EUROPEAN CENTRAL BANK

EUROSYSTEM

Evaluation of output gap estimates

SUERF workshop

The views expressed in this presentation and the related paper are those of the authors' and do not necessarily reflect the views of the ECB.

27/03/2023

Ana Seco Justo, Bela Szörfi
DG-E/SSL

Summary

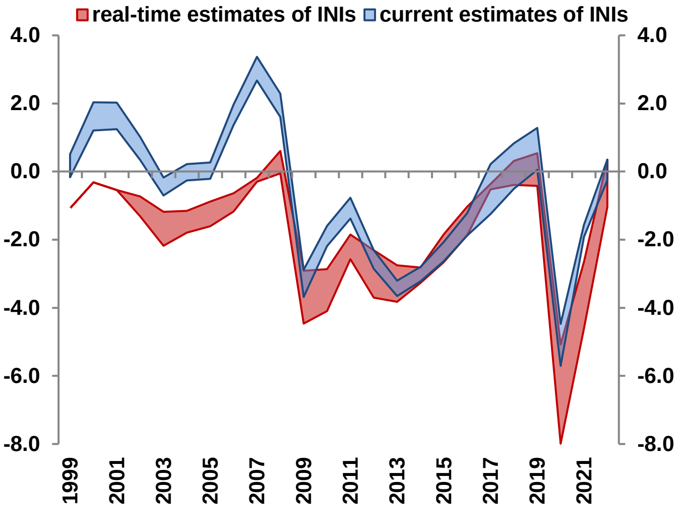
- **Aim of the analysis:**
  - To propose assessment criteria for output gap and PO estimates;
  - To illustrate these with a suite of models;
  - To assess the reliability of estimates by international institutions;
- **Main findings:**
  - A visual inspection of output gaps is insufficient to decide if they are “good” or “bad”, we need more formal assessment.
  - There is no single method that would produce a “best” output gap estimate, each method has advantages and disadvantages.
  - Amongst international institutions, output gap estimates of the EC revise the least, those of OECD revise the most.
  - Data revisions do play a significant role. In addition, real GDP forecast errors are transmitted to revisions of past potential output and output gap.

Overview

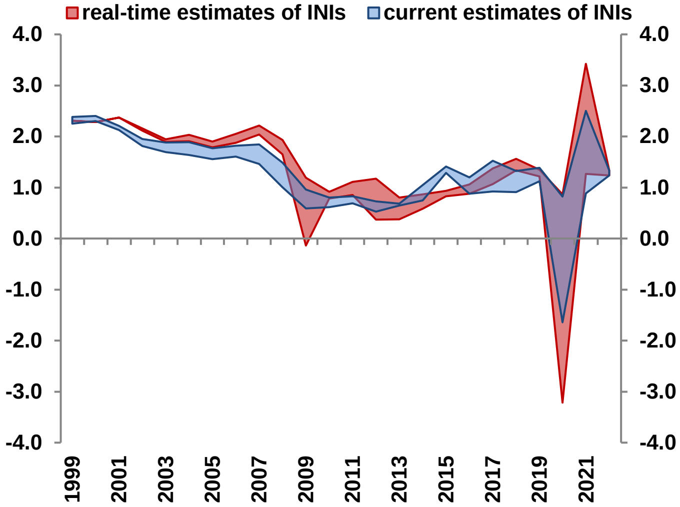
Motivation

- Output gap estimates are uncertain and are often revised, this results in unreliable signals about the business cycle and future inflation.

Output gap estimates, euro area

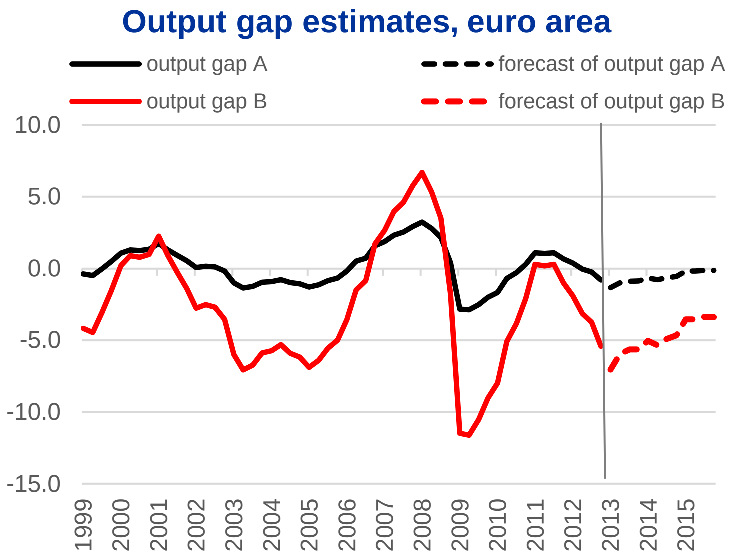


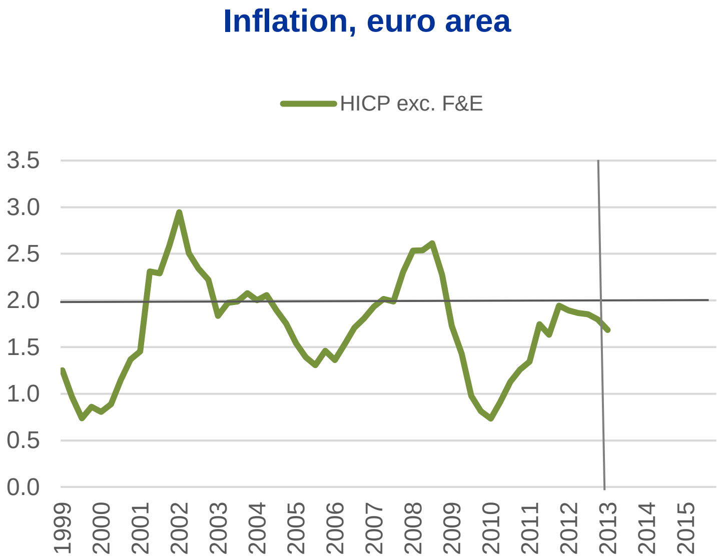
Potential growth estimates, euro area



Motivation

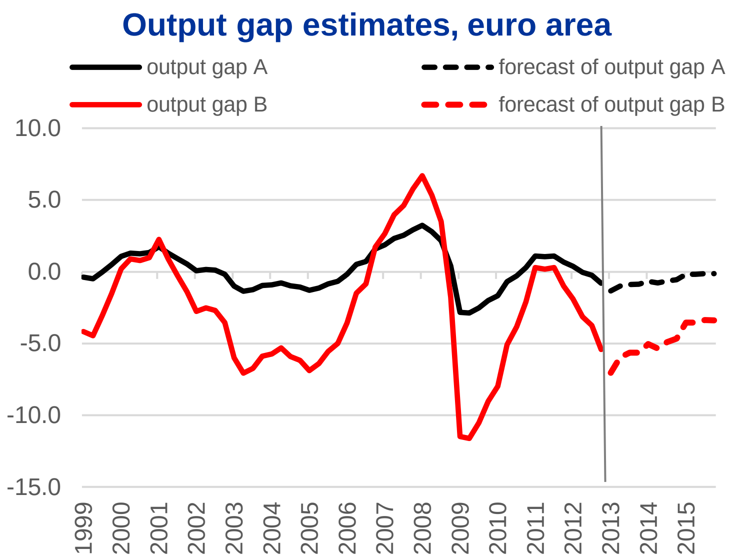
- There are different methods to estimate output gaps, but it is difficult to evaluate them just from visual inspection.

