Mixed Frequency Functional VARs for Nowcasting the Income Distribution in the UK

ESCoE Conference 2025

Andrea De Polis[1,2], Gary Koop[1,2], Stuart McIntyre[1,2] and James Mitchell[3]

21 May 2025

[1] University of Strathclyde
[2] Economic Statistics Centre of Excellence
[3] Federal Reserve Bank of Cleveland

This research has been funded by the ONS as part of the research programme of the Economic Statistics Centre of Excellence. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

Introduction

Introduction

There is a long tradition in microeconomics of studying the effect of policies (e.g., school reform, welfare programs) on different population strata.

- $\hookrightarrow$ Calibrate policies to affect specific demographic
- $\hookrightarrow$ Evaluate policy effect on targeted income quantiles

Introduction

There is a long tradition in microeconomics of studying the effect of policies (e.g., school reform, welfare programs) on different population strata.

$\hookrightarrow$ Calibrate policies to affect specific demographic

$\hookrightarrow$ Evaluate policy effect on targeted income quantiles

More recently, models with firms or household heterogeneity have been used to study the distributional impact of macroeconomic policies.

$\hookrightarrow$ Heterogeneous feature evolve with aggregate quantities

Motivation

Often, micro data that can shed light on a population's characteristics takes a long time to be available to researchers.

$\hookrightarrow$ Prevents detection of how shocks affect the distribution in real-time

$\hookrightarrow$ Delay in policy implementation

Motivation

Often, micro data that can shed light on a population's characteristics takes a long time to be available to researchers.

↪ Prevents detection of how shocks affect the distribution in real-time

↪ Delay in policy implementation

In the UK, survey data about population income has an average publication delay of about 1.5 years.

↪ Income data about 2018 is only available on 22 Sept., 2020

Motivation

Often, micro data that can shed light on a population's characteristics takes a long time to be available to researchers.

↪ Prevents detection of how shocks affect the distribution in real-time

↪ Delay in policy implementation

In the UK, survey data about population income has an average publication delay of about 1.5 years.

↪ Income data about 2018 is only available on 22 Sept., 2020

From the policymaker perspective, it seems relevant to have a timely and accurate measurement of distributional features of the population to improve policy planning.

In this paper

We develop a state-space model to capture the interaction between aggregate fluctuations and distributional dynamics at business cycle frequencies.

↪ Parsimoniously captures salient features of the distributions and models the joint dynamics with macro data

↪ Timely *nowcasts* of the underlying micro-distribution as implied by developments in macroeconomic aggregates

In this paper

We develop a state-space model to capture the interaction between aggregate fluctuations and distributional dynamics at business cycle frequencies.

↪ Parsimoniously captures salient features of the distributions and models the joint dynamics with macro data

↪ Timely *nowcasts* of the underlying micro-distribution as implied by developments in macroeconomic aggregates

We propose a method to evaluate the predictive accuracy of functional forecasts

↪ Generalization of the Continuously Ranked Probability Score (CRPS)

↪ Can highlight predictive accuracy across subset of the prediction support

A quick look at the data

Income distribution

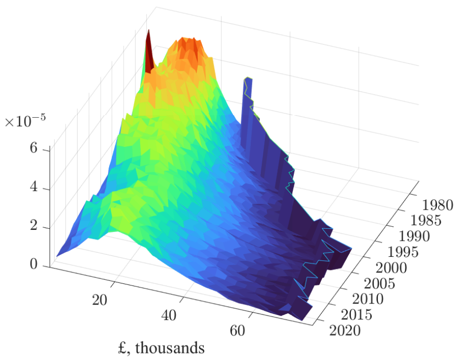
ONS survey data about the income distribution in the UK are available from 1970s.

- Income survey had changed over time
  - FES (1961-2001)
  - Expenditure and Food Survey (2001-2007)
  - Living Costs and Food survey (2007-)
- Granular information about employment status, reference year and quarter, pension benefit, etc.
- We reconstruct calendar years from survey data that refer to fiscal years.

Income distribution

ONS survey data about the income distribution in the UK are available from 1970s.

