# Evolving Takagi Sugeno Modelling with Memory for Slow Processes

**S. McDonald [a] and Dr. P. Angelov [b]**

[a] Corresponding author, Department of Communication Systems, Infolab21, Lancaster University, Lancashire, LA1 4WA, UK. s.mcdonald1@lancaster.ac.uk. Tel: +44(0)1524510525. Fax: +44(0)1524510493

[b] Department of Communication Systems, Infolab21, Lancaster University, Lancashire, LA1 4WA, UK. p.angelov@lancaster.ac.uk

**Abstract**

Evolving Takagi Sugeno (eTS) models are optimised for use in applications with high sampling rates. This mode of use produces excellent prediction results very quickly and with low memory requirements, even with large numbers of input attributes. In this paper eTS modelling is adapted for optimality in situations where memory usage and processing time are not specific requirements. The new method, *eTS with memory*, is demonstrated on two financial time series, both the fullband signals and after decomposition by the discrete wavelet transform. It is shown that the use of previous inputs and multiple iterations in eTS can produce better predictions for signals which are not dominated by the characteristics of noise.

**Keywords**

Evolving Takagi Sugeno, Fuzzy, Modelling, Noise, Discrete Wavelet Transform

## 1. Introduction

Consider the time series of daily foreign exchange closing prices between the British Pound Sterling (GBP) and the European Currency Unit (EUR). An application that predicts the next value based on current knowledge is not subject to processing time restrictions because there is a full day between each new value in the series. Furthermore, such an application may not be required to run on a resource-restricted device such as a microprocessor or field-programmable gate array (FPGA). In this situation it makes sense to utilise the available time and resources to generate a better prediction.

Traditional eTS [1] is a fuzzy rule-based modelling method that occurs in two basic stages: evolution of the rule base and an adapted recursive least squares (RLS) estimation technique. The rule base antecedents for the fuzzy rule-based modelling are produced by online evolving clustering [2] in a recursive and non-iterative way. This method is based on the position of the data vector in the data space relative to any existing cluster centres and the other data vectors. The adapted RLS estimation method in eTS traditionally only uses the most recent data point in order to minimise the memory requirements and to improve the processing time. For certain time series data there will be improvements in prediction when previous data vectors are also used during RLS estimation. This should certainly be the case for slowly changing time series as opposed to bursty, noise-like series. Finally, classical eTS schemes only perform one iteration of the recursive least squares algorithm in order to minimise the processing time. Since RLS is known to converge with the target on iteration [3, 4] given certain conditions on the input data, it seems prudent to iterate further if there is time available to do so and the predictions are improving.

These modifications are implemented in eTS and are demonstrated by the modelling and prediction of GBP v EUR closing price data using the original time series and subbands of scale produced by the discrete wavelet transform [5, 6]. As a second example the experiment is repeated for quarterly US Gross Domestic Product Data. The mean square prediction errors for each time series are compared for different combinations of previous inputs and numbers of iterations in the new eTS with memory model.

**2. Methods**

The classical eTS method that is modified in this paper is outlined in [1] as a seven stage scheme:

1. Initialization of the rule base with the first input (generation of the rule antecedents)
2. Arrival of the new data point
3. Recursive calculation of the informative potential of the new data point
4. Recursive recalculation of the informative potentials of the previous data points given the arrival of a new data point in the data space
5. Possible adjustment or update of the rule base dependent on the potential of the new point relative to the potentials of the existing points
6. Recursive calculation of the consequent parameters
7. eTS prediction of the value at the next time step

This method assumes that each data point is a possible cluster centre and then mathematically quantifies this possibility as the potential, based on the Euclidean distances, $d_{ij}$, from all other points.

$$P_i = \sum_{j=1}^{n} e^{-\alpha d_{ij}}$$

$P_i$ is the potential of point *i* to be a cluster centre and *n* is the number of data points. The constant α defines the radius value which specifies what fraction of the normalised data space represents a cluster centre's range of influence in each dimension. Effectively this limits the maximum size of the cluster from the cluster centre. The value $d_{ij}$ is the distance between points *i* and *j* defined by:

$$d_{ij}(x_i, x_j) = \| x_i - x_j \|^2$$

The double bar ( ‖ ) operator indicates the distance between the two data points in each dimension of the data space.

