Representing Uncertainty about Response Paths: the Use of Heuristic Optimisation Methods

February 28, 2006

Anna Staszewska

Chair of Econometric Models & Forecasts

University of Łódź

Łódź

Poland

emfans@uni.lodz.pl

Keywords : bootstrapping, confidence bands, genetic algorithm, impulse response analysis

Abstract

In impulse response analysis the construction of intervals for the response at a particular time is a familiar topic. This paper considers the construction of confidence bands for the path of responses. It investigates the feasibility of procedures based on heuristic optimisation methods for constructing bootstrap confidence bands and evaluates the coverage properties of these bands for a stylised empirical VEC model.

1 Introduction

When a vector autoregressive model is fitted it is routine to carry out impulse response analysis. There is a considerable literature on how best to construct confidence intervals for the responses, that is for the response at a *particular* time. By contrast, the construction of confidence bands for the entire response path has received very little attention although it is a subject of considerable interest. Confidence intervals for responses are often presented as plots of intervals against time with the highest points in successive intervals and the lowest points joined so that the reader is almost invited to interpret the area between the two lines as a confidence band for the entire path. This is a rather hazardous proceeding for a coverage of 95%, say, for each of the intervals will not necessarily produce anything like a 95% coverage for the band. The present paper takes off from the belief that there should be some gain in taking a more considered approach than joining the intervals and hoping for the best. The methods proposed combine bootstrapping, which is now widely used for the construction of intervals, with the use of heuristic optimisation methods for finding the narrowest band for the entire path.

A recent paper (Staszewska (2006a)) on the construction of confidence intervals for impulse responses included a brief consideration of the construction of confidence bands for impulse response paths and gave some preliminary results on the performance of different procedures. The present paper has a more thorough discussion of the issues involved in constructing such bands and evaluates the performance of different methods when applied to a VEC model of the kind used in empirical work. It is found that the coverage probability of the ad hoc procedure of joining up the confidence intervals is substantially less than the desired value while the proposed methods for constructing confidence bands perform well for short horizons, at least, and always achieve much higher coverage probabilities than the joined up intervals.

2 Confidence intervals and confidence bands

We begin by noting the inconclusiveness of reasoning from the properties of intervals for particular responses to the properties of bands for the entire response path. The main point is that knowing the confidence levels for the intervals for each period may tell us little about the confidence level of the band produced by joining up the intervals.

The standard result linking joint probabilities to marginal probabilities is Bonferroni's inequality; see Miller (1980). It states that for any collection of events $\{A_i\}$

$$P\left(\cap A_i\right) \geq 1 - \left(\sum P\left(\overline{A}_i\right)\right)$$

and it applies to the present situation if we interpret A_i as being in the i-th confidence interval and $\cap A_i$ as being in all the intervals. If $P(A_i) = 0.95$ for $i = 1, .., 6$ then

$$P\left(A_i\right) \geq P\left(\cap A_i\right) \geq 1 - (6 \times 0.05) = 0.70$$

If we look ahead 12 periods then the lower bound falls to 0.40, while if we look ahead 24 periods the bound is uninformative. However the inequality can be used more positively to construct a confidence band, albeit a conservative one: thus to obtain a path coverage of 0.95 for 24 intervals a stringent probability of 0.998 is required for each interval. Another benchmark is the case when the $\{A_i\}$ are *independent*. In this case $P\left(\cap A_i\right) = \prod P\left(A_i\right)$ which for $P(A_i) = 0.95$, $i = 1, .., 24$ generates $P\left(\cap A_i\right) = 0.29$. In impulse response analysis a 'hit' in one interval is not independent of a hit in another and the result does not apply; rather successive hits are strongly positively associated.

The obvious condition $P\left(A_i\right) \geq P\left(\cap A_i\right)$ should not be overlooked for it implies that the performance of the joined up method for the whole path cannot be better than the performance for an individual period. In Monte Carlo experiments similar to those described in sections 5 and 6 below Staszewska (2006a) computed intervals for responses over a period of 24 months and found that the lowest coverage probability was around 0.7 (for a nominal level of 0.95); the joined up intervals had a coverage probability for the path of around 0.5. So the overall path performance was considerably worse than any of the individual interval performances.

