ARTICLE

Graeme Chamberlin
Office for National Statistics

Flash estimates of European labour costs

SUMMARY

Flash estimates are early or advance estimates of statistics, usually based exclusively or partly on forecasts. They are demanded by policy-makers and analysts who have to make decisions in real time and before enough information is collected to publish the data conventionally. As such there is a trade-off between timeliness and accuracy. This article outlines the approach taken by the Office for National Statistics (ONS) to produce flash estimates of European labour costs, as part of a wider Eurostat-led project on flash estimates in Europe. A general-to-specific methodology is used to select the best combination of indicators to use in a forecast model. Flash estimates are presented for the groups of Euro-Zone 15 and European Union 27 countries.

Introduction

Flash estimates are early or advance estimates of statistics. As they are usually produced before a full set of information is available they require, to some extent, forecasts or imputations. This reflects the timeliness versus accuracy trade-off that exists in the production of most economic data. For this reason there is often little difference between a forecast of a particular estimate and a flash estimate – all that is important is that they are produced more quickly than the data would otherwise be available.

Demand for flash estimates comes from policy-makers and analysts who have to make quick decisions in real-time. For example, conventional wisdom is that it takes two years for an interest rate change to fully feed through the economy, in which case policy has to be set pre-emptively. If economic data were only available with a significant lag then policy-makers would be faced with making these decisions in the dark. For them it is better to have some data in a timely fashion, even if it is based on more limited information, than none at all.

This article describes investigative work undertaken by the Office for National Statistics (ONS), as part of a wider project based at Eurostat, to look at producing flash estimates of European labour costs. At present these are available 85 days after the end of the reference quarter. The aim is to produce an estimate within 40 days using available information. To achieve this, a general-to-specific (GETS) methodology is applied to model the European Labour Cost

Index (LCI). Flash estimates of this variable are then formed from the one-step-ahead forecasts of these models. So for example, to obtain a flash estimate for 2008Q2 would see the LCI modelled up to 2008Q1, and then a one period ahead forecast produced using this model.

In estimating the European LCI a large number of potentially relevant indicators are available. The general problem is therefore one of model selection. Given that a small number of variables can generate a large number of different models, the researcher faces high search costs in trying to identify the optimal one from among all the possibilities. Various techniques are available to make this problem more tractable. Cluster analysis, principal components and dynamic factor models all work on the basic principle that the information in a large number of indicators can be adequately represented by a subset of those variables or a smaller number of factors. If there is strong correlation between the competing variables, these techniques can be very effective in reducing the set of possible indicators to a more manageable size without losing too much information.

The GETS modelling approach essentially works the other way around. Instead of trying to limit the number of explanatory variables, the starting point is a very general and over-fitted dynamic statistical model. Standard testing procedures are then used to reduce its complexity by eliminating statistically insignificant variables and hone in on the true underlying data generating process. Proponents of GETS would argue

that by commencing with an over-specified model they are less likely to discard potentially useful information.

However, the order in which the variables are deleted generally matters, so the final model is dependent on the path taken to get there, and a multitude of terminal models can result from the same starting point. Traditional argument against GETS modelling therefore focus on the high 'search costs' involved in finding the optimal model. Fortunately recent advances in econometrics software have reduced search costs by automating the GETS procedure enabling multiple deletion paths to be explored simultaneously and using encompassing tests to select between different terminal models. One such example, and used in this article, is the PcGets software by Hendry and Krolzig (2001). Without doubt the automation of GETS modelling has significantly reduced search costs, but one of the key conclusions of this study is that its effectiveness is still limited by a number of factors.

For the purposes of flash estimation the most important feature of models is their out-of-sample performance. In general, it is relatively easy to fit a model to past data that achieves a satisfactory goodness of fit if appropriate indicators are available. However the past does not necessarily predict the future and it is easy to over-fit a model. A particular indicator may achieve significance by accounting for an idiosyncratic feature of past data, but if that feature is not repeated in the future then that variable will become redundant and compromise the model's forecast accuracy. So deriving a model that is good for forecasting requires a subtle balance between explaining the past data well, but not so well that it is an over-specification of the true data generating process.