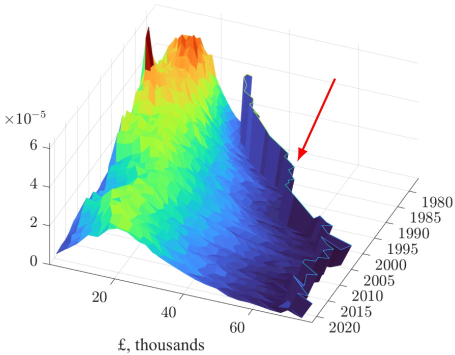
- Income survey had changed over time
  - FES (1961-2001)
  - Expenditure and Food Survey (2001-2007)
  - Living Costs and Food survey (2007-)
- Granular information about employment status, reference year and quarter, pension benefit, etc.
- We reconstruct calendar years from survey data that refer to fiscal years.



Income distribution

ONS survey data about the income distribution in the UK are available from 1970s.

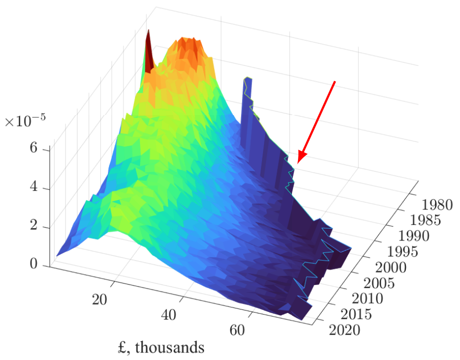
- Income survey had changed over time
  - FES (1961-2001)
  - Expenditure and Food Survey (2001-2007)
  - Living Costs and Food survey (2007-)
- Granular information about employment status, reference year and quarter, pension benefit, etc.
- We reconstruct calendar years from survey data that refer to fiscal years.
- Data top-coded from 2006.



Income distribution

ONS survey data about the income distribution in the UK are available from 1970s.

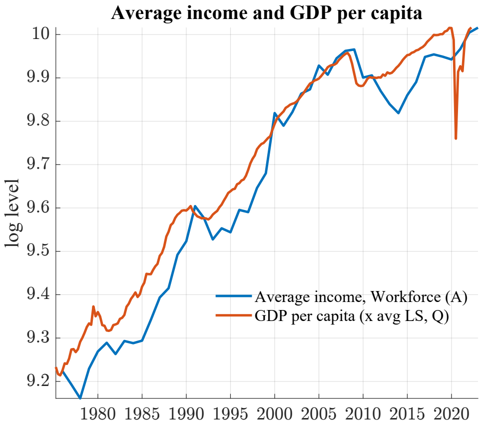
- Income survey had changed over time
  - FES (1961-2001)
  - Expenditure and Food Survey (2001-2007)
  - Living Costs and Food survey (2007-)
- Granular information about employment status, reference year and quarter, pension benefit, etc.
- We reconstruct calendar years from survey data that refer to fiscal years.
- Data top-coded from 2006.



We use income data for the workforce (employed & unemployed people)

Micro and Macro data

Increase in average income seems proportional to GDP per capita



Average income and output per capita - ONS survey data

Model & Methods

Model - the problem

Let $Y_t = [y_{1,t}, \ldots, y_{n_{macro},t}]'$ contain n_{macro} variables. We can model their joint dynamics as a Vector Autoregression (VAR)

$$Y_t = \Phi_1 Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

Model - the problem

Let $Y_t = [y_{1,t}, \ldots, y_{n_{macro},t}]'$ contain n_{macro} variables. We can model their joint dynamics along with any distribution as a functional-VAR

$$Y_t = \Phi_{yy} Y_{t-1} + \int_{\mathbb{R}^+} \Phi_{y\delta}(\tilde{x}) \delta_{t-1}(\tilde{x}) d\tilde{x} + \varepsilon_{y,t}$$

$$\delta_t(x) = \Phi_{\delta y}(x) Y_{t-1} + \int_{\mathbb{R}^+} \Phi_{\delta\delta}(x, \tilde{x}) \delta_{t-1}(\tilde{x}) d\tilde{x} + \varepsilon_{\delta,t}(x)$$

Good news!

$\hookrightarrow$ We can add the entire *distribution of micro-outcomes* at time t as $\delta_t(x) = \ln p_t(x)$ and expand the model.