3 The bootstrap procedure

The objective of impulse response analysis is to discover how the VAR responds to typical economic shocks. This objective has been interpreted in different ways, corresponding to different views of what is economically reasonable and intelligible. Three distinct forms of economic experiment have been proposed leading to traditional, generalised and orthogonalised impulse response analysis; the terms are from Koop, Pesaran & Potter (1996). The confidence bands proposed below can be applied to all forms of response analysis and results on the performance of the bands have been obtained for all three forms but, as the present interest is in the methods, most of the results reported relate to one form of economic experiment, that leading to the traditional analysis. The traditional analysis is based on the vector moving average representation of the VAR and the economic experiment involves altering the error associated with one variable, holding the other errors fixed. Staszewska (2006a) considered intervals for all three forms of experiment.

Impulse responses are functions of the autoregressive parameters and estimates of the responses are calculated using the same functions with estimates of the coefficients instead of the true parameter values. Recently bootstrapping has been used to obtain confidence intervals. However, although bootstrapping has undergone spectacular development since its introduction by Efron in 1979, its use in complicated models remains something of an art and numerous ways of proceeding have been proposed.

A brief account of interval construction is in order for, though our interest is in constructing bands, we exploit the methodology for constructing intervals and indeed we want to evaluate the practice of joining intervals to form bands. The interval estimates considered here are non-parametric versions of the standard construction introduced by Efron and a modification due to Hall (1992) for dealing with biassed estimates. Computations were also done for the parametric version but they are not reported here. Staszewska (2006a) constructed intervals using both versions and found very little difference in the intervals obtained. Although the parametric method uses information about the Monte Carlo DGP–that the disturbances are normal–it does not perform better than the nonparametric method which does not. Further methods for interval construction are described by Benkwitz, Lütkepohl and Wolters (2001).

Following the notation of Benkwitz, Lütkepohl and Wolters (2001), let ϕ_i be some general impulse response coefficient at period i, $\widehat{\phi}_i$ the estimator implied by the estimators of the model coefficients and $\widehat{\phi}_i^*$ the corresponding bootstrap estimator. The *standard method* of constructing a $(1-\gamma)\%$ confidence interval uses $s^*_{\gamma/2}$ and $s^*_{(1-\gamma/2)}$, the $\gamma/2$ and $(1-\gamma/2)$-quantiles of the bootstrap distribution of $\widehat{\phi}_i^*$ and is given by:

$$CI_S = [s^*_{\gamma/2}, s^*_{(1-\gamma/2)}].$$

The interval CI_S is the construction as it appears in Efron and Tibshirani (1993). They point out, however, that the interval may not have the desired coverage probability, for instance, if $\widehat{\phi}_i$ is a biased estimator of ϕ_i. One of the modifications of CI_S developed to deal with this problem is Hall's (1992) method.

In Hall's method the confidence intervals are constructed using $t^*_{\gamma/2}$ and $t^*_{(1-\gamma/2)}$, the $\gamma/2$ and $(1-\gamma/2)$-quantiles of the distribution of the difference $\widehat{\phi}_i^* - \widehat{\phi}_i$. The interval is defined as

$$CI_H = [\widehat{\phi}_i - t^*_{(1-\gamma/2)}, \widehat{\phi}_i - t^*_{\gamma/2}].$$

This method is the one preferred by Benkwitz, Lütkepohl and Wolters (2001). It performed well in the experiments conducted by Staszewska (2006a).

4 Constructing confidence bands

The basic method for constructing a 95% confidence interval for the response for a given period is to order the B bootstrap values for that period and identify the top and bottom 2.5 percentiles. In the case of *paths* there is no underlying ordering because the paths typically cross. The guiding idea of discarding extreme paths can be retained but the implementation has to be changed; in fact, the idea can be implemented in several different ways.