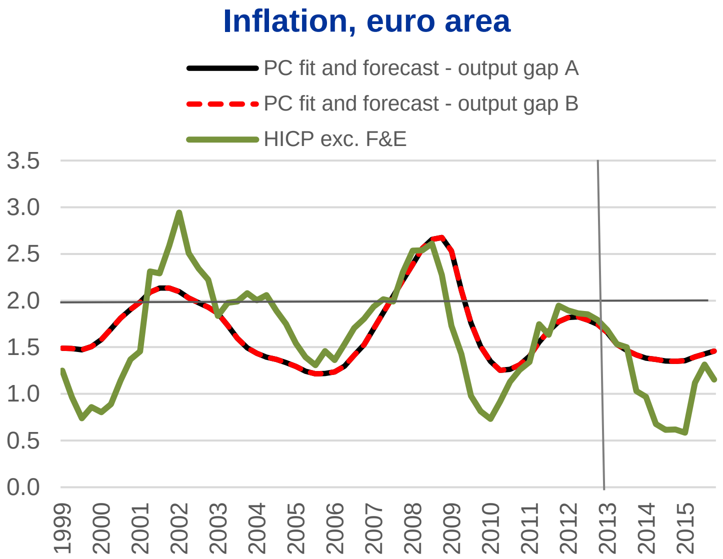




Motivation

- There are different methods to estimate output gaps, but it is difficult to evaluate them just from visual inspection.
- Departing from the desired properties of output gaps, we introduce 4 quantitative (and one qualitative) criteria to evaluate the different methodologies.





Evaluation criteria and an illustration

- **We propose 4 quantitative and one qualitative evaluation criteria:**
- **Reliability**
  - **Average revisions**
  - **Absolute revisions**
  - **Correlation between final and real-time estimates**
  - **Sign change**
- **Cyclicality**
- **Symmetry**
- **Inflation forecasting**
- **Plausibility**

Evaluation criteria and an illustration

- Ten methods ranging from simple but widely used to more complex;
- Each of them estimated over a sample of 1995q1-2019q4;
- Where needed, sub-samples / quasi real-time estimates are created, with samples ending 2005q1 – 2019q4: 60 sub-samples.

Methods to evaluate

method	
Hodrick-Prescott (lambda=1)	Univariate statistical filter
Hodrick-Prescott (lambda=1600)	Univariate statistical filter
Hodrick-Prescott (lambda=5000)	Univariate statistical filter
Blanchard-Quah	Multivariate structural VAR
Hamilton	Univariate time series regression
Beveridge-Nelson	Univariate statistical filter
Christiano-Fitzgerald	Univariate statistical filter
Jarocinski-Lenza	Dynamic Factor Model
Survey-Based Measure of Slack	Multivariate statistical filter
Unobserved Components Model	semi-structural model based multivariate filter

Evaluation criteria and an illustration

- Average revisions:
  - HP does not seem to revise much, but this masks larger revisions in different periods;
  - More than 1pp average revisions of the J-L estimates;
  - Revisions generally larger in the pre-2013 period.
- Absolute revisions:
  - J-L and smooth HP revise the most.
- Correlation:
  - J-L turns out to revise such that the revised estimate remains fairly strongly correlated with the original estimate
  - Beveridge-Nelson, HP with lambda=1 and UCM are the most stable
- Sign change:
  - B-Q, C-F change sign the most often;
  - If an estimate is almost always negative/positive, it doesn’t change sign often but still might revise a lot (J-L case in point).

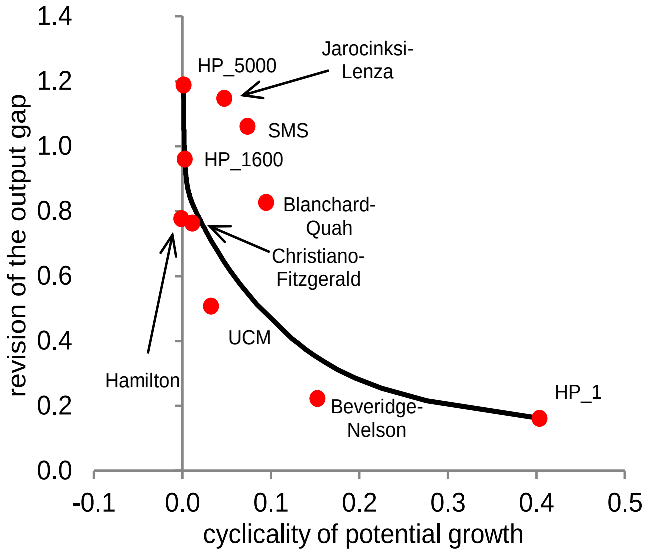
Evaluation criteria and an illustration

- Volatile HP trend is of course the most cyclical;
- Hamilton, J-L are non-cyclical.
- There is a trade-off between the cyclicality of PO and revisions of OG estimates

Cyclicality of potential growth estimates

method	cyclicality 1 (Measure 5)	cyclicality 2 (Measure 6)
HP_1	0.11	0.40
HP_1600	0.84	0.00
HP_5000	0.98	0.00
Blanchard-Quah	1.63	0.04
Hamilton	1.85	0.00
Beveridge-Nelson	0.55	0.15
Christiano-Fitzgerald	0.85	0.01
Jarocinksi-Lenza	2.53	0.05
SMS	1.24	0.07
UCM	1.49	0.03

Trade-off between cyclicality and revisions



Note: lighter shades represent less cyclical estimates, darker shades represent more cyclical estimates

Evaluation criteria and an illustration

- All OG estimates improve the inflation forecast compared to real GDP in the PC;
- J-L, C-F do particularly well;
- The UCM also does well, especially 2 quarters ahead.