Flash estimates of the LCI presented in this article for the Euro-Zone 15¹ (EZ-15) and European Union 27² (EU-27) are the one-step-ahead forecasts of the models estimated by the GETS procedures. Timeliness of data is therefore an important issue. Flash estimates can only be formed using the indicators that are available at the time of the forecast which often means lagged and not contemporaneous data are considered. For each model, out-of-sample performance and hence the accuracy of the flash estimate can be assessed in the conventional way by comparing the one-step-ahead forecasts with the actual data outturn.

Here there is a slight problem with a small sample. LCI data is available for the period 1996Q2–2008Q2, but considering

that a four quarter lag of the data could be a plausible indicator reduces the effective sample to 1997Q2–2008Q2. With only 44 data points it is harder to use an adequate number of observations to fit a good model and have sufficient left-over for the purpose of testing its forecast potential. This problem is particularly acute with GETS procedures because the starting point is a deliberately over-specified model which further tightens the available degrees of freedom. Despite this issue, one-step-ahead forecasts/flash estimates are presented for the period 2003Q2–2008Q2.

The rest of the article proceeds as follows. The next section outlines the GETS methodology and describes how PcGets resolves some of the historical problems affecting its use. The third section, taking an economics perspective, identifies a set of indicators that might be potentially important determinants of European labour costs. These include both 'hard' or official data sources, and 'soft' data which consists of non-official consumer and business surveys. The main results are presented in the fourth section consisting of flash estimates of the LCI for both the EZ-15 and EU-27 country groups from various different data sources. Finally, some concluding comments are offered.

PcGets: Improving the efficiency of GETS modelling

The standard GETS procedure is to delete the variable with the lowest t-ratio at each stage until a model is arrived at where all the remaining variables have achieved a given level of significance. However, this is not the only possible selection path. GETS procedures typically suffer from path dependence which is where the results obtained depend on the simplification sequence adopted.

The costs of search are the additional mistakes introduced by commencing from an initial model that is larger than the data generating process. They comprise any retained irrelevant variables which are significant by chance, as well as omitting relevant variables.

High search costs can easily be understood from the theory of repeated testing. Conducting 20 independent tests at a 5 per cent significance level should on average lead to one significant rejection by chance (a type 1 error). As the model selection process often requires many tests it may seem that spurious results will be obtained with high probability. In fact the probability that none of the tests rejects at 5 per cent is $(1-0.05)^{20} = 0.36$. A type 1 error

is the probability of rejecting a hypothesis that is true. In this case it is the probability of accepting variables that are insignificant, that is maintaining irrelevant variables in the regression. Viewed as a type 1 error $1-0.36=0.64$ is quite large.

Failing to reject irrelevant variables means that they may stay in the regression and act as proxies for variables that do matter, which are subsequently omitted. A possible solution to combating the accumulation of type 1 errors is to raise the size of the test by using larger critical values. For example, for 40 tests the probability that no test rejects by chance at the 5 per cent level is $P(0.05)=0.13$, but using a 0.5 per cent significance level raises this to $P(0.005) = 0.89$. Here the consequence of increasing the size of the test is to lower the probability of type 1 errors from 0.87 to 0.11.

Unfortunately more stringent criteria for avoiding rejections when the null is true (a variable is insignificant) lowers the power of rejection when it is false (a variable is significant). That is, in attempting to lower the probability of maintaining irrelevant variables by raising critical values we increase the chance probability of rejecting relevant ones (a type 2 error). The size versus power trade-off is a well-known phenomenon in econometric modelling.

Therefore, under repeated testing the probability of retaining variables that should not enter a relationship would be high because a multitude of tests on irrelevant variables must deliver some significant outcomes by chance. This indicates how the results obtained from a particular starting model need not be independent of the modelling procedure. Due to path dependence a multitude of terminal models can result from the same starting point.

The main advancement in automated GETS procedures such as PcGets is to reduce the search costs associated with exploring multiple deletion paths from a general model and choosing between alternative terminal models. Starting from the general unrestricted model (GUM) the PcGets algorithm then works like a series of sieves, searching multiple deletion paths, checking that congruence is maintained at each stage and then using encompassing tests to select between terminal models.