Model - the problem

Let $Y_t = [y_{1,t}, \ldots, y_{n_{macro},t}]'$ contain n_{macro} variables. We can model their joint dynamics along with any distribution as a functional-VAR

$$Y_t = \Phi_{yy} Y_{t-1} + \int_{\mathbb{R}+} \Phi_{y\delta}(\tilde{x}) \delta_{t-1}(\tilde{x}) d\tilde{x} + \varepsilon_{y,t}$$

$$\delta_t(x) = \Phi_{\delta y}(x) Y_{t-1} + \int_{\mathbb{R}+} \Phi_{\delta\delta}(x, \tilde{x}) \delta_{t-1}(\tilde{x}) d\tilde{x} + \varepsilon_{\delta,t}(x)$$

Good news!

$\hookrightarrow$ We can add the entire *distribution of micro-outcomes* at time t as $\delta_t(x) = \ln p_t(x)$ and expand the model.

However

$\hookrightarrow$ This is an infinite-dimensional model, that poses several challenges.

Model - the (approximate) solution

We follow Chang et al. (2024) and approximate $\delta_t(x)$ with a K-dimensional cubic spline, which allows us to keep the dimension of the model tractable. That is, *we only need* K $\alpha_{k,t}$ *factors* to recover the full density as

$$\delta_t(x) = \sum_{k=1}^{K} \alpha_{k,t}\zeta_k(x)$$

where $\zeta(x)$ is spline functions.

Model - the (approximate) solution

We follow Chang et al. (2024) and approximate $\delta_t(x)$ with a K-dimensional cubic spline, which allows us to keep the dimension of the model tractable. That is, *we only need K $\alpha_{k,t}$ factors* to recover the full density as

$$\delta_t(x) = \sum_{k=1}^{K} \alpha_{k,t} \zeta_k(x)$$

where $\zeta(x)$ is spline functions.

Therefore,

$$Z_t = \sum_{p=1}^{P} \Phi_p Z_{t-p} + u_t, \quad u_t \sim \mathcal{N}(0, \Xi)$$

where $Z_t = [Y_t \;\; \alpha_t]'$ is a $n = n_{macro} + K$ vector of observables, and Ξ is unrestricted.

Quick digression - Few alternatives

Functional VARs with spline basis are quite flexible and relatively easy to work with, but this is not the only way! Alternative measure that can be used are:

- measures of inequality, Gini coefficient (Mumtaz and Theophilopoulou, 2017)
  ↪ this coefficient is defined over [0,1], and they miss info about the distribution
- quantiles of the distribution
  ↪ similar to our approach, but difficult to impose non-crossing restriction
- pseudo-individuals (Koop et al., 2024)
  ↪ this approach amounts to slicing the income distribution by individuals' characteristics.

Methods - Sieve Estimation

Chang et al. (2024) show that the filtering problem for this type of functional-VAR (fVAR) can easily be broken into a two-step procedure.

1. Estimate the sieve coefficients, α_t for each year by fitting the earning distribution to a K-knots cubic spline via ML.
2. Add α_t to the VAR and estimate the joint dynamics.

Additional details:

- $\hookrightarrow$ Top-coding issue addressed via penalized ML.
- $\hookrightarrow$ Earnings data are detrended by a factor of $\approx (.6\, GDP_t)^{-1}$
- $\hookrightarrow$ We apply an inverse hyperbolic sine transformation to the earning data to ensure non-negativity of the density functions.

Methods - Data

We use quarterly data for the period 1975Q1 to 2021Q4 for the macro variables:

- Unemployment
- Inflation rate
- Labour share
- Hours per capita
- Output per hour
- Real GDP
- Consumption
- Bank rate

Data are appropriately transformed to guarantee *stationarity* and demeaned.

Methods - Data

We use quarterly data for the period 1975Q1 to 2021Q4 for the macro variables:

- Unemployment
- Inflation rate
- Labour share
- Hours per capita
- Output per hour
- Real GDP
- Consumption
- Bank rate

Data are appropriately transformed to guarantee *stationarity* and demeaned.

Income survey data are released once a year with information about yearly income.

$\hookrightarrow$ Distributional data enter the model at the annual frequency, in Q4 of each year, and treat the previous three quarters as missing data.

Methods - Data

We use quarterly data for the period 1975Q1 to 2021Q4 for the macro variables:

- Unemployment
- Inflation rate
- Labour share
- Hours per capita
- Output per hour
- Real GDP
- Consumption
- Bank rate

Data are appropriately transformed to guarantee *stationarity* and demeaned.

Income survey data are released once a year with information about yearly income.

$\hookrightarrow$ Distributional data enter the model at the annual frequency, in Q4 of each year, and treat the previous three quarters as missing data.

We have a mixed frequency problem!