The eTS rule-base takes the form:

**if($x_1$ is $N_{i1}$) and … and ($x_n$ is $N_{in}$) then ($y_i = a_{i0} + a_{i1}x_1 + \dots + a_{in}x_n$); i={1,2,…,R}**

where **x** is the input data vector of size **n** (and also the number of antecedents for each rule), $\mathbf{y_i}$ is the output of the $\mathbf{i^{th}}$ rule, derived from the antecedents (cluster centres) and the model parameters $\mathbf{a_{il}}$, **l={1,2,…,n},** and **R** is the number of rules. Each rule has a firing level which is determined by the product of the membership functions, **μ**, in this case Gaussians, and the values in the input vector:

$$\tau_i = \prod_{j=1}^{n} \mu_{ij}(x_j)$$

where $\mu_{ij} = e^{-\alpha \| x_j - x_{ij}^* \|^2}$**,** i={1,2,…,R}, and j={1,2,…,n}.

The final output from the model is then a combination of each rule's input values, $\mathbf{x_e^T}$, and consequent parameters, $\boldsymbol{\pi_i} = [\mathbf{a_{i0}\ a_{i1} \dots a_{in}}]^{\mathbf{T}}$, weighted by that rule's ***normalized*** firing level [1], $\boldsymbol{\lambda_i} = (\boldsymbol{\tau_i} / \boldsymbol{\Sigma^{R}_{j=1}\tau_j})$:

$$y = \sum_{i=1}^{R} \lambda_i x_e^T \pi_i \qquad \{\text{equation } 1\}$$

The locally optimal scheme for finding the parameters in traditional eTS is implemented as the minimisation of each rule's cost function [1]:

$$J_{Li} = (Y - X^T \pi_i)^T \Lambda_i (Y - X^T \pi_i)$$

where $\Lambda_i$ is a diagonal matrix holding the $\boldsymbol{\pi_i(x_k)}$ values in the main diagonal. The minimisation is achieved by the recursive weighted (wRLS) method derived in [1]:

$$\hat{\pi}_{ik} = \hat{\pi}_{ik-1} + c_{ik} x_{ek} \lambda_i (x_k)(y_k - x_{ek}^T \hat{\pi}_{ik-1}) \qquad \{\text{equation 2}\}$$

note that the subscript ***e*** denotes the input vector has been extended with a unitary value at the start. The ($y_k$- $x_{ek}^T \pi_{ik-1}$) term is the expression of the error of prediction using the parameters from the previous stage. The covariance matrix is also updated recursively [1]:

$$c_{ik} = c_{ik-1} - \frac{\lambda_i (x_k) c_{ik-1} x_{ek} x_{ek}^T c_{ik-1}}{1 + \lambda_i (x_k) x_{ek}^T c_{ik-1} x_{ek}} \qquad \{\text{equation 3}\}$$

Note that, for traditional eTS, equation {2} is only evaluated once, ie. undergoes one iteration, and then only using the most recent data vector, $\mathbf{x_{ek}}$, and the $\boldsymbol{\pi_{ik-1}}$ consequent parameters and $\mathbf{c_{ik-1}}$ covariance matrix from the previous step.

The modifications presented here affect stage six of this scheme. The recursive clustering of the data space and adjustment of the rule base antecedents (steps two to five) and the eTS prediction (step seven) proceed exactly as before. The new developments involve storing previous data points up to a certain number (ie. the memory) which could be a fixed value or, indeed, all of the previous data vectors, ie. growing window eTS. Let the number of previous inputs be **Q**. Equations {2} and {3} above are carried out at stage six using the most recent rule base antecedents, the most recent consequent parameters, and the most recent covariance matrix, *but using data vectors starting at (k-Q) through to k successively*. Furthermore, equation {2} is iterated for each of the Q data points, based on some previously supplied limit, say G. Pseudocode for the new weighted RLS is presented to clarify this method:

For each of the **Q** previous data points from the oldest to the most recent

    Update the normalised firing levels, $\boldsymbol{\lambda_i}$

    For each rule

        Update the covariance matrix, see equation {3}

        For **G** iterations

            Update the consequent parameters, see equation {2}

        End loop

    End loop

End loop

Finally, at stage seven, the prediction of time step ***k+1*** is produced from equation {1} above, and the process begins from stage two at the next time step.

## 3. Application

Recall that eTS represents the data space by its rules which are the centres of clusters and best represent the data in each cluster in terms of informative potential. Given the new developments explained in this paper, old input data can be reassessed in terms of new rules, or rules that may have changed after the data vector was first presented to the model. Essentially the previous data is being assessed in terms of a newer, and hopefully more representative, rule base. The logical extreme of such eTS with memory would be to form a rule base on all the previous data and **then** use all those rules to arrive at better consequent parameters for every vector in the data space in order of time steps. This makes sense because a more complete rule base should lead to better consequent parameters and the parameters for any given data point are calculated from the parameters at the previous stage. Thus the final prediction for the next, as yet unknown, value should be improved. This will, of course, depend on the nature of the signal being modelled. Stationary ergodic processes are better modelled by minimum mean square error methods, as used in the Wiener filter. In fact, for a sufficiently large stationary data set, the RLS method, as used in the Kalman filter, converges with the Wiener solution. Thus eTS should be chosen for modelling non-stationary processes. However, research has suggested [7] that RLS convergence depends on the input signal to noise ratio (SNR) and the value of the forgetting factor for previous inputs. Thus, signals with a high SNR or with a broadband frequency spectrum *may* not benefit, and *may* even suffer, from further iteration of RLS. The expected improvements in prediction accuracy for low frequency signals (trends and low noise data) are shown in figure one where increasing the number of iterations reduces the prediction errors, ie. the model converges with the process, to some limiting value which is signal dependent.