Searching many reduction paths is designed to mitigate the problem of path dependence, leading to misspecification as important variables are deleted and irrelevant variables are retained as proxies. Exploring several paths gives the opportunity for error correction in the

light of wrong decisions by increasing the probability that some models will retain the variables that matter while eliminating those that do not. Congruence is maintained as the algorithm only undertakes reductions which do not lead to failures in diagnostic tests.

Each search path is terminated where there are no further possible reductions or when deletion induces a diagnostic test failure. Encompassing is the notion of being able to account for the results obtained by rival models given one's own findings. Therefore, if model A encompasses model B then model A accounts for all of the variance in the dependent variable explained by model B. In this sense encompassing implies variance-dominance, so a badly-fitting model cannot account for the variance of a well-fitting model.

In the encompassing stage of the PcGets algorithm all distinct non-nested models are collected and encompassing is used to eliminate those which are dominated. If a unique choice does not result it implies that the remaining models are incomplete, so that each explains some variance in the dependent variable not accounted for by other models, but no model is dominant. The PcGets algorithm then forms the union of resulting models which becomes the new starting point for path searches. The algorithm repeats until the union is unchanged between successive rounds. The final model, if it is not already unique is then selected by information criterion.

Simply choosing the best-fitting terminal model based on adjusted R^2 or other information criterion offers no protection against picking a spurious relationship. When a given path eliminates a variable that matters, other variables proxy such an effect leading to spuriously large and misspecified models. However, some other paths will retain that variable and in the encompassing tests the proxies will be frequently revealed as conditionally redundant inducing a smaller final model focussed on the genuine causal factors. This is potentially very important as over-specified models may be more prone to forecast failure.

An innovation in PcGets is the use of sub-sample information to check constancy of parameters and model specification. In this case it provides further information on overall significance. Since a central t-test wanders around the origin, the probability is low that an effect which is significant by chance in the full sample will be significant in two sub-samples. This information can be used to refine the detection of spurious

regressors without losing too much power. Again, models that consist of indicators that are stable in their significance are more likely to have better out-of-sample forecasting performance.

Although PcGets is an automatic procedure there is still a vital role to be played by the practitioner. First in choosing the significance levels of the variable deletion and diagnostic tests, which acts as constraints on the paths the algorithm explores and therefore has an important bearing on the terminal models produced. If faced with a relatively large number of indicators and the aim is to produce models based on a small number of variables, as in this case, paths can be set up to minimize non-deletion. Second, the researcher is responsible for selecting the initial set of indicators and transformations that form the GUM which is the starting point of all path searches. Here economic theory may guide on the selection of relevant variables. The next section identifies possible determinants of labour costs from an economic theory standpoint.

At this point it is worth noting that although PcGets dramatically lowers the search costs that have historically hampered the use of GETS modelling it is not a complete solution. Two problems remain which have compromised its operational effectiveness in this work. First, all models start from a GUM which must have sufficient degrees of freedom to be estimated. When the set of indicators, including lags, approaches the number of observations the GUM cannot be estimated. Castle and Hendry (2006) suggest dividing the indicator set into smaller groups and collecting the surviving indicators as the basis for a new GUM can, according to their Monte Carlo analysis, get close to recreating the true data generating process when there are more indicators than observations. This approach is adopted here. Second, multicollinearity in the GUM seems to reduce the efficiency of the PcGets algorithm, leading to over-specified terminal models. This problem is most pertinent in the consumer and business surveys where many variables are highly correlated. Therefore, the allocation of indicators to each subgroup should be motivated by the aim of reducing multicollinearity in the GUM.

Possible indicators: the economic determinants of labour costs

The empirical success of PcGets depends crucially on the creativity of the researcher

in specifying the general model and the feasibility of estimating it from the available data. Economic theory could guide on the inclusion of relevant variables and the exclusion of irrelevant effects. Also data transformations are up to the researcher so that appropriate specifications are inputted into the model.

