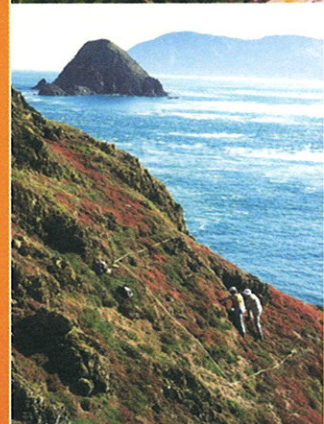
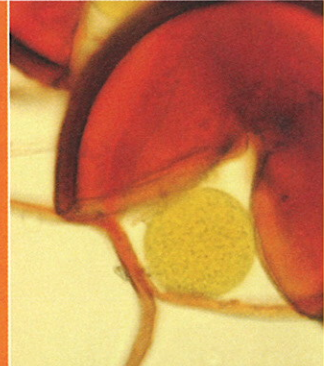


Bio-Protection & Ecology Division



Further Sensitivity Analysis of Simple Evolving Connectionist Systems Applied to the Lincoln Aphid Data Set

Michael J. Watts and S.P. Worner



Bio-Protection & Ecology Technical Report No. 3

CHRISTCHURCH
NEW ZEALAND
www.lincoln.ac.nz



Lincoln
University
Te Whare Wānaka o Aoraki

ISSN: 1177-7435 ISBN: 0-86476-177-5 978-0-86476-177-5

Further Sensitivity Analysis of Simple Evolving Connectionist Systems Applied to the Lincoln Aphid Data Set

Michael J. Watts and S.P. Worner

wattsm2, worner@lincoln.ac.nz

Chapter 1

Introduction

This report presents two further experiments over the Aphid data set. The first is a more detailed investigation of the sensitivity of Simple Evolving Connectionist System (SECoS) networks to the exclusion of various combinations of inputs. This is in contrast to the previous work (Watts, 2004), where only the effect of excluding single variables was investigated.

The second experiment investigates a hypothesis that attempts to explain the results found in the first experiment.

Chapter 2

Sensitivity Analysis

2.1 Introduction

The goal of this experiment is to further investigate the anomalous results in (Watts, 2004). These anomalous results were that, while excluding time-stepped variables at $w - 1$ and w for the average rain variable, and at $w - 2$ and $w - 2$ for the cumulative rain variable, produced a significant decrease in performance, excluding the entire variables (that is, excluding all measures of average rain or cumulative rain) did not significantly effect the performance of the SECoS. Anomalous results were also found for the aphid count variable, whereby excluding any of the previous time-steps did not cause a significant change in performance, but excluding the entire variable caused a serious decrease in performance. It was hypothesised in (Watts, 2004) that the reason for the results over the average and cumulative rain variables was because the *change* in these variables was more important than the actual value of the variables. It was also hypothesised that the presence of any previous aphid count was sufficient for the network to learn.

2.2 Method

The hypothesis described above was tested in the following manner. For each variable investigated, every possible combination of presence and absence of the time stepped variables was investigated. These combinations are listed in Table 2.1. Note that combinations A, D, F, G and H have already been investigated in previous work.

Ten fold cross-validation was used to investigate each combination, and the same data sets were used as in previous work.

2.3 Results

The results for the combinations of the average rain variable are presented in Table 2.2. The results are presented as the mean of the Mean Absolute Error (ME) and the standard deviation (σ). The results for the cumulative rain variable are presented in Table 2.3, and for the aphid count variable in Table 2.4.

Combination	$w - 2$	$w - 1$	w
A	0	0	0
B	0	0	1
C	0	1	0
D	0	1	1
E	1	0	0
F	1	0	1
G	1	1	0
H	1	1	1

Table 2.1: Possible combinations of time-stepped variables.

Combination	Mean	σ
A	2.42	0.26
B	2.43	0.38
C	2.26	0.28
D	2.43	0.23
E	2.36	0.23
F	5.61	1.17
G	5.71	1.28
H	2.46	0.30

Table 2.2: Results for omitting combinations of the average rain variable

2.4 Discussion

An exhaustive comparison of the accuracies of each combination of variables was performed. These comparisons consisted of two-tailed t -tests. The results presented in this section are for $p = 0.01$ and are ‘accept’ if the null hypothesis was not rejected, and ‘reject’ if the null hypothesis was rejected.

The results of these tests are presented in Table 2.5 for the average rain variable, Table 2.6 for the cumulative rain variable, and Table 2.7 for the previous aphid counts variable.

Inspection of the results in Tables 2.2 and 2.5 shows that the average rainfall vari-

Combination	Mean	σ
A	2.42	0.26
B	2.43	0.39
C	2.27	0.29
D	5.50	1.21
E	2.36	0.23
F	2.26	0.47
G	2.30	0.23
H	2.46	0.30

Table 2.3: Results for omitting combinations of the cumulative rain variable

Combination	Mean	σ
A	13.82	0.47
B	2.30	0.26
C	2.03	0.35
D	2.46	0.36
E	2.04	0.24
F	2.21	0.33
G	2.15	0.34
H	2.46	0.30

Table 2.4: Results for omitting combinations of the aphid count variable

	H	G	F	E	D	C	B
A	accept	reject	reject	accept	accept	accept	accept
B	accept	reject	reject	accept	accept	accept	
C	accept	reject	reject	accept	accept		
D	accept	reject	reject	accept			
E	accept	reject	reject				
F	reject	accept					
G	reject						

Table 2.5: Results of hypothesis tests for average rain variable.

ables at time w and $w - 1$ are both necessary only if $w - 2$ is present. If $w - 2$ is present and both w and $w - 1$ are absent, then there is not a significant change in accuracy. If $w - 2$ is present and either w or $w - 1$ are present, then there is a significant degradation in accuracy. This lends further support to the hypothesis that it is the change in this variable that is significant.