Formal optimisation methods can be used. There are two obvious possibilities. For a 95% confidence band choose from the B bootstrap paths (24 component vectors if the analysis is done for 24 periods) the $0.95 \times B$ that are closest together; of course different metrics produce different selections of vectors. Another possibility is to find the narrowest band which contains $0.95 \times B$ of the paths. We investigated the feasibility of doing this using the genetic algorithm (GA) of Carroll (2001).

In the implementation of the genetic algorithm B, the number bootstrap replications, is 1000 and the confidence level is 95%. Finding the narrowest band containing the $0.95 \times B$ paths is equivalent in the present application to rejecting 50 paths and obtaining the band as an outline or envelope of the remaining 950 paths. The 50 paths rejected were chosen from 150 pre-selected paths. The pre-selection was based on the following reasoning. The candidate paths for rejection are those providing the 50 biggest and 50 smallest bootstrapped responses for all periods. For the data generated in the Monte Carlo experiments and for the original WKK data (described below) the number of such paths varied roughly between 300 and 600 paths for the 24 period horizon, excluding 400-700 "internal" paths from consideration. Choosing 50 paths out of 600 or even 300 proved to be computationally very demanding, with slow genetic algorithm convergence. In the present study the computations are not done once but done as many times as there are Monte Carlo replications and for this exercise to be tractable it seemed judicious to limit the set of candidate-paths for rejection. After some experimentation it was decided to limit the set to 150 paths. They were chosen as the ones that would most enlarge the area of the band produced by the set of "internal" paths.

Because these formal optimisation methods are computationally demanding some heuristic procedures were considered as well. Heuristic method 1 (H1) was an iterative procedure with steps:

a) For each period for which the paths are constructed look at the most extreme of the bootstrap values (the smallest and the largest) and identify the paths they belong to. In the case of 24 periods there are 48 extreme points to look at and maximally 48 paths they belong to (it is likely however that some of the extreme points belong to one and the same paths).

b) Find a path which provides the biggest number of extreme points and remove this path from the set of B paths.

c) Repeat steps a) and b) now for $B-1$ remaining paths, then for $B-2$ and so on until we've removed $(1-0.95)\times B$ paths

d) Obtain the 95% confidence band as the outline (envelope) of the remaining $0.95\times B$ paths.

Heuristic method 2 (H2) is a more elaborate iterative procedure which uses information about how far the extreme paths are from the others. (Distance was interpreted both in the l_1 sense and in the l_2 sense):

a) For each period identify the most extreme of the bootstrap values and the paths they belong to.

b) For each path providing any extreme points calculate the sum of distances from these points to the next most extreme points (or, in the squared variant of this method, the sum of squared distances).

c) Remove the path with maximal sum.

d) Repeat steps a), b) and c) until $(1 - 0.95) \times B$ paths are removed.

e) Obtain the 95% confidence band as the outline of the remaining $0.95 \times B$ paths.

The methods just described make *no* use of the original estimated response path $\widehat{\phi}_1, ..., \widehat{\phi}_{24}$. In the case of confidence intervals the Hall method which used the $\widehat{\phi}_i$ values as well as the bootstrap values $\widehat{\phi}_i^*$ was quite effective. So a method which used these values was considered; this involved identifying paths near the estimated response path. We refer to it as the neighbouring paths method (NP).

a) For each bootstrap path $\left\{\widehat{\phi}_i^*\right\}$ compute the distance from the estimated path $\left\{\widehat{\phi}_i\right\}$.

b) From the B bootstrap paths $\left\{\widehat{\phi}_i^*\right\}$ choose the $0.95 \times B$ that are closest to $\left\{\widehat{\phi}_i\right\}$.

c) Obtain the 95% confidence band as the envelope of these paths.

As before 'closeness' was interpreted both in a squared error sense and in an absolute error sense. This method differs from the other heuristic methods in that it is not trying to mimic the working of the optimisation methods.