Methods - Intertemporal restrictions

We cast the mixed-frequency problem in state-space

$$f_t = \sum_{p=1}^{P} \Phi_p f_{t-p} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma) \tag{1}$$

$$Z_t = \Lambda f_t \tag{2}$$

f_t contains *high-frequency estimates* of the micro-density and *spline approximation errors*, and Λ imposes intertemporal restriction to recover annual sieve factors, α_t.

Methods - Intertemporal restrictions

We cast the mixed-frequency problem in state-space

$$f_t = \sum_{p=1}^{P} \Phi_p f_{t-p} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma) \tag{1}$$

$$Z_t = \Lambda f_t \tag{2}$$

f_t contains *high-frequency estimates* of the micro-density and *spline approximation errors*, and Λ imposes intertemporal restriction to recover annual sieve factors, α_t.

$\hookrightarrow$ We only observe sieve coefficients $\rightarrow$ highly nonlinear mapping to density

Methods - Intertemporal restrictions

We cast the mixed-frequency problem in state-space

$$f_t = \sum_{p=1}^{P} \Phi_p f_{t-p} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma) \tag{1}$$

$$Z_t = \Lambda f_t + G\boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim \mathcal{N}(0, \Omega) \tag{2}$$

f_t contains *high-frequency estimates* of the micro-density and *spline approximation errors*, and Λ imposes intertemporal restriction to recover annual sieve factors, α_t.

$\hookrightarrow$ We only observe sieve coefficients $\rightarrow$ highly nonlinear mapping to density

$\hookrightarrow$ Average accumulator: $\alpha_t = \frac{1}{4}\sum_{q=1}^{4} f_{t,q} \rightarrow$ Annual frequency. ▸ Fit

Application

Nowcasting exercise

We employ the model within a nowcasting exercise to monitor developments in the income distribution in real-time.

↪ Pseudo-vintages since 1990, and real-time vintages for the macro data from mid 2015 (ONS data)

↪ Micro data are released at the date of the first version of the survey results

Nowcasting exercise

We employ the model within a nowcasting exercise to monitor developments in the income distribution in real-time.

↪ Pseudo-vintages since 1990, and real-time vintages for the macro data from mid 2015 (ONS data)

↪ Micro data are released at the date of the first version of the survey results

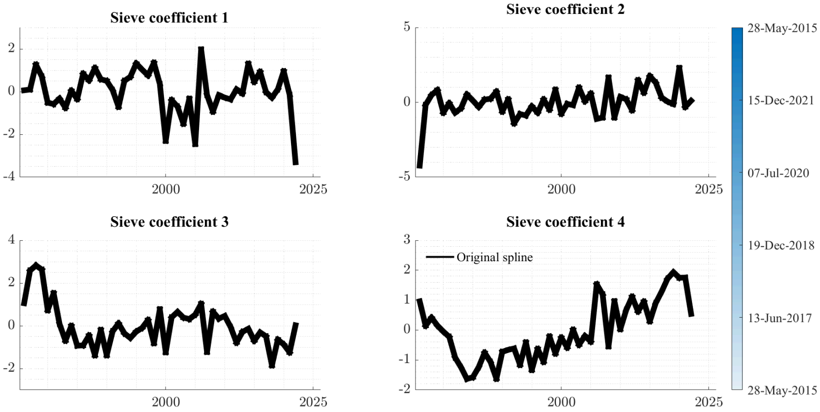
We compare our benchmark fVAR with K=4 against:

- a simple "micro-VAR" with annual sieve coeffitients only (mVAR)
- a VAR on K quantiles of the data (qVAR)
- a VAR on K quantiles of the data and macro variables (qmVAR)

Sieve coefficients

The benchmark model *only* observe the K=4 annual sieve coefficients estimated in the first stage.