Signals which are not dominated by low or high frequency should be improved by iteration but not to the same extent as trend data. Noisy signals could benefit from RLS iteration. When high frequency energy dominates the signal then all outcomes are possible: better prediction, worse prediction, no change, or a combination of all three at different time steps.

It has already been suggested that a more representative rule base should lead to better parameter estimation for previous data vectors, but this should only hold true for as long as the rules are relevant to the current ‘state’ of the signal for a non-stationary process. In fact, in standard time-based RLS a forgetting factor is used to handle the non-stationarity. However, because of the firing levels of rules in the eTS system any rules which were not previously relevant will not affect the predictions significantly. Thus the use of previous data points in wRLS gives the model the chance to produce better results depending on the dynamics and statistics of the data. It is expected that lower frequency signals will benefit most from the use of previous data points in eTS with memory. This is because there is a clear relationship between consecutive data points in such signals. Adjacent points in high frequency signals are often independent of each other, so eTS with memory may not improve the prediction of such signals.

As with many fuzzy systems the choice of membership function can drastically affect the rate of convergence to the underlying process function and the quality of prediction [8]. Also, the data space of some processes may be better represented by clusters formed with some partition criterion other than the informative potential.

**4. Experimentation**

The test data used in the first experiment is a series of 981 closing prices of GBP against EUR as quoted on the London Stock Exchange between 20$^{th}$ October 2004 and 16$^{th}$ July 2007. This data is used with the eTS with memory system where values one to 980 are inputs and values two to 981 are the prediction targets. This is the fullband data set. The same data is also decomposed by the discrete wavelet transform, using the Symlet number eight mother wavelet, into one approximation subband and ten detail subbands, each of which is handled by a different eTS with memory model. This is the subband data set. In every case the data is standardised to a zero mean and unitary standard deviation before modelling. There is no pre-training of the models. Again the current value is used to predict the closing value at the next day. The tests are run for one, ten, and one hundred iterations of the RLS algorithm using one to ten previous data vectors. The results for a single iteration and one previous input are exactly the results that traditional eTS would have produced. All the other results derive from the new developments presented here. All errors are expressed as mean square errors (MSE). Table one shows the basic statistics of the financial fullband data and one large-, medium-, and small-scale subband (approximation, subband five, and subband one respectively) ***before*** standardisation. It is immediately clear from table one that the majority of the energy is in the approximation subband. Note also that the detail subbands have means very close to zero. It is also clear that variability relative to the mean is higher in the lower level detail subbands. This is illustrated in figure two where the ***standardised*** financial approximation subband is shown alongside the standardised detail subbands five and one, where the latter exhibit much more variation and are also non-linear. The prediction MSEs for the fullband data set are presented in figure three. These results are very interesting. The smallest prediction MSE is produced by traditional eTS (ie. one previous data point and one RLS iteration). The use of previous data does not improve predictions at all when one or 100 iterations are performed. When ten iterations are used the MSE is higher using one to five previous inputs but the minimum occurs when ten previous data vectors are used.

The results are quite different for the wavelet-decomposed data, see a subset of these results in figures four, five, and six. For the approximation subband, figure four, and detail subbands ten to seven, not shown, the smallest MSEs are produced by eTS with memory using 100 iterations, as expected with lower frequency signals. The use of ten iterations performs only marginally worse and minimum MSEs are reached using three (for ten iterations) and four (for 100 iterations) previous data points in the RLS scheme. Using one RLS iteration, the smallest MSE occurs using ten previous data points for the approximation and subbands ten and nine, as expected, and with two previous data points for detail subband eight. For subband seven a minimum MSE occurs using one previous input. For detail subbands six and five the best results are still achieved using more RLS iterations, also as expected, although the MSEs generally increase as more previous inputs are used. The use of one iteration produces the smallest prediction errors when more than one previous input is used in the RLS scheme. There is also a noticeable difference between the MSEs achieved using ten and 100 iterations. The results for subband five are shown in figure five. The results are different again for the smallest scale (equivalent to highest frequency) subbands, ie. subbands four to one. eTS with one RLS iteration produces smaller prediction errors regardless of how many previous data points are considered. Furthermore the MSEs from one iteration generally decrease as more previous data points are used. The smallest errors for subband one occur when one previous data vector is used, ie. using traditional eTS. The results for subband one are shown in figure six.