Thus we have six methods to compare, the genetic algorithm and five heuristic methods, heuristic 1, heuristic 2 in two variants and neighbouring paths in two variants. In evaluating these methods two points are of interest: closeness to the designed coverage probability of 0.95

and width of the band. The optimal band would be the narrowest band with the correct coverage probability. In the present study most attention is paid to the closeness to the nominal value. Before proceeding to a comparison of the characteristics of the different methods we give an example of their use with real data.

Figures 1-3 show bands constructed for the VEC model for wages, prices and labour productivity developed by Welfe, Karp and Kębłowski (2004); see section 5 below for details of this study which forms the basis of the Monte Carlo experiments performed in this paper. The three figures show responses to a shock to w, wages, in wages, prices and productivity following the traditional analysis; in this instance the actual historical series are used, not simulated series. For clarity only three bands are shown: for one joined-up interval method (standard nonparametric) and for two joint methods, the genetic algorithm and one heuristic (neighbouring paths absolute); the omitted graphs for the joint methods are similar to those included. The original estimated response is the line in the middle.

The three plots show that the bands produced by the new joint methods are generally wider than the bands produced by joining up intervals. Close examination of the paths reveals, however, that it is not the case that the former bands lie outside the latter for all periods; that is so in Figure 3 but not in the other two cases. Whether any of the lines are far enough apart to achieve the designed coverage probability is considered in section 6.

5 The Monte Carlo design

The coverage properties of the confidence bands obtained by the techniques described in the previous section were investigated by Monte Carlo methods. The specification of the data generating process used for the Monte Carlo analysis, like that in Staszewska (2006a), follows the

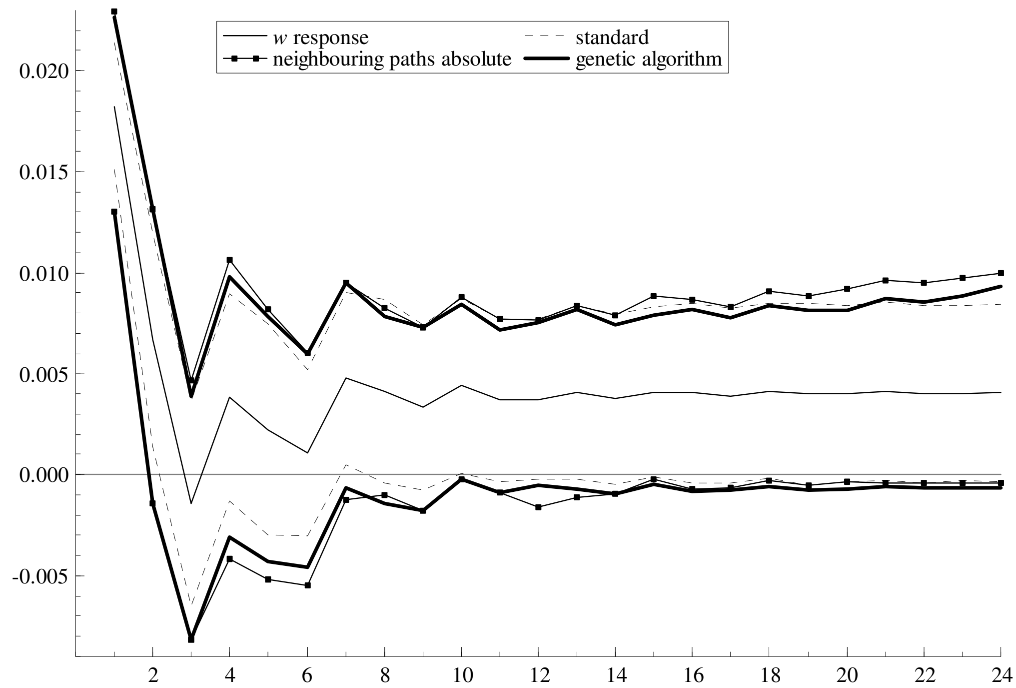

Figure 1: Traditional impulse response analysis: shock to w, response from w.