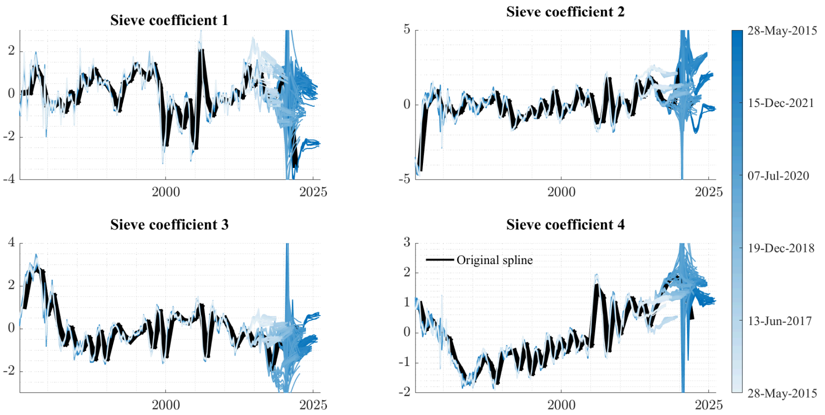
▸ Recovering the density



Sieve coefficients

The benchmark model *only* observe the K=4 annual sieve coefficients estimated in the first stage. When new data are released, we produce quarterly predictions for the coefficients.

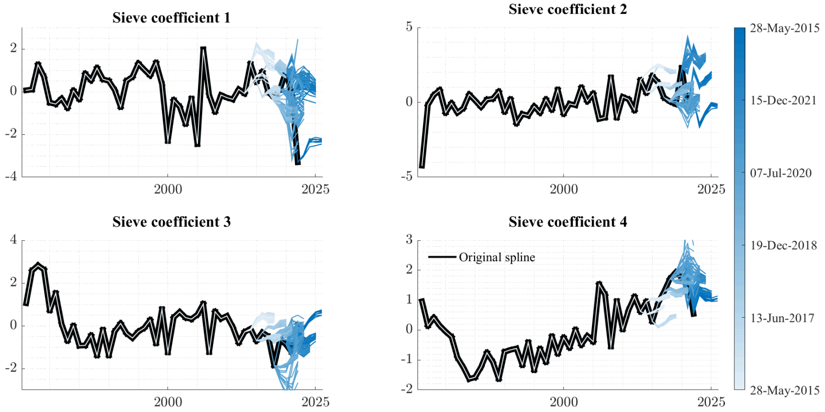
▸ Recovering the density



Sieve coefficients

The benchmark model *only* observe the K=4 annual sieve coefficients estimated in the first stage. When new data are released, we produce quarterly predictions for the coefficients. Equation (2) generates annual now- and forecasts.

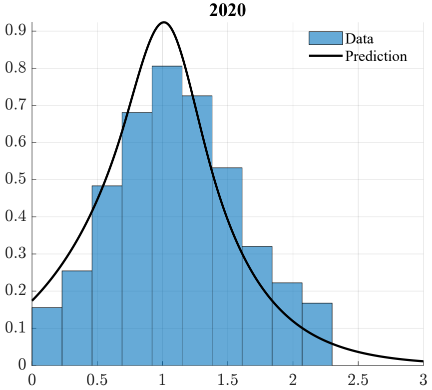
Recovering the density



Recovering predictive densities

From the sieve coefficients, it's easy now to compute proper densities as

$$p^{(K)}\left(x|\alpha_t\right) = \frac{\exp\left(\zeta_t'(x)\alpha_t\right)}{\int \exp\left(\zeta_t'(x)\alpha_t\right) dx}$$



How to evaluate functional forecasts?

Consider the Continuously Ranked Probability Score (CRPS)

$$CRPS(\hat{y}, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{I}_{\{y \leq z\}}\right)^2 dz \tag{3}$$

How to evaluate functional forecasts?

Consider the Continuously Ranked Probability Score (CRPS)

$$CRPS(\hat{y}, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{I}_{\{y \leq z\}}\right)^2 dz \tag{3}$$

Our target is the *whole distribution* of UK earnings, F_Y, not a single realization, y!

How to evaluate functional forecasts?

Consider the functional Continuously Ranked Probability Score (CRPS)

$$fCRPS(\hat{Y}, Y) = \int_{\mathbb{R}} \omega(q) \left(F(z) - F_Y\right)^2 dz \tag{3}$$

Our target is the *whole distribution* of UK earnings, F_Y, not a single realization, y!

$\hookrightarrow$ This metric is akin to the *Cramér-von Mises* criterion.

$\hookrightarrow$ $\omega(q)$ is a weighting function specified over the quantiles of the predictive distributions that allows to highlight different regions of interest.

Recovering empirical CDF from histograms

To avoid giving unfair advantage to any of the competing models, we recover the empirical CDF from the data as

$$F(X_t) = \{v_j\}_{i=0}^{J} \tag{4}$$

$$v_i = \sum_{j=0}^{i} \frac{c_j}{N}, \tag{5}$$

where v_j are monotonically increasing *bin values*, and c_i is the number of observations falling within each bin i.