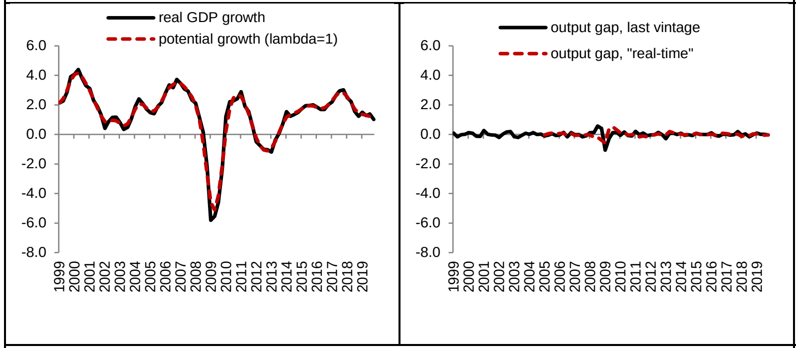
Inflation forecasting power of output gap estimates

method	1q rRMSE	2q rRMSE	4q rRMSE
HP_1	0.66	0.66	0.73
HP_1600	0.97	0.77	0.76
HP_5000	0.99	0.81	0.76
Blanchard-Quah	0.69	0.73	0.87
Hamilton	0.74	0.66	0.73
Beveridge-Nelson	0.69	0.61	0.71
Christiano-Fitzgerald	0.79	0.60	0.65
Jarocinksi-Lenza	0.68	0.56	0.69
SMS	0.86	0.72	0.69
UCM	0.77	0.61	0.81

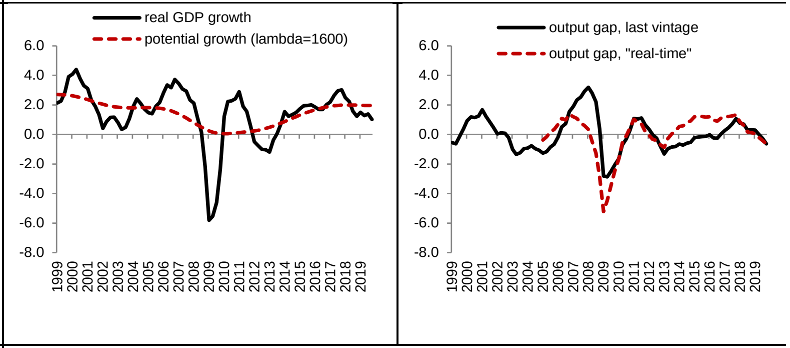
Note: lighter shades represent better inflation forecast, darker shades represent worse inflation forecast