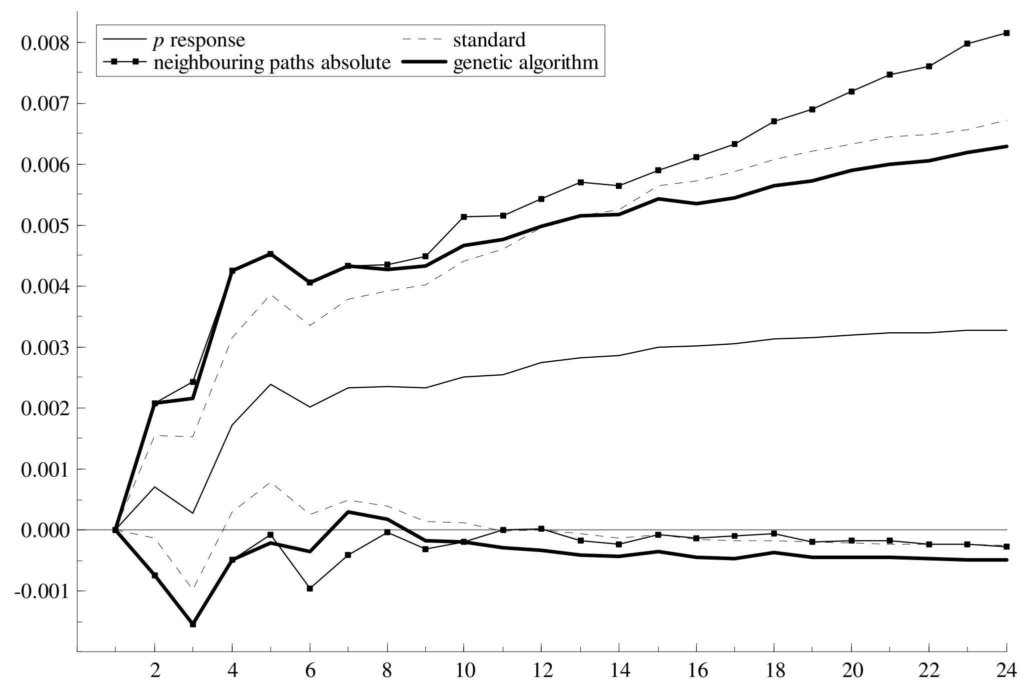

Figure 2: Traditional impulse response analysis: shock to w, response from p.