For each country the labour cost index (LCI) consists of a wage (W) and non-wage (NW) components. Non-wage costs mainly refer to social security contributions but also include the differences between taxes and subsidies employers receive from employing labour.

$$LCI_t = W_t^\alpha NW_t^{1-\alpha}$$

Wages and salaries typically account for over three-quarters of the index. As seen in **Figures 1** and **2** for the EZ-15 and EU-27 country groups, the quarterly growth rate of labour costs (seasonally adjusted) closely follows the trend in the wages and salaries component. In previous work it was discovered that no significant improvement in fit could be achieved by modelling the two components separately. This is a convenient finding, because while there are a broad number of available indicators on wage costs and its determinants, non-wage costs tend to be more 'institutionally' determined and harder to effectively model.

In selecting variables timeliness is an important issue. All potential indicators are required comfortably before the T+40 target. Most of the data from official sources do not meet this requirement so can only be used as lagged variables. Business and consumer surveys though are more timely, and these can enter the regression models as contemporaneous relationships.

The use of official and non-official data, or 'hard' and 'soft' data is also an important consideration. Official or hard data is provided to Eurostat by national statistics institutions, and as such should be reliable in terms of its publication frequency and the methods used to construct it. Non-official or soft data refer to the consumer and business surveys published by trade groups and other various associations. This data tends to be timely, and many surveys are well-established, but it is also based on smaller samples and tends to be qualitative. The use of business and consumer survey data in official statistics, even for the purposes of forecasting, is not an unanimously accepted convention. For this reasoning, this study has a preference for

Figure 1
Wage and non-wage components of the EZ-15 LCI

Per cent, quarter on quarter growth

Source: Eurostat PEEI

Figure 2
Wage and non-wage components of the EU-27 LCI

Per cent, quarter on quarter growth

Source: Eurostat PEEI

using hard data, but also looks at whether or not soft data improves the outcome.

Direct data on labour costs

Apart from the Labour Cost Index itself, many countries publish official data on total wages and salaries and on the compensation of employees (CoE). These are standard Principal European Economic Indicators (PEEIs) from the national accounts and labour market and published on the Eurostat website at a main industry disaggregation. Using corresponding data on employees these can be transformed into average wages or compensation figures. Data on hourly earnings are also available from the OECD's Monthly Economic Indicators (MEI) database, usually for the manufacturing and private sectors of the economy. These data are available for both European aggregates and the major individual European countries.

Indirect data on the determinants of wages

Factors that determine wages could be, albeit indirectly, significant indicators of movements in labour costs. Standard economic theory argues that wages are the outcome of a bargain between employers and workers (who may bargain collectively as unions).

Workers are expected to bargain in terms of their expected real wage (W_t/P_t^e) where W_t are nominal wages and P_t^e the expected price level.

$$\frac{W_t}{P_t^e} = A_t - bu_t$$

These are assumed to be a positive function of a set of autonomous factors (A_t) such as the level of unemployment benefits, trade union power, wage and employment legislation and so on, and negatively related to unemployment (u_t).

Workers are assumed to respond to expected higher prices by asking for compensating wage increases to maintain the real value of earnings. As price expectations are often governed by current prices, official data on consumer prices such as the GDP deflator, consumption deflator, import deflator and the harmonised index of consumer prices (HICP) are possible indicators. These prices data are all official PEEIs. There is also a multitude of information from consumer and business surveys on price expectations.

The tightness of the labour market is also a factor, with workers likely to exert greater bargaining power when available labour is scarce. The opposite would be true in times of higher unemployment, when workers might moderate demands in order to protect jobs. Therefore official data on unemployment is widely considered to be a sensible indicator. This could be

supplemented with unofficial data from business surveys on capacity utilisation, productive capacity, labour shortages and employment intentions.

It is difficult to find many timely and readily available indicators for the more 'structural' factors, especially pertaining to legislation and trade union power. However, one set of useful information might come from business survey data on the competitiveness situation facing firms (published as stemming from domestic, EU and outside of the EU sources). Those facing less competitive pressure may have greater scope to concede to wage demands and pass the costs onto consumers.