An illustration with ten methods - plausibility

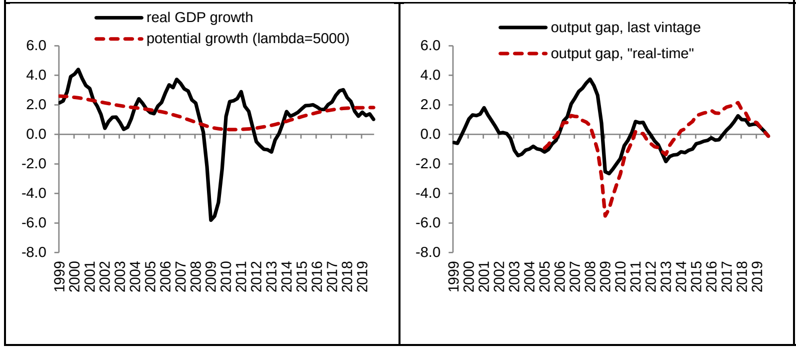
HP filter, lambda = 1



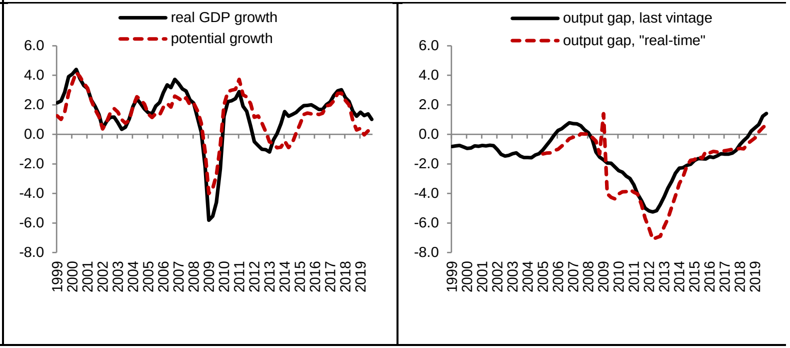
HP filter, lambda = 1,600



HP filter, lambda = 5,000

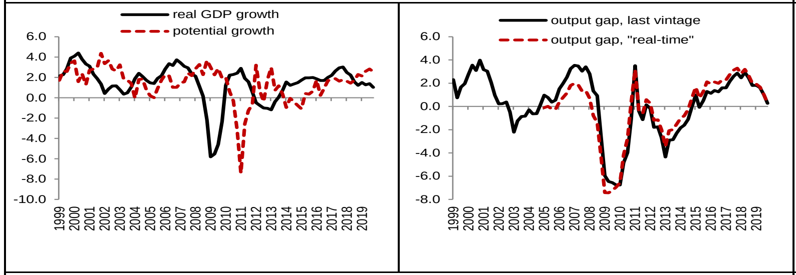


Blanchard-Quah

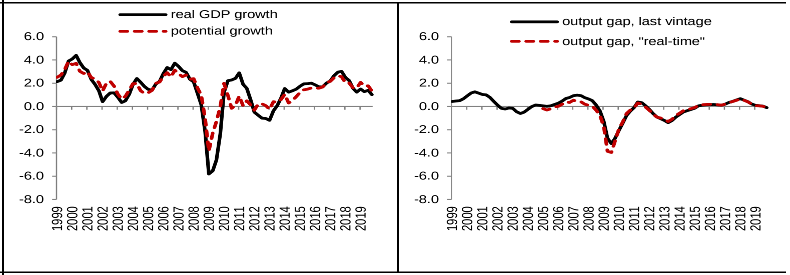


An illustration with ten methods - plausibility

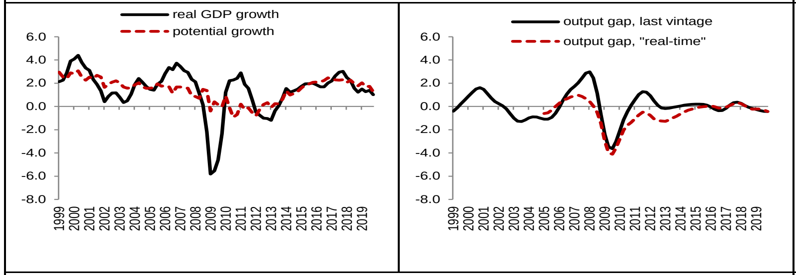
Hamilton



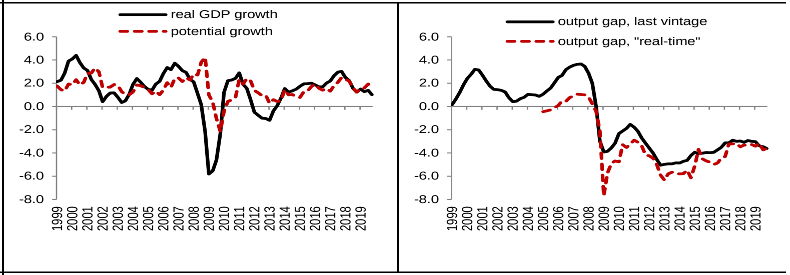
Beveridge-Nelson



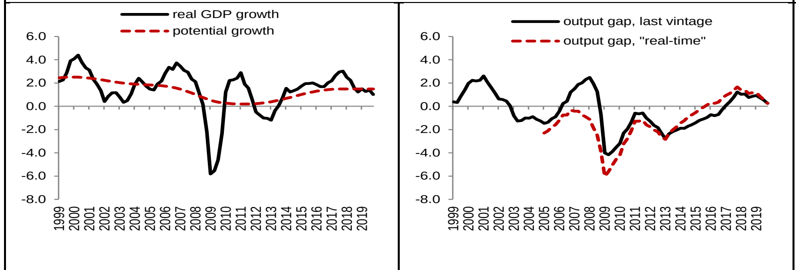
Christiano-Fitzgerald



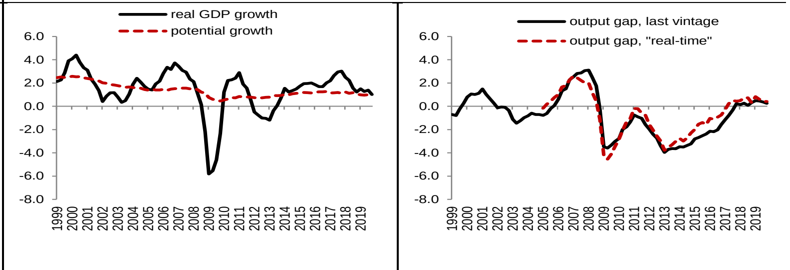
Jarocinski-Lenza



Survey-based Measure of Slack



UCM of Tóth



An illustration with ten methods – summary table

- Methods perform differently according to different criteria – no single “best”;
- The UCM provides solid performance according to all criteria;
- Plausibility also matters!
- Methods could be ranked – weighting the importance of criteria is subjective

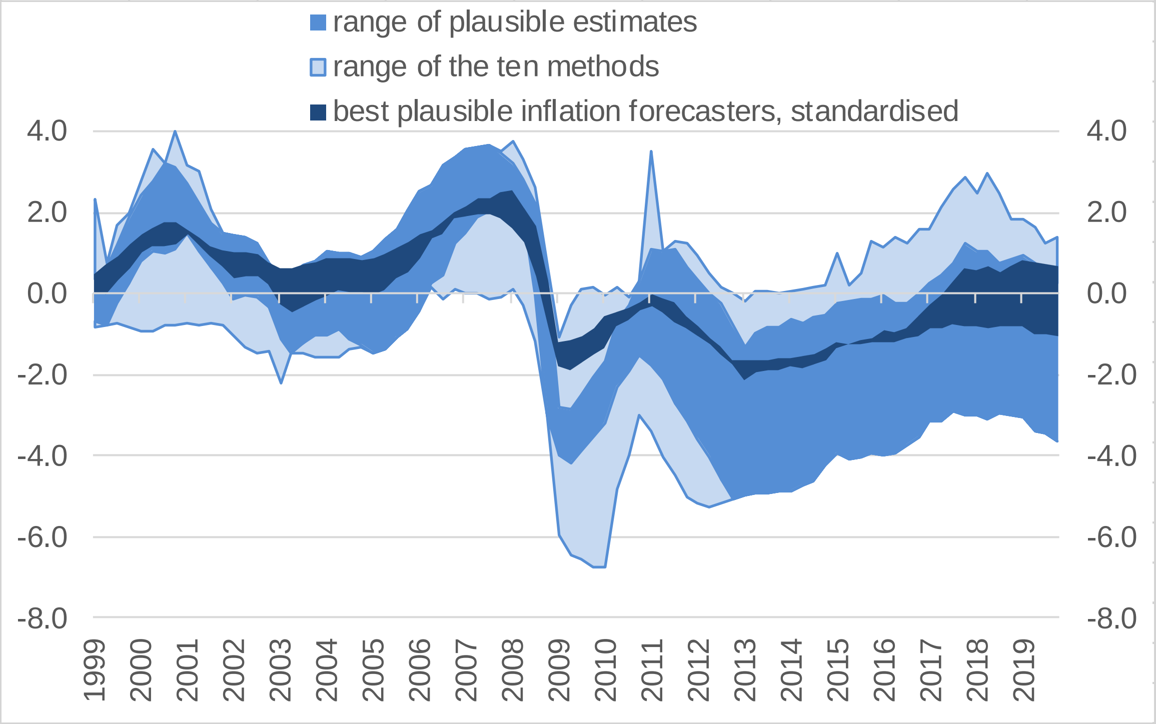
Ranking of the methods based on the assessment criteria

method	PC-relevant revisions	cyclicality	inflation forecast	symmetry	overall
HP_1	2	10	6	1	2
HP_1600	6	3	9	3	5
HP_5000	8	2	10	4	8
Blanchard-Quah	9	8	8	10	10
Hamilton	7	1	5	2	1
Beveridge-Nelson	1	9	4	5	2
Christiano-Fitzgerald	10	4	2	6	7
Jarocinksi-Lenza	5	6	1	9	5
SMS	4	7	7	7	9
UCM	3	5	3	8	2

Conclusions

- It is advised to follow and monitor a range of estimates, in addition to the headline figures (the BMPE in our case).

Range of euro area output gap estimates



Assessing the reliability of estimates by international institutions

- International institutions: European Commission, IMF, OECD
- Data: real-time vintage annual data from 2002/2004 to 2019 (EA, DE, FR, IT, ES, NL). Final estimate is Autumn 2019.
- Estimates of OECD revise the most, EC the least; smaller revisions post-GFC
- Decomposition:
  - Revision of the output gap in any year can be decomposed into:
    - revision of real GDP growth + revision of potential growth + revision of the initial OG

      → *Data revision + nowcast error*
  - Decomposition based on a BdE WP, but two major improvement:
    - Correct definition of "real-time" estimate, that correctly assigns the revisions to data revisions and forecast/nowcast errors
    - More precise decomposition by eliminating the revision of initial output gap / residual as much as possible

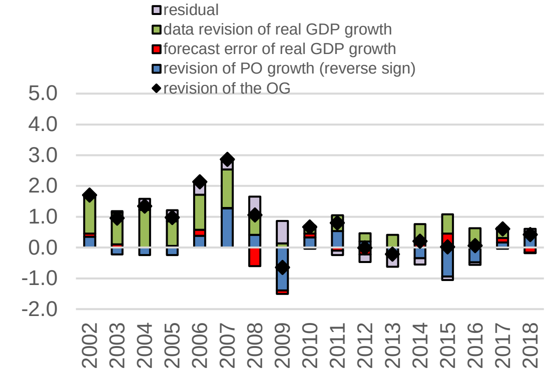
Assessing the reliability of estimates by international institutions - decomposition

- Data revisions do play a significant role;
- Potential growth was also revised strongly – especially by the OECD;
- The nowcast error of real GDP is fairly small.
- Some cross-country differences, but the main story is the same.