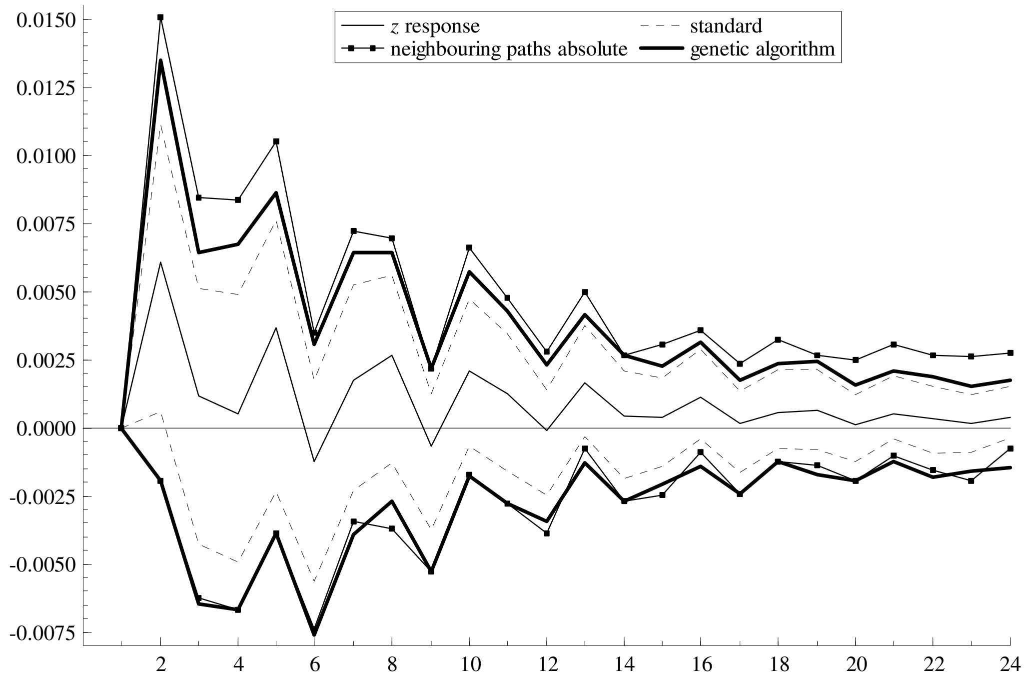

Figure 3: Traditional impulse response analysis: shock to w, response from z.

VEC model for wages, prices and labour productivity developed by Welfe, Karp and Kębłowski (2004). The WKK model is of a size and complexity common in modern applied work. It has 3 modelled or endogenous variables, wages, prices and labour productivity, 3 cointegrating relations and 5 weakly exogenous variables. The system with 4 lags in the VAR was re-estimated and the long-run parameters restricted, leaving the short-term dynamics unrestricted. These estimated values were used as the true values in the Monte Carlo data generating process. In the Monte Carlo DGP the errors are assumed to be normally distributed; the sample size is 120, as in WKK.

The Monte Carlo DGP is a VECM with the long-run relationships:

$$\begin{aligned} w - p &= 8.793 + 0.233z - 0.015U \\ p &= -6.135 + 0.61w - 0.61z + 0.39pm \\ z &= -8.423 + 0.76kl + 0.039fd + 0.195V. \end{aligned}$$

The variables are w - wages, p - consumer price index, z - productivity, U - unemployment rate, pm - import prices, kl - capital-labour ratio, fd - cumulated foreign direct investments (FDIs) per employee, V - share of GDP generated in the private sector in total GDP measuring privatisation where small letters denote natural logarithms. The following variables are assumed to be weakly exogenous: U, pm, kl, fd, V and the values of these variables are the historical values from I. 1993-XII.2002 as used in the WKK study to which the reader is referred for further details. The data is monthly and seasonally unadjusted.

For each of the methods we follow the same procedure:

- In a single Monte Carlo replication: generate data from the Monte Carlo DGP and estimate the model:
  - start bootstrapping: treating the estimated model as the bootstrap DGP, generate data from it 1000 times and construct confidence bands for the response path;
  - check whether the true response path from the Monte Carlo DGP falls into the confidence band.
- Produce 1000 Monte Carlo replications and count the proportion of times that the band contains the true path.

In constructing the bootstrap samples we do not re-estimate the cointegrating relationships every time but use the original estimates; it has been found (see Staszewska (2006) and Benkwitz, Lütkepohl and Wolters (2001)) that re-estimation makes very little difference to the intervals created while it takes a lot of computing time.

6 Experimental results and evaluation

This section reports the results of the main Monte Carlo experiments investigating coverage probabilities and also the results from some subsidiary studies. Path coverage probabilities were obtained for the band methods introduced above and for the naive joined-up (JU) intervals procedure. Table 1 shows the performance of the various methods for constructing confidence bands for 24 periods ahead; the economic experiment is the same as that behind Fig. 1, viz. the traditional impulse response analysis considering the effects on the three endogenous variables of a one standard deviation shock to w.

Table 1: coverage probabilities compared

	traditional experiment		
	w	p	z
JU (stand)	0.562	0.646	0.483
JU (Hall)	0.507	0.624	0.55
GA	0.75	0.755	0.85
H1	0.708	0.758	0.816
H2 (abs)	0.731	0.763	0.823
H2 (sq)	0.747	0.763	0.826
NP (abs)	0.745	0.79	0.873
NP (sq)	0.747	0.778	0.884

While the probabilities are all substantially less than 0.95 all of the joint methods perform much better than the naive methods, JU (stand) and JU (Hall). The six joint methods offer similar improvements over the naive methods but the last two neighbouring paths (NP) methods seem to be the best overall. In the case of shocks to other endogenous variables the joint methods offered similar improvements over the joined-up methods. In all the cases the neighbouring paths methods were somewhat better than the other joint methods.