In bargaining models firms are typically assumed to set prices as a mark-up over labour costs.

$$P_t = (1 + \mu) \frac{W_t}{MPL_t}$$

where μ is a mark-up and MPL_t is the marginal productivity of labour. This means firms can pay higher wages without changing prices if labour becomes more productive. An implicit wage equation can be found by re-arranging so that:

$$\frac{W_t}{P_t} = \frac{MPL_t}{(1 + \mu)}$$

So the real wages that firms are prepared to pay respond positively to labour productivity. An estimate of labour productivity can be proxied using data on (real) output per head (or real Gross Value Added (GVA) per head). GVA and employment are also official PEEIs.

The complete set of indicators used to model European labour costs, divided into hard and soft categories, are listed in **Table 1**. Lags of the dependent and independent variables can pick up valuable information of incomplete adjustments, inertia, transitional effects, adaptive expectations formations, wage and price rigidities and so on, and are also considered in the modelling process. One of the advantages of using PcGets is the determination of suitable lags in the simplification of dynamic regression models as economic theory often poses little evidence for specifying lag lengths in statistical models.

Results: flash estimates of the European LCI

As a first step, using the full sample 1997Q2–2008Q2, three models are estimated with PcGets for the LCIs of the EZ-15 and EU-27 country groups. The baseline model is a simple autoregressive

model, using only lags of the LCI along with appropriate dummy variables to correct for outliers. The hard data model is based on the set of official data in Table 1. Finally the soft data model uses the unofficial survey data also listed in Table 1.

Table 1
Indicators of European Labour Costs

Hard data

Eurostat Principal European Economic Indicators (consumer prices, labour market and national accounts)

- Labour Cost Index (C–K: production and non-government services)
- Gross Value Added (C–K): proxy for labour productivity (real GVA / total employment)
- Compensation of Employees (C–K): Average compensation per employee (total CoE / total employees)
- Wages and Salaries (C–K): Average wages per employee (total W&S / total employees)
- Unemployment
- GDP deflator
- Consumption deflator
- Import deflator
- Consumer prices (HICP)

Earnings data from OECD Monthly Economic Indicators (MEI) database

- EU manufacturing
- Euro-zone manufacturing
- France private sector
- Germany private sector
- UK private sector
- Euro-zone private sector

Soft data

Eurostat Principal European Economic Indicators (business and consumer surveys)

- Industry employment expectations – months ahead
- Industry production expectations – months ahead
- Industry selling prices expectations – months ahead
- Industry productive capacity
- Industry capacity utilisation
- Industry domestic competition
- Industry EU competition
- Industry external (outside EU) competition
- Industry labour shortage limiting production
- Construction employment expectations – months ahead
- Construction price expectations – months ahead
- Construction labour shortage limiting production
- Retail employment
- Consumer price trends – last 12 months
- Consumer price trends – next 12 months
- Consumer unemployment expectations – next 12 months
- Services employment – recent months
- Services employment – next months

Sources: Eurostat, OECD

Table 2
Fitted models of the European LCI using PcGets, 1997 Q2 – 2008 Q2

Model	EZ-15	EU-27
Baseline	Constant Labour Cost Index (–1)	Labour Cost Index (–2), (–3), (–4) Dummy 1998Q1
Hard	Labour Cost Index (–1) GDP deflator (–2)	Labour Cost Index (–2) Manufacturing earnings- EU (–3) Dummy 1998Q1
Soft	Labour Cost Index (–3) Industry Productive Capacity Construction Employment - next months (–1) Services Employment - next months (–1)	Constant Labour Cost Index (–1), (–3), (–4) Industry competition domestic Industry prices- next months (–1) Industry labour shortage Dummy 1998Q1

Source: Author's estimations

Figure 3
EZ-15 models of the Labour Cost Index

Per cent, quarter on quarter growth

Source: Author's estimations

Figure 4
EU-27 models of the Labour Cost Index

Per cent, quarter on quarter growth

Source: Author's estimations

Figure 5
EZ-15 one step a head forecasts (2003 Q2 – 2008 Q2)

Per cent, quarter on quarter growth

Source: Author's estimations

Figure 6
EU-27 one step a head forecasts (2003 Q2 – 2008 Q2)

Source: Author's estimations

fitted values from each regression against the respective published LCI.