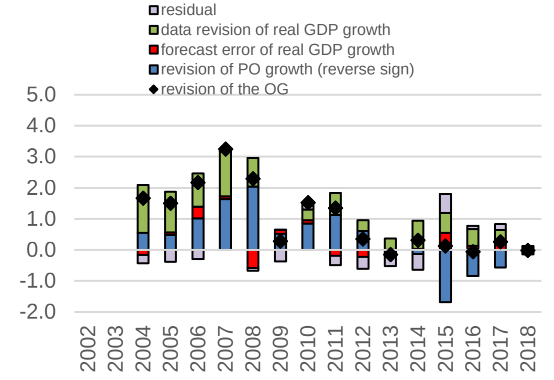
A decomposition of the revisions of the output gap

Euro area

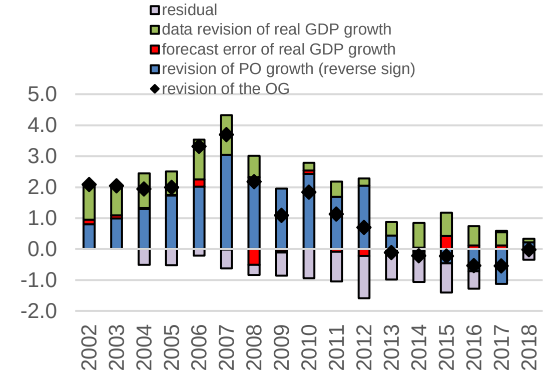
European Commission



IMF



OECD



Assessing the reliability of estimates by international institutions – PO revisions

- Again, data revisions have a sizeable and statistically significant role in explaining potential growth revisions;
- IMF and OECD: ~40% of the real GDP growth forecast error is transferred into potential growth.

Explaining the revision of potential growth

	all 3 institutions				EC				IMF				OECD			
	eq1	eq2	eq3	eq4	eq1	eq2	eq3	eq4	eq1	eq2	eq3	eq4	eq1	eq2	eq3	eq4
c	-0.583***	-0.579***	-0.405**	-0.433***	-0.336***	-0.340	-0.364	-0.397	-0.747**	-0.768***	-0.388	-0.300	-1.286***	-1.285***	-1.362***	-1.220***
	(0.143)	(0.143)	(0.168)	(0.135)	(0.236)	(0.237)	(0.312)	(0.245)	(0.290)	(0.282)	(0.313)	(0.239)	(0.184)	(0.186)	(0.192)	(0.190)
forecast error	0.385***	0.467***	0.779***	0.468***	0.253**	0.213	0.382	-0.038	0.423***	0.755***	1.189***	0.694***	0.396***	0.405***	0.385**	0.352**
	(0.063)	(0.102)	(0.153)	(0.123)	(0.106)	(0.173)	(0.292)	(0.232)	(0.117)	(0.191)	(0.274)	(0.206)	(0.084)	(0.131)	(0.172)	(0.170)
data revision	0.387***	0.392***	0.554***	0.270***	0.620***	0.620***	0.776***	0.429**	0.448**	0.493**	0.720***	0.167	0.448***	0.448***	0.533***	0.319***
	(0.094)	(0.094)	(0.096)	(0.094)	(0.151)	(0.152)	(0.173)	(0.163)	(0.198)	(0.194)	(0.195)	(0.179)	(0.114)	(0.115)	(0.098)	(0.113)
2007 dummy		0.773				-0.371				3.096**				0.092		
		(0.748)				(1.261)				(1.426)				(0.935)		
country f.e.	no	no	no	yes	no	no	no	yes	no	no	no	yes	no	no	no	yes
period f.e.	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes
adj. R-squared	0.171	0.171	0.284	0.605	0.206	0.196	0.131	0.539	0.173	0.220	0.344	0.686	0.308	0.298	0.575	0.652
no. of observations	225	225	225	225	75	75	75	75	65	65	65	65	75	75	75	75

Conclusions

- Data revisions do play a significant role in OG and PO growth revisions – there is not much we can do about it, other than acknowledging it;
- Real GDP forecast errors are often transferred into PO growth and OG revisions – we should be mindful about revising the history if we make a forecast error.
- Of course this is extremely difficult, especially in uncertain and volatile periods like large supply side shocks (Covid, war).

Thank you!

Evaluation criteria

- **Reliability**
  - Ideally, real time estimates should not deviate much from “final” estimates – otherwise wrong signal sent to policymakers in real time.
  - 4 qualitative criteria proposed:
    - Average revisions:
      - $Measure\ 1 = \frac{\sum_{j=1}^{k}\left(\hat{y}_{t-k+j-1}^{F} - \hat{y}_{t-k+j-1}^{j}\right)}{k}$
    - Absolute size of revisions:
      - $Measure\ 2 = \sqrt{\sum_{j=1}^{k}\left(\hat{y}_{t-k+j-1}^{F} - \hat{y}_{t-k+j-1}^{j}\right)^{2}}$
    - Correlation of “real-time” and “final” estimates:
      - $Measure\ 3 = \frac{\sum_{j=1}^{k}\left(corr(\hat{Y}^{F}, \hat{Y}^{j})\right)}{k}$
    - Frequency of sign change ($Measure\ 4$)

Evaluation criteria

- **Cyclicality**
  - Are trend and cycle independent? Statistical and economic explanations why this might not be the case:
    - Working age population and capital stock often not filtered;
    - Hysteresis;
    - Cyclicality of innovation and adoption of technologies.
  - What is the "ideal" degree of cyclicality of potential output?
    - We don't know (politically driven considerations depend on where we are in the cycle);
    - Does it really matter? (see slide in Motivation)
    - But it can be measured, and different estimates can be compared:
      - $Measure\ 5 = \sum_{t=1}^{k} \sqrt{\hat{y}_t^2}$
      - $Measure\ 6 = d\bar{y}_t = \beta_0 + \beta_1 \hat{y}_t + \epsilon_t$

Evaluation criteria

- **Symmetry**
  - It might be desirable, in particular for fiscal policy, that output gap estimates are symmetric over complete economic cycles. Can be related to a zero mean.
    - $Measure\ 7 = \frac{\sum_{t=1}^{k} \hat{y}_t}{k}$
- **Inflation forecasting**
  - In a central banking context, output gap estimates ideally provide information about the business cycle and inflationary pressures.
    - $Measure\ 8\text{:}\ \pi_t = \beta_0 + \beta_1 \pi_{t-1} + \beta_2 ip_t + \beta_3 slack_{t-k} + \varepsilon_t$
- **Plausibility**
  - Potential output and output gap should provide a meaningful narrative of trend developments, and economic and labour market slack.
    - Qualitative assessment

An illustration with ten methods - reliability

- On average, HP does not seem to revise much, but this masks larger revisions in different periods;
- More than 1pp average revisions of the J-L estimates;
- Revisions generally larger in the pre-2013 period.