	H	G	F	E	D	C	B
A	accept	accept	accept	accept	reject	accept	accept
B	accept	accept	accept	accept	reject	accept	
C	accept	accept	accept	accept	reject		
D	reject	reject	reject	reject			
E	accept	accept	accept				
F	accept	accept					
G	accept						

Table 2.6: Results of hypothesis tests for cumulative rain variable.

Inspection of the results in Tables 2.3 and 2.6 shows that the cumulative rainfall variable at time $w - 2$ is necessary only if w and $w - 1$ are also present. If w and $w - 1$ are not present, then omitting $w - 2$ will not cause a significant change in accuracy. This again lends further support to the hypothesis that it is the change in this variable that is significant, rather than the values of the variable *per se*.

The results in Tables 2.4 and 2.7 show that the presence of the aphid count variable at time w is not needed for accurate prediction. In two cases (combinations B and E)

	H	G	F	E	D	C	B
A	reject	reject	reject	reject	reject	reject	reject
B	accept	accept	accept	accept	accept	accept	
C	reject	accept	accept	accept	accept		
D	accept	accept	accept	accept			
E	reject	accept	accept				
F	accept	accept					
G	accept						

Table 2.7: Results of hypothesis tests for aphid count variable.

removal of time w measurement significantly improved the prediction accuracy.

Chapter 3

Replacing Variables with the Change in Variables

3.1 Introduction

It was hypothesised in (Watts, 2004) that the change in variables was more important than the actual values of the variables. The results in the previous chapter did not disprove this. To further investigate this hypothesis the work in this section was carried out.

3.2 Method

The hypothesis discussed above was investigated by replacing the average rain and cumulative rain variables with the “delta” values of those variables. That is, instead of including measurements at times w , $w - 1$ and $w - 2$, the differences between these variables were used. The average rain variable was replaced by the deltas listed in Table 3.1. The cumulative rain variable was replaced by the deltas listed in Table 3.2. The results of the sensitivity analysis indicated that the delta $w - (w - 1)$ was not needed for the cumulative rain variable.

$w - (w - 2)$
$w - (w - 1)$
$(w - 1) - (w - 2)$

Table 3.1: Average Rain Deltas

$w - (w - 2)$
$(w - 1) - (w - 2)$

Table 3.2: Cumulative Rain Deltas

The creation, training and evaluation of the SECoS networks was performed as in the original work with SECoS (Watts, 2004). That is, a network was trained on Set

A, and tested on Sets A, B and C. The network was then further trained on Set B, and again tested on Sets A, B and C. Ten-fold cross-validation was again used.

3.3 Results

The results of this experiment are presented in Table 3.3. The accuracy is reported as the Mean Absolute Error (ME) over each data set. Table 3.3 presents the mean and standard deviation over all ten folds of the data.

Recall Set	Train Set	
	A	B
A	2.43 / 0.37	2.56 / 0.37
B	11.39 / 1.70	1.00 / 0.35
C	10.91 / 1.43	10.97 / 1.56
Full	4.17 / 0.35	3.25 / 0.32

Table 3.3: Results of SECoS trained with delta variables.

3.4 Discussion

Two sets of statistical tests were performed. The first compared the accuracy of the SECoS networks before and after further training was carried out. This was to evaluate the adaptation and forgetting of the networks. The second compared the accuracy of the networks to those that were trained using the original data set. In both cases, two-tailed t -tests were used, and the hypothesis evaluated at $p = 0.01$.

The results of the t -tests for the first set of comparisons are presented in Table 3.4, while the results for the second set of comparisons are presented in Table 3.5. In both cases, an entry of ‘reject’ indicates that the null hypothesis was rejected, while an entry of ‘accept’ indicates that the null hypothesis was not rejected.

δA	δB	δC	δF
accept	reject	accept	reject

Table 3.4:

The results in Tables 3.3 and 3.4 indicate that the SECoS networks were able to adapt to Set B without significantly forgetting the previous data. This matches the results over the original data set.

AA	AB	AC	AF
accept	accept	accept	accept
BA	BB	BC	BF
accept	accept	accept	accept

Table 3.5:

Inspection of the results in Table 3.5 shows that SECoS networks trained with the modified data set (that is, with delta values instead of the original values) performed with an accuracy that was not significantly different to that of the SECoS networks trained over the original data set.

Chapter 4

Conclusion

The experiments described in this report have investigated the effect of various time-steps of three variables of the aphid prediction problem, namely the average rainfall, cumulative rainfall, and previous aphid counts. It was found that for the average and cumulative rainfall variables, the difference between timesteps is more important than the actual values of the variables. That is, the rate at which the variables change is the most important contribution of these two variables. For the aphid data variable, it was found that the aphid counts at time w are not needed for accurate predictions.

Chapter 5

Future Work

The next step in this project will be to calculate prediction intervals (Baxt and White, 1995) for the ANN. Preliminary work has also commenced towards reducing the number of variables used, by identifying and removing variables that are highly correlated.

Bibliography

- Baxt, W. G. and White, H. (1995). Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction. *Neural Computation*, 7:624–638.
- Watts, M. (2004). Comparison of multi-layer perceptrons and simple evolving connectionist systems over the aphid data set. Technical report, National Centre for Advanced Bio-Protection Technologies.