Mean squared errors (MSE), which are the average squared difference between the fitted values of each model and the actual LCI data gives a rudimentary goodness of fit statistic. These are presented in **Table 3** below. Using hard data does not appear to significantly improve the model fit for either the EZ-15 or the EU-27 LCI. However, soft data models are more successful in this regard.

The main focus of interest though is how these model strategies perform out-of-sample as this is more representative of how the LCI flash estimate would be produced. A truer approach would be to use 'real-time' data which is the unrevised data that the researcher would see at the time of the forecast. Unfortunately a complete real-time database for the PEEIs was unavailable, so one-step-ahead forecasting was the closest this study could come to replicating real-time conditions.

Flash estimates are formed by re-estimating the model each quarter and then generating a one-step-ahead forecast. For both the EZ-15 and EU-27 country groups, recursive estimation of the baseline, hard data and soft data models generated flash estimates of the LCI for the period 2003Q2–2008Q2. In the course of this strategy, PcGets would often update its best model. Table 4 shows the evolution of indicator selection across the different models over time. Each model represents the optimal indicator selection made by PcGets when the LCI was estimated on that particular sample (**Table 4**).

Plots of the out-of-sample forecasts/Flash estimates relative to the actual LCI data are presented in **Figure 5** for the EZ-15 and

Figure 6 for the EU-27. Corresponding measures of mean-squared (forecast) errors are displayed in **Table 5**. The results indicate that the hard data models don't offer any improvement over the baseline approach. Soft data again appears to improve the accuracy of the flash estimates, but the out-of-sample improvement is smaller than in-sample.

Concluding comments

This article has explored the effectiveness of PcGets in producing flash estimates of aggregate European Labour Cost Indices. The results tend to suggest simple baseline autoregressive models are the most effective for this purpose, although the conclusion may have been partly determined by some limitations within PcGets.

It is not a huge surprise that hard data models do not outperform the baseline. All the hard data suffers from the timeliness issue of not being fully and consistently available within the T+40 days threshold, hence only lags of the data can be used. There is no reason to expect lags of the factors that might determine labour costs to out-perform lags of the Labour Cost Index itself. Expressed in another way, lags of the Labour Cost Index seem to encapsulate the information held in the lags of the hard data indicators.

Soft data from unofficial consumer and business surveys is timely and can

be used contemporaneously. It is also qualitative enabling it to measure factors such as competitiveness effects, capacity and expectations of the future which are difficult for national statistics institutions to identify and record. Out-of-sample testing suggests that these variables do offer a slight improvement over the baseline. This is particularly so for the EU-27 country group where official data for some countries is more patchy/unreliable, increasing the usefulness of softer data.

Another factor explaining why autoregressive models may do so well is the sample period. Low inflation has generally prevailed in the European Union, meaning there has been an unprecedented era of nominal stability. Simple linear models like autoregressive models tend to perform well when the data they are trying to predict is fairly stable. If the data were less stable then other indicators may become more important, especially if there is a turning point or a change in regime in the data.

The use of soft data in official statistics, even for the purposes of forecasting, tends to create heated debate. It is mainly for this reason that the analysis in this article splits the use of hard and soft data in forming Flash estimates. Objections range from simple points of principle to concerns about the reliability and the quality of the data. Given that these arguments are

Table 3
MSE of different models of the LCI

Model	Baseline	Hard data	Soft data
EZ-15	0.0143	0.0139	0.0069
EU-27	0.0262	0.0231	0.0089