Reliability of output gap estimates – average revisions of the output gap

method	2005q1 - 2019q4	2005q1 - 2013q1	2013q2 - 2019q4
HP_1	0.00	0.00	0.00
HP_1600	0.00	0.60	-0.73
HP_5000	0.10	1.12	-1.16
Blanchard-Quah	0.52	1.03	-0.11
Hamilton	0.07	0.59	-0.56
Beveridge-Nelson	0.13	0.26	-0.03
Christiano-Fitzgerald	0.61	0.96	0.18
Jarocinksi-Lenza	1.18	1.63	0.62
SMS	0.49	1.37	-0.58
UCM	-0.20	0.07	-0.53

Note: lighter shades represent more reliable estimates, darker shades represent less reliable estimates

An illustration with ten methods - reliability

- Absolute revisions provide a more accurate picture, but more difficult to interpret;
- J-L and smooth HP revise the most.

Reliability of output gap estimates – absolute revisions of the output gap

method	2005q1 - 2019q4	2005q1 - 2013q1	2013q2 - 2019q4
HP_1	0.16	0.16	0.04
HP_1600	0.96	0.82	0.50
HP_5000	1.19	0.96	0.70
Blanchard-Quah	0.83	0.80	0.20
Hamilton	0.78	0.70	0.33
Beveridge-Nelson	0.22	0.22	0.03
Christiano-Fitzgerald	0.76	0.74	0.20
Jarocinksi-Lenza	1.15	1.07	0.41
SMS	1.06	1.00	0.36
UCM	0.51	0.38	0.34

Note: lighter shades represent more reliable estimates, darker shades represent less reliable estimates

An illustration with ten methods - reliability

- J-L turns out to revise such that the revised estimate remains fairly strongly correlated with the original estimate
- Beveridge-Nelson, HP with lambda=1 and UCM are the most stable

Reliability of output gap estimates – correlation of “real-time” and “final” estimates

method	2005q1 - 2019q4	2005q1 - 2013q1	2013q2 - 2019q4
HP_1	98.31	97.05	99.87
HP_1600	96.29	94.30	98.85
HP_5000	95.12	93.09	97.76
Blanchard-Quah	92.37	86.97	99.23
Hamilton	96.00	93.56	99.08
Beveridge-Nelson	98.67	97.65	99.95
Christiano-Fitzgerald	89.18	82.74	97.03
Jarocinksi-Lenza	96.49	93.86	99.81
SMS	96.82	94.58	99.67
UCM	97.95	96.57	99.70

Note: lighter shades represent more reliable estimates, darker shades represent less reliable estimates

An illustration with ten methods - reliability

- B-Q, C-F change sign the most often;
- If an estimate is almost always negative/positive, it doesn't change sign often but still might revise a lot (J-L case in point).

Reliability of output gap estimates – frequency of sign change

method	sign change
HP_1	0.68
HP_1600	0.90
HP_5000	0.90
Blanchard-Quah	1.69
Hamilton	0.45
Beveridge-Nelson	0.17
Christiano-Fitzgerald	1.30
Jarocinksi-Lenza	0.34
SMS	0.85
UCM	0.56

Note: lighter shades represent more reliable estimates, darker shades represent less reliable estimates

An illustration with ten methods - symmetry

- Some estimates are symmetric by design and have a zero mean;
- Those that are not symmetric tend to have a negative mean over 1995-2019.

Symmetry of output gap estimates

method	mean	max	min	std. dev.	skewness	kurtosis	Jarque-Bera	probability
HP_1	0.000	0.571	-1.065	0.172	-2.023	17.332	924.1	0.000
HP_1600	0.000	3.203	-2.867	1.118	0.386	3.903	5.9	0.053
HP_5000	0.000	3.742	-2.653	1.275	0.745	3.734	11.5	0.003
Blanchard-Quah	-1.508	0.791	-5.247	1.329	-0.902	4.143	19.0	0.000
Hamilton	0.246	3.988	-6.766	2.415	-1.084	4.163	25.2	0.000
Beveridge-Nelson	0.011	1.265	-3.196	0.802	-1.643	6.711	102.4	0.000
Christiano-Fitzgerald	-0.097	2.988	-3.628	1.182	-0.161	4.126	5.7	0.058
Jarocinksi-Lenza	-1.097	3.648	-5.049	2.656	0.245	1.704	8.0	0.018
SMS	-0.237	2.614	-4.164	1.519	-0.334	2.797	2.0	0.362
UCM	-0.758	3.164	-3.930	1.689	0.222	2.690	1.2	0.542

Note: lighter shades represent less cyclical estimates, darker shades represent more cyclical estimates

Assessing the reliability of estimates by international institutions

- International institutions: European Commission, IMF, OECD
- Data: real-time vintage annual data from 2002/2004 to 2019 (EA, DE, FR, IT, ES, NL). Final estimate is Autumn 2019.
- Estimates of OECD revise the most, EC the least; smaller revisions post-GFC

Revision of output gaps (common sample)

	European Commission			**IMF**			**OECD**		
	2004-2018	2004-2012	2013-2018	2004-2018	2004-2012	2013-2018	2004-2018	2004-2012	2013-2018
	euro area								
average revisions	0.7	1.0	0.2	1.0	1.6	0.1	1.1	2.0	-0.3
absolute revisions	4.3	4.3	0.8	5.5	5.5	0.4	6.7	6.6	0.8
correlation	0.941	0.904	0.998	0.879	0.802	0.995	0.831	0.721	0.995
	Germany								
average revisions	0.2	0.0	0.5	0.5	0.9	-0.1	0.6	0.9	0.1
absolute revisions	3.5	3.1	1.6	3.9	3.8	0.6	4.5	4.1	1.8
correlation	0.973	0.963	0.988	0.946	0.926	0.975	0.845	0.774	0.952
	France								
average revisions	1.3	1.8	0.6	1.6	1.9	1.2	1.5	2.3	0.3
absolute revisions	6.3	6.1	1.6	7.0	6.1	3.4	7.5	7.4	1.5
correlation	0.796	0.664	0.993	0.803	0.722	0.924	0.846	0.751	0.990
	Italy								
average revisions	1.0	1.8	-0.1	1.4	2.2	0.2	1.2	2.3	-0.5
absolute revisions	6.6	6.5	1.0	7.2	7.1	0.9	8.1	8.0	1.5
correlation	0.822	0.709	0.991	0.580	0.325	0.963	0.855	0.764	0.992
	Spain								
average revisions	0.3	1.2	-0.9	1.1	2.7	-1.3	0.1	1.7	-2.4
absolute revisions	8.8	7.9	3.8	11.7	10.9	4.3	11.1	8.2	7.5
correlation	0.799	0.668	0.994	0.615	0.368	0.986	0.736	0.565	0.993
	The Netherlands								
average revisions	0.4	0.7	0.1	1.0	1.1	0.8	1.3	2.1	0.1
absolute revisions	3.0	2.9	0.8	5.2	3.9	3.4	7.0	6.8	1.8
correlation	0.977	0.965	0.995	0.938	0.945	0.928	0.950	0.924	0.988