How to evaluate functional forecasts?

We consider four specifications for $\omega(q)$

- fCRPS: $\omega\left(q \in [0\ 1]\right) = 1$
- Left tail: $\omega(q < 0.2) = 1$
- Center: $\omega\left(0.2 \leq q \leq 0.8\right) = 1$
- Right tail: $\omega(q > 0.8) = 1$

How to evaluate functional forecasts?

We consider four specifications for $\omega(q)$

- fCRPS: $\omega\left(q \in [0\ 1]\right) = 1$
- Left tail: $\omega(q < 0.2) = 1$
- Center: $\omega\left(0.2 \leq q \leq 0.8\right) = 1$
- Right tail: $\omega(q > 0.8) = 1$

	mVAR	qVAR	qmVAR
fCRPS	1.050 (0.152)	**1.698** (0.000)	**1.673** (0.000)
Center	**1.225** (0.001)	1.126 (0.274)	0.988 (0.527)
Right	1.015 (0.395)	**1.802** (0.000)	**1.805** (0.000)
Left	0.895 (0.715)	**2.604** (0.000)	**2.521** (0.003)

Scores are relative to the fVAR's; Diebold and Mariano p-values are in parentheses.

Quantile prediction

We evaluate the point predictions quantile by quantile.

We consider the 94^{th} quantile as beyond that data are top-coded.

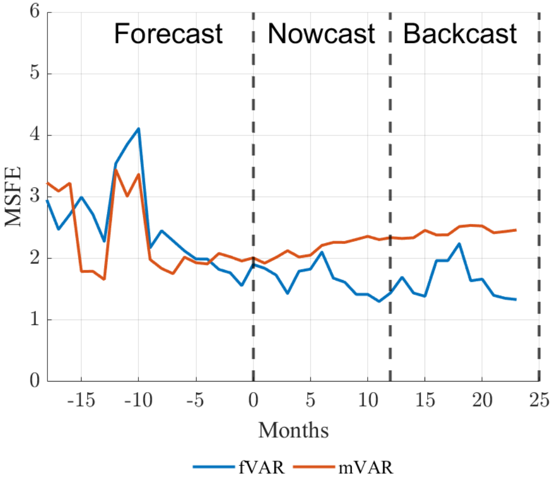
MSFE by quantile

	q05	q25	q50	q75	q94
mVAR	0.941	1.906	1.970	2.700	1.916
qVAR	1.578	5.092	5.715	14.364	4.357
qmVAR	1.670	2.897	5.148	12.818	4.057

Note: The results are reported as the average relative performance of the models over that of the fVAR.

Do macro information improve nowcasts?

We compare the Mean Squared Forecast Error for selected quantiles across models.



Comparison for median predictions.

Conclusions

What we have so far

We have developed a functional-VAR model to jointly estimate the dynamics of aggregate data and the income distribution in the UK over the last 50 years.

$\hookrightarrow$ Macro data add relevant information to predict micro outcomes

$\hookrightarrow$ Our nowcasting model seems to fare well compared to competitor models, and can easily address the mixed-frequency nature of the problem

Next steps

$\hookrightarrow$ How can this methodology be used for policy analysis?
We can identify shock and compute the response of different quantiles of the population.

Mixed Frequency Functional VARs for Nowcasting the Income Distribution in the UK

ESCoE Conference 2025

Andrea De Polis[1,2], Gary Koop[1,2], Stuart McIntyre[1,2] and James Mitchell[3]

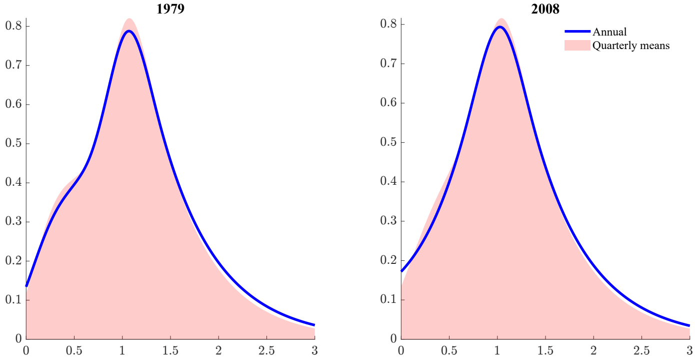
21 May 2025

[1] University of Strathclyde
[2] Economic Statistics Centre of Excellence
[3] Federal Reserve Bank of Cleveland