The second experiment uses annualised and seasonally adjusted US GDP data from January 1$^{st}$ 1947 to April 1$^{st}$ 2005. The data represents billions of U.S. dollars for a total of 234 quarterly observations. The fullband and subband (eleven subbands using the Symlet number eight mother wavelet) data are modelled and predicted exactly as in experiment one. The statistics of some of the data before standardisation are presented in table two. Again, the majority of the energy is found in the approximation subband and the variability about the mean increases for the smaller scale subbands (ie. those with lower indices). The prediction MSEs for the fullband data are very interesting, see figure seven. In this experiment fullband predictions ***are*** improved using eTS with memory. In fact, eTS with ten iterations produces markedly lower errors for every case and the minimum occurs using two previous data points. When using one to five previous points the predictions are better with one RLS iteration but from six to ten the results are better with one hundred RLS iterations.

The subband results for the GDP data set are very similar to those from experiment one. For the largest scale subbands (approximation and details ten and nine) the use of one hundred iterations always produces the best results with ten iterations only slightly worse. One iteration produces much larger MSEs. The improvement using previous data vectors is less marked than in experiment one. For detail subbands eight to six, one hundred iterations again produces the best results with ten iterations producing larger MSEs and one iteration worst in every case. The use of previous data points degrades the predictions. This was not expected for these signals. Subband five of the GDP data is best modelled by one hundred iterations for all number of previous points except two, where ten iterations are better. From subbands four to one the use of one iteration yields the best predictions, as was the case in experiment one. Furthermore, one hundred iterations produce the worst results. Interestingly, for subbands two and one in experiment two the use of more previous data vectors improves the prediction MSEs in every case.

**5. Results Summary**

The modelling of subbands in both experiments clearly demonstrates that multiple RLS iterations in eTS with memory improves the prediction of the large scale subbands where there is a significant amount of energy but relatively little variation of the signal around the mean. The power spectra of this type of signal have the majority of the energy in the low frequency range. For the smaller scale subbands, where there is more high frequency energy, extra RLS iterations lead to poorer predictions. Furthermore, the use of previous data points also improves the modelling of the large scale subbands. The relationship between the small scale subbands and the use of previous data vectors is less clear. In experiment one MSE reductions were seen for each subband when more previous data points were used from the approximation to detail subband nine. From there the smaller scale subband predictions generally became worse with more previous data vectors. In experiment two the use of previous data points improved the modelling of the approximation subband and detail subbands ten, nine, two, and one. This is particularly interesting because subbands one and two from experiment two have a greater deviation about the mean than the same subbands in experiment one.

Modelling of the fullband signals produced different results in the two experiments. For the GBP v EUR data set traditional eTS performed best with MSEs generally increasing with more eTS iterations and more previous data points. For the GDP data the best results were achieved by ten RLS iterations. The best MSEs for one and ten iterations occurred with two previous data vectors whereas the poorly-performing 100 iteration models actually improved with more previous data points. As with the small scale subbands discussed above, it seems that a broader range of energy, as with the experiment one fullband data set, leads to no improvement from the use of extra iterations in the new eTS method. It seems that there is a complex relationship between the use of previous data points and the dynamics of the signal being modelled in terms of improving the prediction accuracy of eTS models with memory. The optimal number of RLS iterations can be found using the standard techniques in machine learning: a target MSE performance, a minimum MSE performance gradient improvement between successive iterations, or a maximum number of iterations.

**6. Conclusions**

Applications which are not restricted by high sampling frequencies or reduced resources can benefit from the use of the new eTS with memory model presented here. If the signal being modelled is not dominated by high frequency energy then prediction accuracy should increase with more iterations of the RLS algorithm. Using previous data points in the weighted RLS scheme can further improve prediction, particularly for lower frequency signals. It seems, however, that there are other factors affecting the use of previous inputs such that these improvements depend on the dynamics of the signal and the time-ordered distribution of the data in the data space. Applications with low sampling rates should certainly be tested with eTS with memory to see if these benefits can be exploited.

With the development of eTS with memory it can be seen that applications with very different requirements and/or restrictions can take advantage of the predictive power of eTS modelling. Time and resources can, in effect, be balanced against the required prediction accuracy.

Future work should try to identify how the optimum number of previous data points can be calculated rather than using trial and error. In fact, this optimum may change over time and require some degree of adaptivity.