Source: Author's estimations

Table 4
Indicator selection for European LCI Flash estimates

Model	EZ-15	EU-27
Baseline	1997Q2–2003Q4: Labour Cost Index (–1) 1997Q2–2005Q2: Labour Cost Index (–1), (–2)	1997Q2–2003Q1: Constant Labour Cost Index (–1) Dummy 1998Q1 1997Q2–2005Q2: Constant Dummy 1998Q1 1997Q2–2005Q4: Labour Cost Index (–2), (–3) Dummy 1998Q1 1997Q2–2007Q2: Labour Cost Index (–2), (–3), (–4) Dummy 1998Q1
Hard	1997Q1–2003Q1: Constant Labour Cost Index (–4) Wages and salaries (–4) Unemployment (–1) 1997Q1–2004Q4: Labour Cost Index (–1) GDP deflator (–2)	1997Q2–2003Q1: Constant Labour Cost Index (–4) Wages and Salaries (–4) Manufacturing earnings - EU (–4) 1997Q2–2004Q4: Gross Value Added (–1) Manufacturing earnings- EU (–3) Private sector earnings- EZ (–1)
Soft	1997Q2–2003Q1: Labour Cost Index (–3) Industry Capacity Utilisation (–1) Construction Employment - next months (–1) Construction Prices - next months (–1) 1997Q2–2003Q3: Constant Labour Cost Index (–3) Construction Employment - next months (–1) 1997Q2–2005Q2: Labour Cost Index (–2), (–3) Construction Employment - next months (–1) Services Employment - next months (–1) Industry Prices - next months (–1) Construction Prices - last months (–1) 1997Q2–2006Q2: Labour Cost Index (–3) Industry Productive Capacity Construction Employment - next months (–1) Services Employment - next months (–1)	1997Q2–2003Q1: Constant Industry employment next (–1) Industry capacity utilisation (–1) 1997Q2–2003Q2: Labour Cost Index (–1), (–2), (–3) Industry capacity utilisation (0), (–1) Construction employment next (0), (–1) 1997Q2–2004Q4: Constant Labour Cost Index (–1), (–2) Industry competition ex-EU (–1) Industry capacity utilisation (–1) Industry production next (–1) Industry labour shortage (–1) Dummy 1998Q1 1997Q2–2005Q3: Constant Labour Cost Index (–1), (–3), (–4) Industry competition domestic Industry labour shortage Dummy 1998Q1 1997Q2–2006Q4: Labour Cost Index (–1), (–3), (–4) Industry competition domestic (0), (–1) Industry labour shortage Industry prices - next months (–1) Consumer prices - last months (–1) Consumer unemployment - next months (–1) Dummy 1998Q1 1997Q2–2007Q4: Constant Labour Cost Index (–1), (–3), (–4) Industry competition domestic Industry prices - next months (–1) Industry labour shortage Dummy 1998Q1

Source: Author's estimations

Table 5
Mean squared forecast errors of LCI Flash estimates

Model	Baseline	Hard data	Soft data
EZ-15	0.0127	0.0199	0.0117
EU-27	0.0332	0.0332	0.0269

Source: Author's estimations

unlikely to be resolved soon implies that, as the reported gains to using soft data are not clear-cut, it is unlikely they will be mutually accepted.

A further problem with the soft data models is evident from Table 5. It is clear that the variables making up the optimal model, as decided by PcGets, tend to be quite unstable. Legitimate concerns have recently been raised about the performance of business and consumer surveys in out-of-sample and real-time settings on these grounds; that they can tend to work well in some periods but not others.

Part of the instability in model selection though undoubtedly reflects the working of PcGets. The algorithm tends to really struggle if a number of highly correlated variables are presented in the starting GUM. The result is an over-specified terminal model where the parameters tend to be unstable over time. Again Table 5 would tend to confirm this as an issue. Soft data models, especially for the EU-27 model, are generally much larger in the number of right hand side variables, and more prone to being updated.

Notes

1. The group of EZ-15 countries includes: Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, Netherlands, Portugal, Slovenia and Spain. Slovakia joined in January 2009 after this work had been completed.
2. The group of EU-27 countries includes: EZ-15 group, Slovakia (EZ-16), Bulgaria, Czech Republic, Denmark, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Sweden and the UK.

CONTACT

✉ elmr@ons.gov.uk

REFERENCES

Castle J and Hendry D (2006): Extending the boundaries of automated model selection: non-linear models. Department of Economics working article, Oxford University.

Hendry D and Krolzig H-M (2001): Automatic econometric model selection using PcGets. Timberlake Consultants Press.