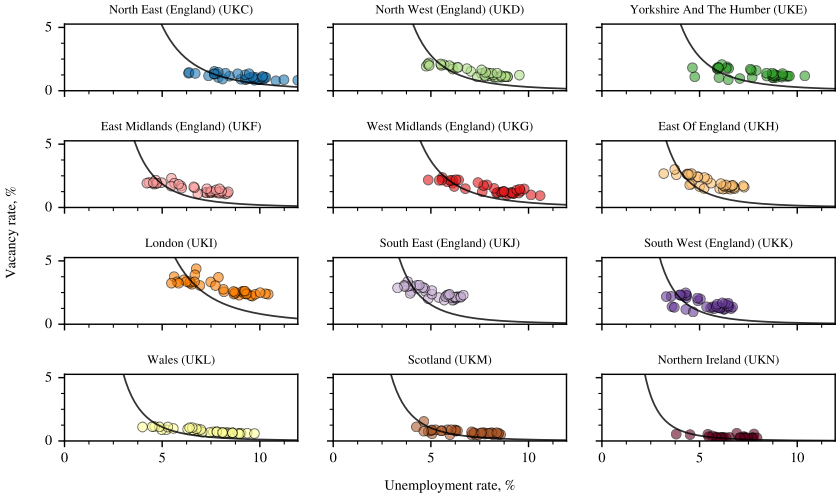
	1-digit SOC	2-digit SOC	3-digit SOC	1-digit NUTS	Aggregate data
Elasticity parameter (α)					
Point estimate (least squares)	.396	.427	.431	.254	.367
Standard error	.075	.050	.037	.020	.030
Point estimate (IV)	.392	.442	.371	.275	.350
Standard error	.073	.061	.048	.026	.031
Cross-sections	9	25	90	12	-
Observations	324	852	2120	423	35

Matching function parameter estimates. All results are significant at the 1% level.

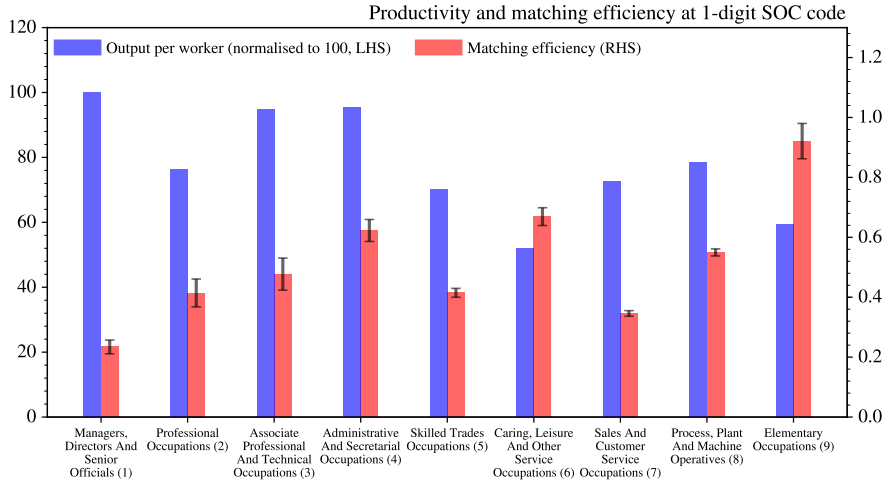
Heterogeneity matters: Beveridge curves by occupation



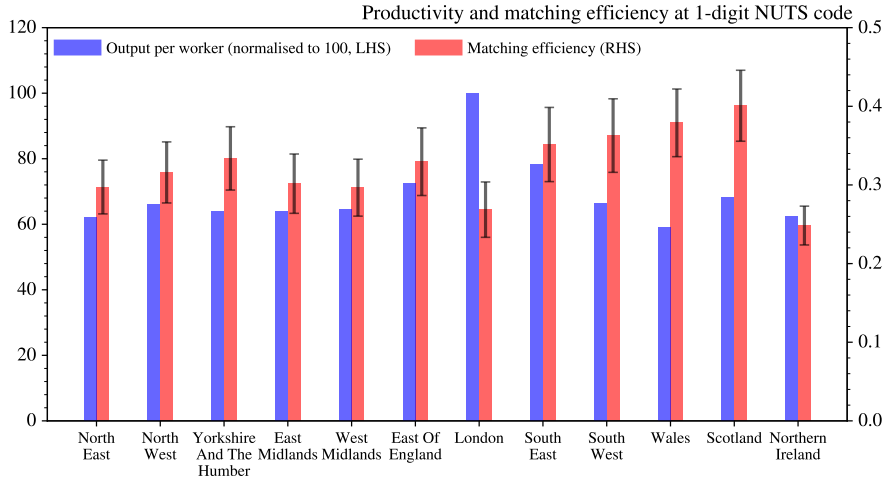
Heterogeneity matters: Beveridge curves by region



Productivity and matching efficiency: occupation



Productivity and matching efficiency: region



Social planner's problem

From Sahin et al., the planner chooses $\vec{u}_t$ to maximise output, a problem which is given by

$$V(u_t, \vec{e}_t; \Xi_t) = \max_{\{u_{i,t}\}} \left\{ \sum_i z_{i,t}(e_{i,t} + \gamma h_{i,t}) - \xi u_t + \beta \mathbb{E}\left[V(u_{t+1}, \vec{e}_{t+1}; \Xi_{t+1})\right] \right\}$$

such that $\sum_i u_{i,t} \leq u_t$ where $u_{t+1} = L_{t+1} - \sum_i e_{i,t+1}$.

This allocation allows for the construction of counter-factual output at each time period t via

$$Y_t^* = \sum_i^I z_{it} e_{it}^* + y_t^*$$

where $e_{it}^* = (1 - \xi_{t-1})e_{i,t-1}^* + h_{it}(v_{it}, u_{it}^*)$. Output per worker in the realised and counter-factual cases is given by Y_t/e_t and Y_t^*/e_t^* respectively.