Assessing the reliability of estimates by international institutions - decomposition

Revision of the initial output gap (t=0):

$$rev(\hat{y}_0) = \hat{y}_0^F - \hat{y}_0^{RT} = \frac{y_0^F - \bar{y}_0^F}{\bar{y}_0^F} - \frac{y_0^{RT} - \bar{y}_0^{RT}}{\bar{y}_0^{RT}} \approx (\ln y_0^F - \ln \bar{y}_0^F) - (\ln y_0^{RT} - \ln \bar{y}_0^{RT})$$

Similarly, for t = 1:

$$rev(\hat{y}_1) = \hat{y}_1^F - \hat{y}_1^{RT} = \frac{y_1^F - \bar{y}_1^F}{\bar{y}_1^F} - \frac{y_1^{RT} - \bar{y}_1^{RT}}{\bar{y}_1^{RT}} \approx (\ln y_1^F - \ln \bar{y}_1^F) - (\ln y_1^{RT} - \ln \bar{y}_1^{RT})$$

Adding and subtracting $\ln y_0^F, \ln y_0^{RT}, \ln \bar{y}_0^F, \ln \bar{y}_0^{RT}$

$$rev(\hat{y}_1) = [(\ln y_1^F - \ln y_0^F) - (\ln y_1^{RT} - \ln y_0^{RT})] - [(\ln \bar{y}_1^F - \ln \bar{y}_0^F) - (\ln \bar{y}_1^{RT} - \ln \bar{y}_0^{RT})] + (\ln y_0^F - \ln \bar{y}_0^F) - (\ln y_0^{RT} - \ln \bar{y}_0^{RT})$$

I.e.: $rev(\hat{y}_1) = [(\Delta \ln y_1^F - \Delta \ln y_1^{RT}) - (\Delta \ln \bar{y}_1^F - \Delta \ln \bar{y}_1^{RT})] + rev(\hat{y}_0)$

Following the same logic for any t>0 until the last but one vintage:

$$\underbrace{rev(\hat{y}_t)}_{\text{Revision of output gap}} = \underbrace{\sum_{j=1}^{t} (\Delta \ln y_j^F - \Delta \ln y_j^{RT})}_{\text{real GDP growth}} - \underbrace{\sum_{j=1}^{t} (\Delta \ln \bar{y}_j^F - \Delta \ln \bar{y}_j^{RT})}_{\text{potential growth}} + \underbrace{rev(\hat{y}_0)}_{\text{initial output gap}}$$

Further decomposition of the revisions of real GDP growth:

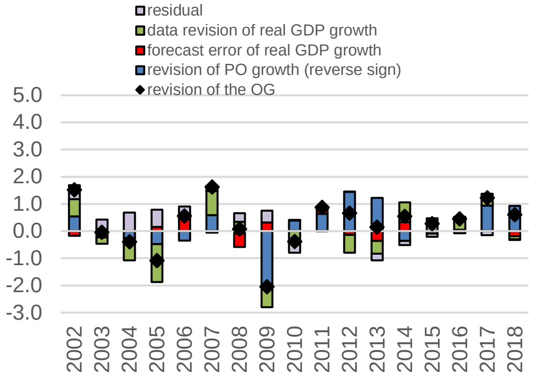
$$\underbrace{\Delta \ln y_t^F - \Delta \ln y_t^{RT}}_{\text{Revision of real GDP growth}} = \underbrace{(\Delta \ln y_t^F - \Delta \ln y_t^{t-1})}_{\text{Data revision}} + \underbrace{(\Delta \ln y_t^{t-1} - \Delta \ln y_t^{RT})}_{\text{Nowcast error}}$$

Assessing the reliability of estimates by international institutions - decomposition

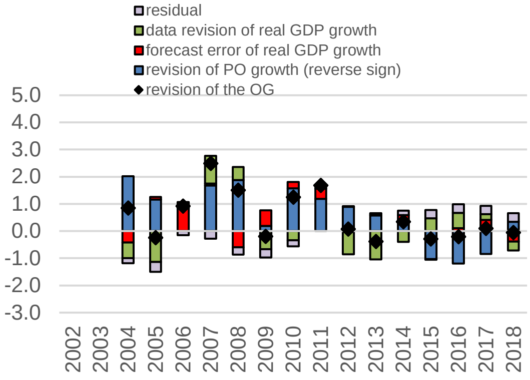
A decomposition of the revisions of the output gap

Germany

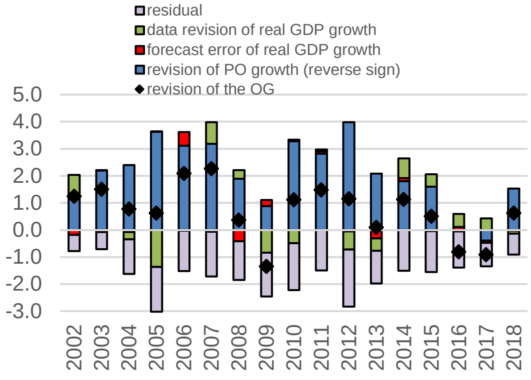
European Commission



IMF

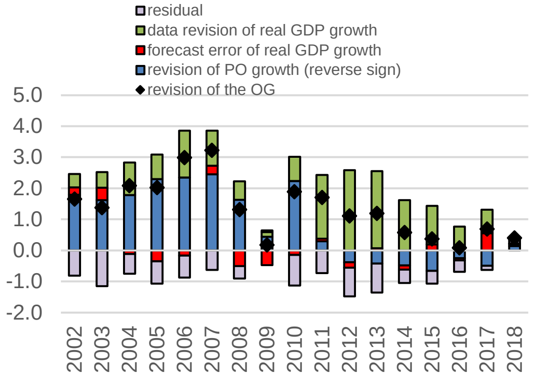
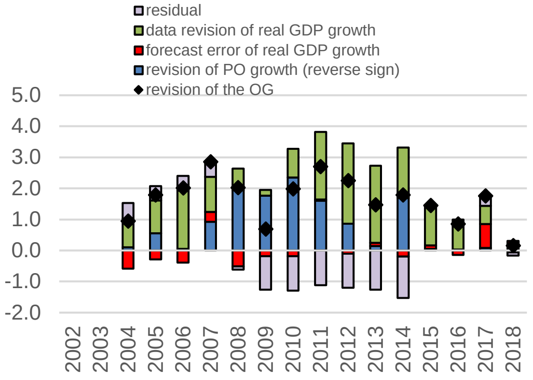
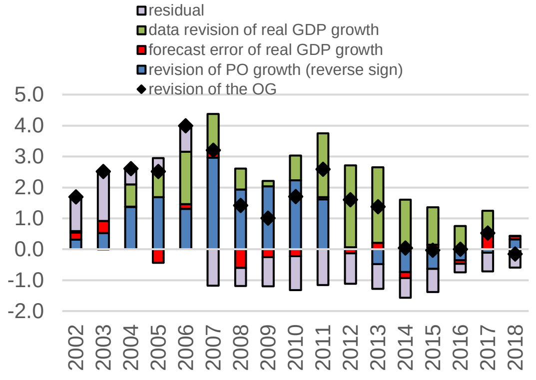