**References**


[1] Angelov P.P. and Filev D.P., "An Approach to Online Identification of Takagi-Sugeno Fuzzy Models" *IEEE Trans. Systems, Man, and Cyber. Part B: Cybernetics*, vol. 34 (1), pp. 484-498, February 2004

[2] Angelov P., "An Approach for Fuzzy Rule-base Adaptation using On-line Clustering", *Inter. Journ. of Approximate Reasoning*, vol. 35 (3), pp. 275-289, 2004

[3] Ljung L., "Analysis of Stochastic Gradient Algorithms for Linear Regression Problems" *IEEE Trans. Inf. Theory*, vol. 30(2), pp. 151-160, March 1984

[4] Eweda E., Macchi O., "Convergence of the RLS and LMS Adaptive Filters", *IEEE Trans. Circ. and Systems*, vol. 34 (7), pp.799-803, July 1987

[5] Haar A., "Zur Theorie der orthogonalen Funktionen-Systeme" *Math. Ann*, vol. 69, pp.331-371, 1910

[6] Daubechies I., "The wavelet transform, time-frequency localization and signal analysis" *IEEE Trans. Inf. Theory*, vol. 36, pp.961-1005, 1990

[7] Bershad N. and Macchi O., "Comparison of RLS and LMS algorithms for tracking a chirped signal" Proc. *ICASSP-89, Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp.896-899. ISSN 1520-6149

[8] Mitaim S. and Kosko B., "The Shape of Fuzzy Sets in Adaptive Function Approximation", *IEE Trans. Fuzzy Systems*, vol. 9 (4), pp.637-656, August 2001

| Data | Mean | Standard Deviation |
|---|---|---|
| **Fullband** | 1.4654 | 0.0202 |
| **Approximation** | 1.4628 | 0.0092 |
| **Detail 5** | -0.000009 | 0.0068 |
| **Detail 1** | 0.00000056 | 0.0015 |

**Table 1: Basic Statistics of the GBP v EUR Data Set before Standardisation**

| Data | Mean | Standard Deviation |
|---|---|---|
| **Fullband** | 5090.0556 | 2697.8189 |
| **Approximation** | 5519.564 | 477.8162 |
| **Detail 5** | 3.7045 | 69.8597 |
| **Detail 1** | 0.0205 | 13.337 |

**Table 2: Basic Statistics of the GDP Data Set before Standardisation**

**Figure 1: Expected MSE Improvements for Low Noise and Trend Signals**
**Figure 2: Standardised Data: Approximation and Detail Subbands 4, and 1**
**Figure 3: Fullband GBP v EUR prediction MSEs**
**Figure 4: Prediction MSEs for the Approximation Subband**
**Figure 5: Prediction MSEs for Subband 5**
**Figure 6: Prediction MSEs for Subband 1**
**Figure 7: Fullband Prediction MSEs for the GDP Data Set**