The improved performance of the band methods over the joined-up intervals is not an artifact of the economic experiment chosen. Table 2 shows the coverage probabilities obtained for a variety of bands–though not including the genetic algorithm–for non-traditional impulse response analyses for a shock to w. Four distinct experiments are reported, the generalised analysis and different orthogonalised analysis corresponding to different ways of ordering the variables; different orderings of the variables do not always generate distinct responses and so in the table some equivalent experiments are shown together.

Table 2: orthogonalised & generalised

	orth wpz, wzp, gen.			orth pzw, zpw			orth pwz			orth zwp		
	w	p	z	w	p	z	w	p	z	w	p	z
JU (stand)	0.491	0.569	0.379	0.522	0.629	0.474	0.465	0.559	0.366	0.553	0.666	0.466
JU (Hall)	0.441	0.52	0.327	0.461	0.597	0.511	0.415	0.504	0.325	0.484	0.623	0.517
H1	0.587	0.684	0.518	0.686	0.74	0.813	0.56	0.667	0.549	0.707	0.771	0.813
H2 (abs)	0.601	0.652	0.572	0.705	0.74	0.818	0.593	0.64	0.593	0.734	0.768	0.812
H2 (sq)	0.62	0.655	0.59	0.718	0.751	0.818	0.607	0.644	0.602	0.747	0.775	0.818
NP (abs)	0.655	0.715	0.613	0.725	0.776	0.872	0.642	0.699	0.636	0.77	0.799	0.877
NP (sq)	0.638	0.689	0.655	0.731	0.76	0.882	0.63	0.683	0.687	0.764	0.816	0.87

In two of these sets of economic experiments the performance of the methods is similar to that of the traditional methods reported in Table 1. However, in the case of the first and the third, performance is inferior for all the methods.

Staszewska (2006a) found that the performance of the bootstrapped confidence intervals–in terms of coverage probability–deteriorated as the horizon was lengthened. So the effect of the horizon on the performance of the band methods was investigated with the experiments repeated for 6 and 12 month horizons. Improved performance is to be expected at these shorter horizons, in the case of the joined up method because the Bonferroni lower bound rises and in all cases because the bootstrapping is less effective for the remoter periods. Table 3 presents results on performance for shorter horizons with the information from Table 1 on the long horizon for comparison.

Table 3: different horizons

	6 months			12 months			24 months		
	w	p	z	w	p	z	w	p	z
JU (stand)	0.679	0.802	0.681	0.596	0.722	0.563	0.562	0.646	0.483
JU (Hall)	0.685	0.813	0.74	0.587	0.723	0.66	0.507	0.624	0.55
GA	0.88	0.924	0.895	0.815	0.834	0.867	0.75	0.755	0.85
H1	0.89	0.978	0.965	0.814	0.866	0.934	0.708	0.758	0.816
H2 (abs)	0.87	0.901	0.872	0.8	0.838	0.849	0.731	0.763	0.823
H2 (sq)	0.872	0.901	0.873	0.804	0.84	0.851	0.747	0.763	0.826
NP (abs)	0.942	0.93	0.94	0.842	0.861	0.906	0.745	0.79	0.873
NP (sq)	0.925	0.921	0.917	0.842	0.86	0.9	0.747	0.778	0.884

All methods show substantially better performance for 6 months ahead than for the longer horizons. The method with best overall performance was the neighbouring paths absolute but even this performed badly for the 24 months case.

To gain insight into the working of the joint methods coverage probabilities for individual periods were calculated, i.e., the proportion of the B paths lying within the band at the particular period 12. Some of these probabilities are shown in the following table (traditional analysis of a shock to w response in w) for the genetic algorithm and the two neighbouring paths methods.