OECD



Assessing the reliability of estimates by international institutions - decomposition

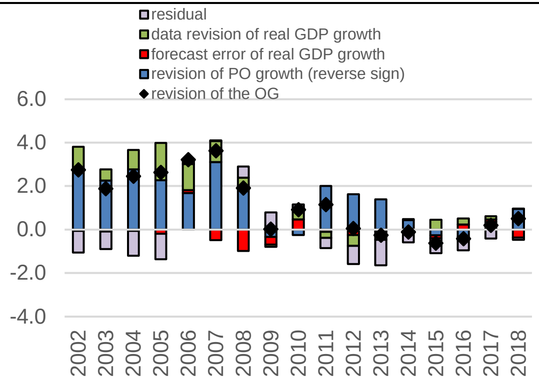
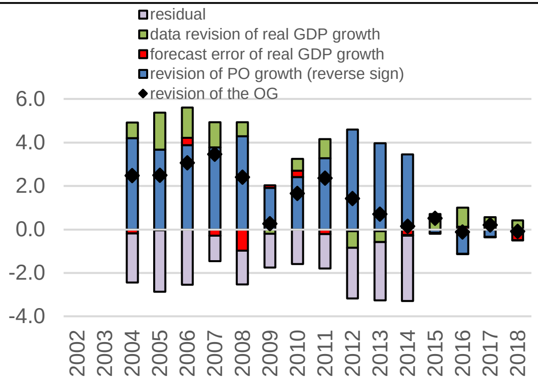
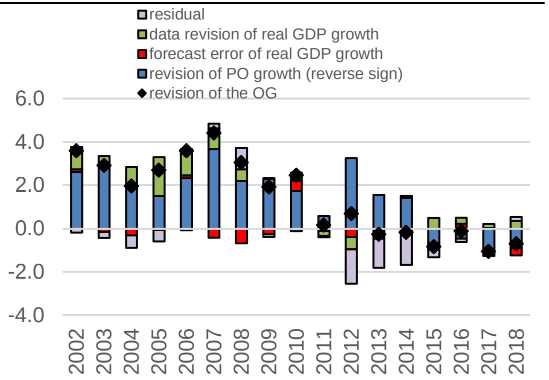
A decomposition of the revisions of the output gap

France		
European Commission	***IMF***	***OECD***
		

Assessing the reliability of estimates by international institutions - decomposition

A decomposition of the revisions of the output gap

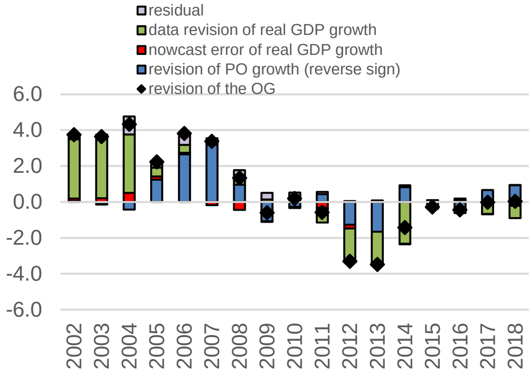
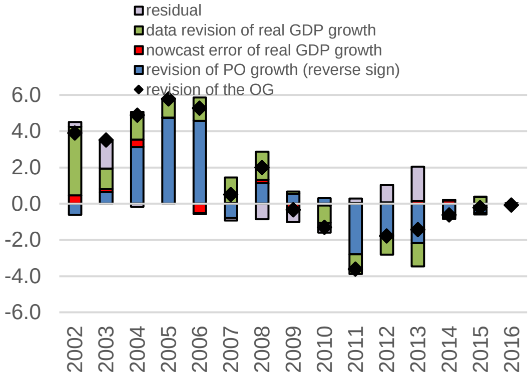
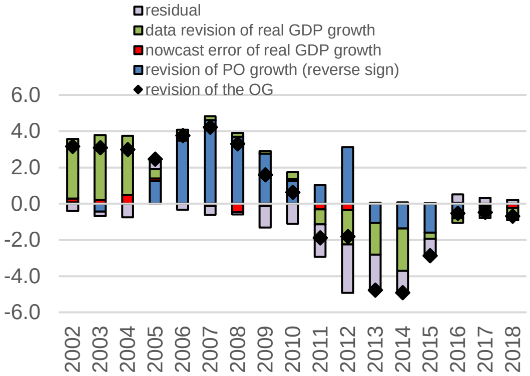
Italy

European Commission	*IMF*	*OECD*
		

Assessing the reliability of estimates by international institutions - decomposition

A decomposition of the revisions of the output gap

Spain

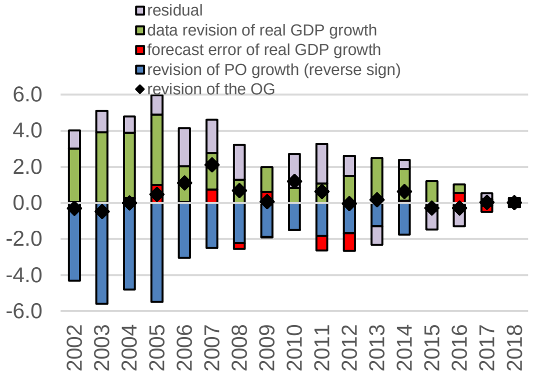
European Commission	*IMF*	*OECD*
		

Assessing the reliability of estimates by international institutions - decomposition

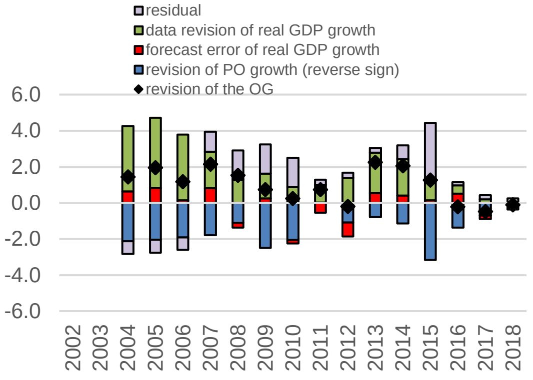
A decomposition of the revisions of the output gap

The Netherlands

European Commission



IMF



OECD