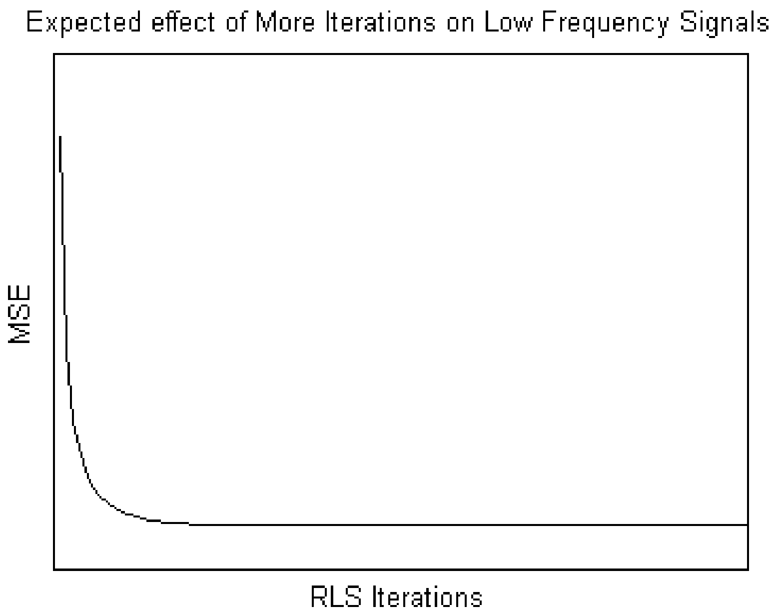


**Figure 1: Expected MSE Improvements for Low Noise and Trend Signals**

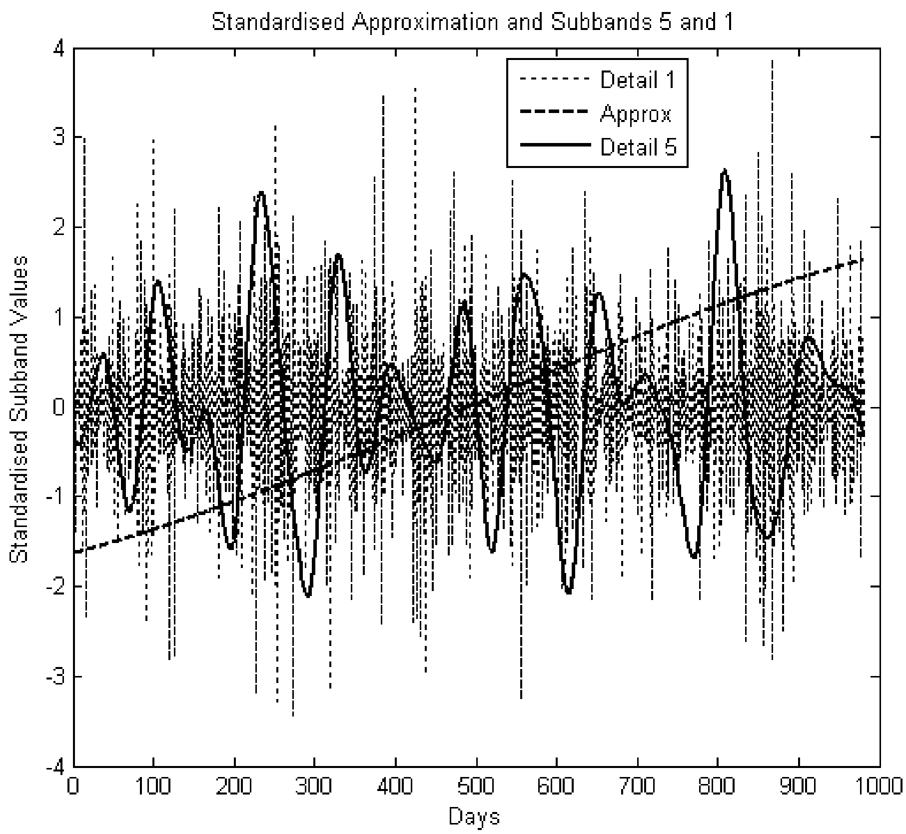


**Figure 2: Standardised Data: Approximation and Detail Subbands 5 and 1**