Table 4: coverage probabilities for individual periods

	NP (abs)	NP (sq)	GA
1	0.992	0.992	0.98
6	0.991	0.991	0.995
12	0.9	0.897	0.905
24	0.8	0.8	0.825

While the early periods show the necessary excess over 0.95 the later periods are disappointingly deficient. In a study of the properties of confidence intervals Benkwitz, Lütkepohl and Neumann (2000) found a similar problem with remote periods and suggested it could be a small sample problem. We are currently investigating that possibility in the present set-up.

Good confidence procedures require not only coverage probabilities that are close to the designed level but also small bands. For the different joint methods the coverage probabilities are very similar–what about the size of the bands? These were calculated for the WKK data (see Figures 1-3 above). The areas of the confidence bands are also very similar in size, as the following table (traditional shock to w response in w) shows.

Table 5: areas associated with different bands

	w	p	z
GA	0.22287	0.11851	0.15673
H1	0.22539	0.11968	0.16205
H2 (abs)	0.22416	0.11982	0.15674
H2 (sq)	0.22578	0.12104	0.15933
NP (abs)	0.23740	0.13054	0.17670
NP (sq)	0.23372	0.12749	0.16994

This table in conjunction with Figures 1-3 suggests that the different methods are doing much the same thing, although only the genetic algorithm has as its explicit objective the minimisation of this area. It can be seen that the methods that did best in the coverage properties evaluation have the largest band areas.

7 Conclusions

The main finding is that it is possible to produce confidence bands which are much more successful in achieving a desired confidence level than the technique of joining up confidence intervals. This conclusion is likely to hold in other circumstances although the detailed findings are subject to the important caveat that they relate to a restricted set of experiments using a particular DGP. Of the methods considered the neighbouring paths methods perform best. However a serious unresolved problem with all of the bands–and the same holds for intervals–is the poor coverage probabilities achieved for responses at long horizons. A possible explanation for this poor performance is the small estimation sample of 120 observations (see Benkwitz, Lütkepohl and Neumann (2000)). The possibility that this is so is being investigated. Another matter under investigation is the narrowness of the band generated by the various methods; the present paper has concentrated on the proximity of the achieved coverage probability to the desired level. For the present, however, the results on the band methods seem promising and justify further research.

References

[1] Benkwitz A., H. Lütkepohl, M. H. Neumann (2000), Problems Related to Confidence Intervals for Impulse Responses of Autoregressive Processes, *Econometric Reviews*, **19** (1), 69-103.

[2] Benkwitz A., H. Lütkepohl, J. Wolters (2001), Comparison of Bootstrap Confidence Intervals for Impulse Responses of German Monetary Systems, *Macroeconomic Dynamics*, **5**, 81-100.

[3] Carroll, D. L. (2001), GA-version 1.7a: FORTRAN Genetic Algorithm Driver, CU Aerospace, Urbana, IL, http://cuaerospace.com/carroll/ga.html.

[4] Efron B., R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

[5] Hall P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer.

[6] Koop G., M. H. Pesaran & S. M. Potter (1996), Impulse Response Analysis in Nonlinear Multivariate Models, *Journal of Econometrics*, **74**, 119-147.

[7] Miller R. G (1980), *Simultaneous Statistical Inference*,.second edition, Springer, New York.

[8] Staszewska A. (2006), Impulse Response Analysis in Conditional Models: Wages, Prices and Productivity in Poland, *Metody ilościowe w naukach ekonomicznych, Szóste Warsztaty Doktorskie z Zakresu Ekonometrii i Statystyki*, red. A. Welfe, Oficyna Wydawnicza SGH, (forthcoming).

[9] Staszewska A. (2006a), Confidence Bands for Impulse Response Paths, Proceedings of the 32nd International Conference *Macromodels* 2005, (forthcoming).

[10] Welfe, A., P. Karp and P. Kębłowski (2004), Modelling Polish Economy: An Application of SVEqCM, in A. Welfe (ed.): *New Directions in Macromodelling*, Elsevier, Amsterdam.