This research has been funded by the ONS as part of the research programme of the Economic Statistics Centre of Excellence. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

Appendix

Intertemporal restrictions fit



Back

Recovering cross-sectional densities

At every period t, we observe N_t^{survey} draws, $x_{i,t}$, from $p_t^{(K)}(x)$. Defining $\delta_t^{(K)}(x) = \log p_t^{(K)}(x)$, we can then compute

$$p_t^{(K)}(x) = \frac{exp\left(\delta_t^{(K)}(x)\right)}{\int_{\mathbb{R}+} exp\left(\delta_t^{(K)}(\tilde{x})\right) d\tilde{x}}.$$

In the model, we only observe α_t, and so

$$p_t^{(K)}(x) = exp\left(N_t^{survey}\mathcal{F}^{(K)}\left(\alpha_t|X_t\right)\right)$$

$$\mathcal{F}^{(K)}\left(\alpha_t|X_t\right) = \zeta_k(X_t)\alpha_t - \varphi(\alpha_t)$$

$$\varphi(\alpha_t) = \int_{\mathbb{R}+} exp\left(\zeta_k(\tilde{x})\alpha_t\right) d\tilde{x}.$$

◂ Back

Is measurement error helpful?

months	q05	q25	q50	q75	q96	q05	q25	q50	q75	q96
	fVAR 4									
	w me									
All	0.583	1.777	2.801	7.245	74.954					
Forecasts	0.586	1.944	2.848	8.095	87.571					
Nowcast	0.522	1.587	2.562	6.657	73.019					
Backcast	0.617	1.751	2.847	6.784	65.943					
	mVAR 4									
	w me					w/o me				
All	0.838	1.269	1.020	1.309	1.302	0.890	1.289	1.016	1.335	1.325
Forecasts	1.036	1.806	1.567	1.709	1.414	1.088	1.825	1.573	1.742	1.436
Nowcast	0.973	1.467	1.115	1.474	1.379	1.047	1.506	1.120	1.511	1.404
Backcast	0.628	0.798	0.607	0.964	1.196	0.657	0.813	0.598	0.984	1.218
	mVAR 8									
	w me					w/o me				
All	1.000	1.538	1.257	1.381	2.545	1.023	1.559	1.267	1.405	2.586
Forecasts	1.183	2.064	1.768	1.693	2.460	1.221	2.102	1.779	1.714	2.438
Nowcast	1.175	1.793	1.344	1.513	2.617	1.214	1.789	1.334	1.527	2.625
Backcast	0.757	1.046	0.887	1.121	2.588	0.775	1.068	0.906	1.161	2.701

Back

Is measurement error helpful?

	fCRPS	fCRPS left	fCRPS center	fCRPS right	fCRPS	fCRPS left	fCRPS center	fCRPS right
	fVAR 4							
	w me							
All	0.770	0.037	0.457	0.276				
Forecast	0.726	0.036	0.421	0.269				
Nowcast	0.780	0.035	0.467	0.277				
Backcast	0.789	0.038	0.472	0.279				
	mVAR 4							
	w me				w/o me			
All	1.051	1.086	1.060	1.039	1.070	1.121	1.071	1.068
Forecast	1.194	1.401	1.251	1.079	1.216	1.422	1.265	1.112
Nowcast	1.069	1.189	1.071	1.054	1.094	1.242	1.092	1.084
Backcast	0.952	0.821	0.934	1.007	0.969	0.854	0.944	1.034
	mVAR 8							
	w me				w/o me			
All	1.636	1.419	1.327	2.176	1.655	1.443	1.313	2.251
Forecast	1.811	1.763	1.541	2.240	1.828	1.815	1.520	2.312
Nowcast	1.615	1.546	1.296	2.166	1.609	1.562	1.260	2.207
Backcast	1.540	1.118	1.216	2.143	1.582	1.133	1.222	2.251

Back

Bibliography

References

Chang, M., X. Chen, and F. Schorfheide (2024): "Heterogeneity and aggregate fluctations," *Journal of Political Economy*.

Mumtaz, H. and A. Theophilopoulou (2017): "The impact of monetary policy on inequality in the UK. An empirical analysis," *European Economic Review*, 98, 410–423.