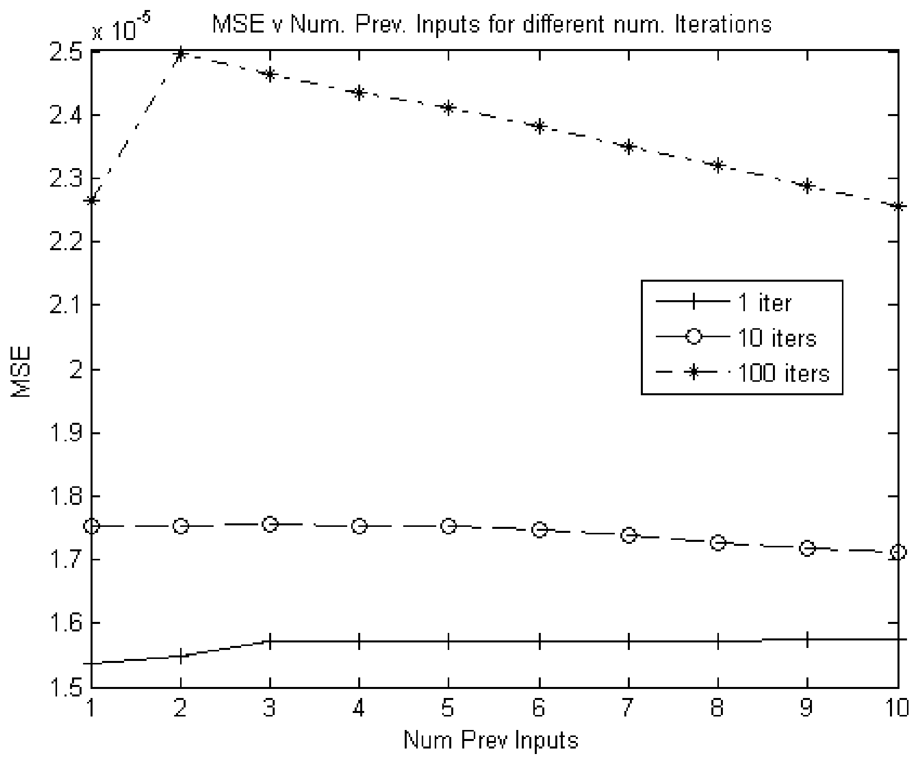


**Figure 3: Fullband GBP v EUR prediction MSEs**

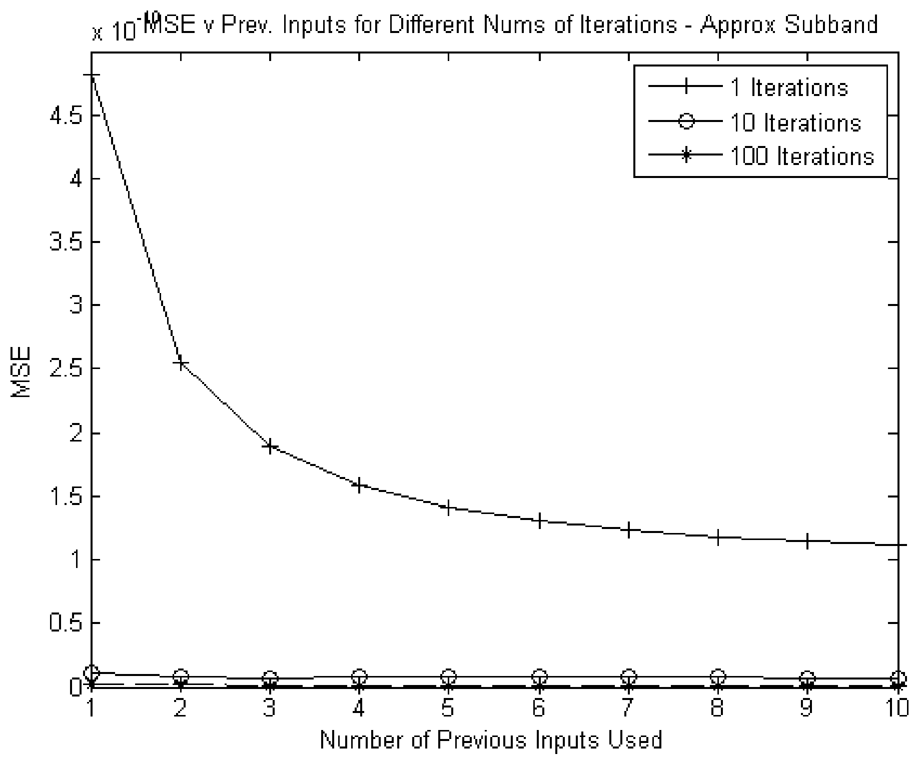


**Figure 4: Prediction MSEs for the Approximation Subband**

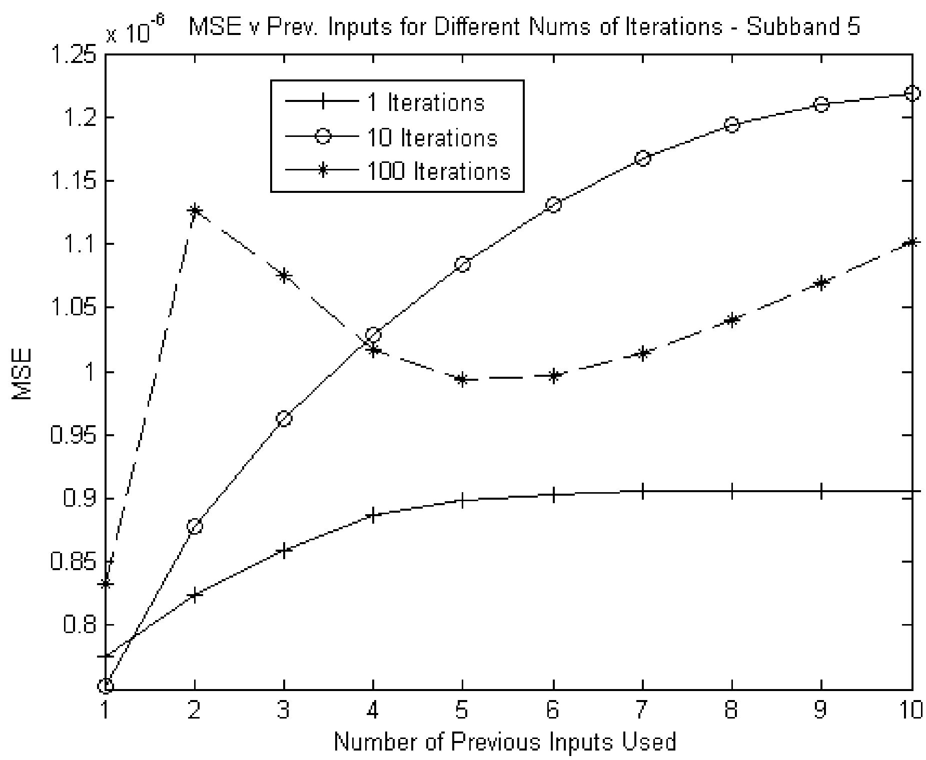


**Figure 5: Prediction MSEs for Subband 5**

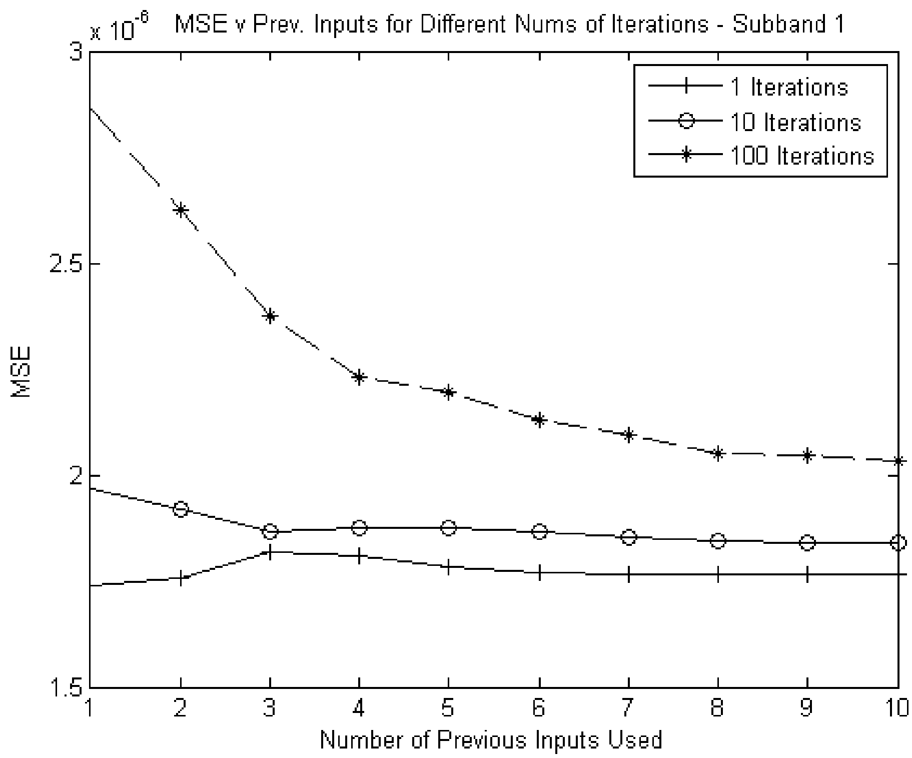


**Figure 6: Prediction MSEs for Subband 1**

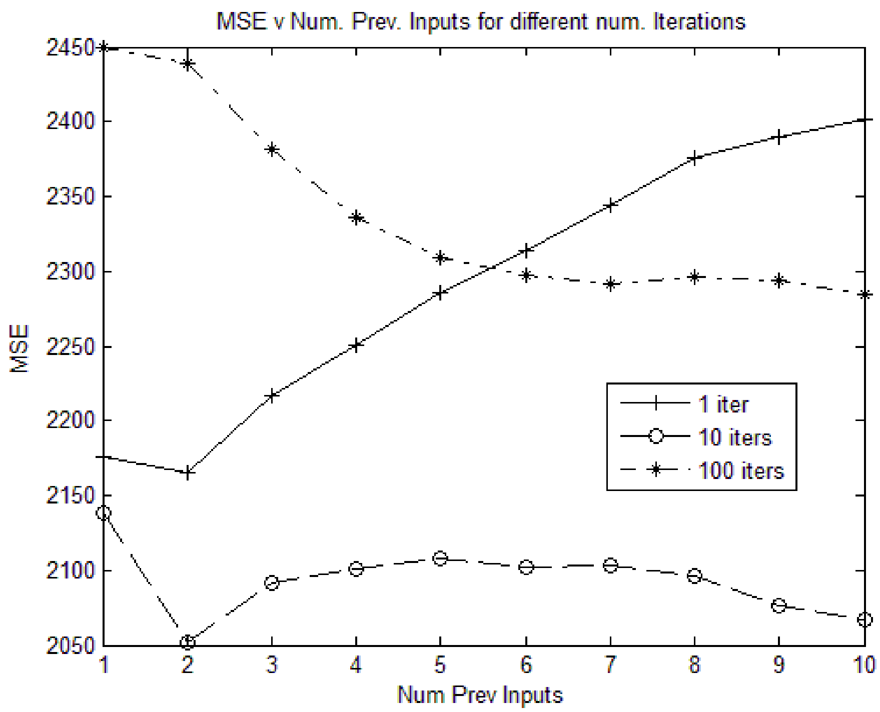


**Figure 7: Fullband Prediction MSEs for the GDP Data Set**